Translating mass spectra to peptides with deep learning

Melih Yilmaz

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Commitee:

William Stafford Noble, Chair

Sewoong Oh

Meliha Yetisgen

Program Authorized to Offer Degree:
Computer Science and Engineering

University of Washington

Abstract

Translating mass spectra to peptides with deep learning

Melih Yilmaz

Chair of the Supervisory Committee:
William Stafford Noble
Department of Genome Sciences

Tandem mass spectrometry is the leading technique to study proteins at scale, and a fundamental challenge in mass spectrometry-based proteomics is the identification of the peptide that generated each acquired tandem mass spectrum. Approaches that leverage known peptide sequence databases cannot detect unexpected peptides and can be impractical or impossible to apply in some settings. Thus, the ability to assign peptide sequences to tandem mass spectra without prior information—*de novo* peptide sequencing—is valuable for tasks including antibody sequencing, immunopeptidomics, and metaproteomics. Although many methods have been developed to address this problem, it remains an outstanding challenge in part due to the difficulty of modeling the irregular data structure of tandem mass spectra. In this work, we describe Casanovo, a machine learning model that uses a transformer neural network architecture to translate the sequence of peaks in a tandem mass spectrum into the sequence of amino acids that comprise the generating peptide. Casanovo is trained on a repository-scale dataset and it significantly advances the state-of-the-art in *de novo* peptide sequencing. We show that Casanovo's superior performance improves the analysis of immunopeptidomics and metaproteomics experiments and allows us to delve deeper into the dark proteome. Finally, we go beyond the *de novo* peptide sequencing

problem and demonstrate Casanovo's capabilities as a foundation model in mass spectrometry proteomics.

Contents

Acknowledgments

To my co-advisors, Bill Noble and Sewoong Oh, who have guided and supported me every step of the way. Choosing your advisor is the most important decision that one has to make in grad school and I am very fortunate to have been mentored by you.

To the rest of my committee, who gave their valuable feedback at major milestones of this process.

To all my collaborators, including Wout Bittremieux and Will Fondrie, who helped build and realize my research agenda.

To the members of Noble Lab, who not only provided useful feedback and insightful discussions but also camaraderie.

To my friends, who never ceased to cheer for me along the way. I am grateful for your friendship.

To my parents, who have been there for me since before day one. I owe so much in life to your unconditional love and support.

Lastly, to my wife, who has been the best companion I could ever ask for in this long journey.

DEDICATION

to my dear wife, İrem

Chapter 1

Introduction

Tandem mass spectrometry provides a high-throughput framework for identifying and quantifying proteins in complex biological samples, making it the most popular analytical technique to characterize proteomes, the complete set of proteins expressed by an organism [1]. A key challenge in the analysis of tandem mass spectrometry data lies with determining the exact protein content from observed mass spectra at scale. At the core of this challenge is the spectrum identification problem, in which given an observed mass spectrum and the associated mass and charge of the peptide (known as the *precursor*) that is responsible for generating the spectrum, we must infer the amino acid sequence of the precursor peptide.

To better illustrate how the spectrum identification problem arises, let us consider the basics of how tandem mass spectrometry works. A tandem mass spectrometer measures mass-to-charge (m/z) ratios of charged peptides in a two-scan process. Proteins from a biological sample are enzymatically digested into peptides and a first scan (MS1) measures the m/z of the intact peptide (also known as the precursor); the peptide is then fragmented and the resulting fragments are analyzed in a secondary scan (MS2). This MS2 scan is carried out on a population of (ideally) homogeneous peptide sequences, each of which is randomly fragmented at one location along

the peptide backbone. As a result, the fragmentation scan contains peaks that correspond to prefixes (called *b-ions*) and suffixes (*y-ions*) of the peptide, each with an associated charge state. Thus, the primary data object, the MS2 spectrum, consists of a bag of peaks, where each peak is characterized by an *m/z* value and its associated intensity. Fundamentally, spectrum identification involves identifying these peaks as b- or y-ions by comparing their m/z and intensity values to recover the amino acid sequence of the generating peptide. This task is challenging because some of the expected b- and y-ion peaks may be missing, and some additional peaks may appear in the spectrum, created by losses of small molecular groups during fragmentation or by multiple cleavage events occurring on the same peptide. Spectra also contain noise, including experimental noise from the instrument as well as chemical noise produced by contaminants, other peptides, or non-peptide molecules.

The standard computational method for solving the spectrum identification problem is enumerative, scoring each observed spectrum with respect to a list of candidate peptides (i.e., peptides whose masses are close to the observed precursor mass associated with the spectrum) and reporting the best-scoring peptide-spectrum match (PSM) per spectrum [49]. However, the drawback to any database search methodology is that it requires that we specify *a priori* which peptides might occur in the sample. Such an approach is often sensible when analyzing samples from a species, such as human, with a well-characterized genome sequence. However, relying on a database prevents the detection of unexpected peptide sequences, such as those that arise from genetic variation. A sequence database also cannot be used for the analysis of some types of immunopeptidomics data [64], in antibody sequencing [63], or in vaccine development when searching for bacterial peptides present on the surface of infected cells [47]. Finally, constructing an accurate database for metaproteomic analyses, such as the human microbiome or environmental samples, is nearly impossible [48]. Such settings require *de novo* peptide sequencing from the acquired mass spectra, i.e. directly from the tandem mass spectra without relying on a reference protein database.

Early *de novo* methods used heuristic search [61] or dynamic programming [44, 16, 26] to score peptide sequences against each observed spectrum. Machine learning has provided state-of-the-art performance on this task since 2015 [43], and the most recent methods when we started working on this problem in 2021 employed deep neural networks [63, 54, 72, 36]. Despite the advancing the frontiers for *de novo* peptide sequencing, these deep learning-based methods suffered from several limitations, with the then state-of-the-art method correctly assigning peptides to only half of the spectra identified with high confidence by a database search [54]. Underlying these limitations, we identified factors such as a lack of large scale high quality training sets that powered the rise of deep learning in other domains as well as techniques to natively encode high-resolution MS2 data, and turned to the sequence learning literature, particularly neural machine translation, for inspiration.

To address these issues, we formulated *de novo* peptide sequencing as a sequence-to-sequence translation problem and developed Casanovo, the first transformer-based *de novo* sequencing method [76, 75]. The chapters of this dissertation focuses on different aspects of our contributions to the *de novo* peptide sequencing literature through Casanovo. Chapter 2 introduces Casanovo and our benchmarking efforts against existing *de novo* tools. Chapter 3 describes an improved version of Casanovo with repository-scale training as well as its application to various problem domains. Chapter 4 goes beyond *de novo* sequencing and demonstrates Casanovo's capabilities as a mass spectrometry proteomics foundation model. Chapter 5 concludes the dissertation with a brief discussion of Casanovo's impact on the *de novo* sequencing literature and future work.

Chapter 2

Casanovo: *De Novo* Mass Spectrometry Peptide Sequencing with a Transformer Model

This chapter is adapted with minimal modification from:
Melih Yilmaz, William Fondrie, Wout Bittremieux, Sewoong Oh, William S Noble *De novo* mass spectrometry peptide sequencing with a transformer model. *Proceedings of the 39th International Conference on Machine Learning*, PMLR 162:25514-25522, 2022.

2.1 Introduction

As outlined in Chapter 1, tandem mass spectrometry is an essential framework to study proteins at scale and the spectrum identification problem remains a fundamental challenge in determining the exact protein content of complex biological samples. The standard method for solving this problem is enumerative, scoring each observed spectrum with respect to a list of candidate peptides (i.e.,

peptides whose masses are close to the observed precursor mass associated with the spectrum) and reporting the best-scoring peptide-spectrum match (PSM) per spectrum.

However, the drawback to any database search methodology is that it requires that we specify *a priori* which peptides might occur in the sample. Such an approach is often sensible when analyzing samples from a species, such as human, with a well-characterized genome sequence. However, relying on a database prevents the detection of unexpected peptide sequences, such as those that arise from genetic variation. A sequence database also cannot be used for the analysis of some types of immunopeptidomics data [64], in antibody sequencing [63], or in vaccine development when searching for bacterial peptides present on the surface of infected cells [47]. Finally, constructing an accurate database for metaproteomic analyses, such as the human microbiome or environmental samples, is nearly impossible [48]. Such settings require *de novo* peptide sequencing from the acquired mass spectra.

Early *de novo* methods used heuristic search [61] or dynamic programming [44, 16, 26] to score peptide sequences against each observed spectrum. Machine learning has provided state-of-the-art performance on this task since 2015 [43], and recent methods employ deep neural networks [63, 54, 72, 36]. Although *de novo* search tools are improving, there is still a long way to go. The most recent report [54] suggests that state-of-the-art methods achieve peptide-level recall (i.e., the percentage of spectra with the correctly assigned peptide) of 39-60%, depending on the dataset. However, this percentage is calculated only with respect to ground truth spectra that, by definition, were previously identified with high confidence by a database search.

Additionally, all of these methods involve complicated modeling schemes (Table 2.1) featuring different neural networks for different sub-tasks such as convolutional neural networks (CNNs) for spectrum peak embedding and spectrum processing, and recurrent neural networks (RNNs) for peptide sequence processing. These methods also include complex post-processing steps that involve either matching the predicted peptide mass with the spectrum's observed precursor mass

	DeepNovo	SMS	PointNovo	Casanovo
CNN for spectrum peak embedding	✓	✓		
CNN for spectrum processing	✓	✓		
RNN for peptide sequence processing	✓	✓	✓	
PointNet			✓	
Transformer				✓
Dynamic programming post-processor	✓		✓	
Database search post-processor		✓		
Precursor *m/z* filter				✓
Discretization of *m/z* axis	✓	✓		

Table 2.1: **Comparison of deep learning methods for *de novo* peptide sequencing.** Casanovo introduces a simpler yet more powerful architecture of a transformer for *de novo* peptide sequencing.

using dynamic programming or refining low confidence predictions with a database search-like procedure. The necessity of discretizing the mass-to-charge (m/z) axis of the mass spectra also remains a setback for all existing methods except PointNovo, necessitating a trade-off between low binning resolution (hence low sequencing accuracy) and higher model complexity (hence longer inference time).

In this work, we propose Casanovo, a transformer framework for *de novo* peptide sequencing (Figure 2.1). Casanovo uses the self-attention mechanism to translate directly from a variable-length sequence of observed spectrum peaks to a variable-length sequence of amino acids, analogous to a neural machine translation model in the natural language processing setting. Importantly, Casanovo takes individual spectrum peaks, together with the precursor mass and charge, as input, without resorting to discretization of the m/z axis, and learns to predict the generating peptide sequence in a supervised setting in which ground truth sequences are obtained with database search. Unlike existing methods, Casanovo does not employ an additional RNN to process peptide sequences and replaces the dynamic programming post-processing step with a simple delta mass filter, offering a simpler yet more powerful framework.

We train and evaluate our model on a multi-species benchmark dataset using an established cross-validation framework which involves testing on spectra with never-before-seen peptide

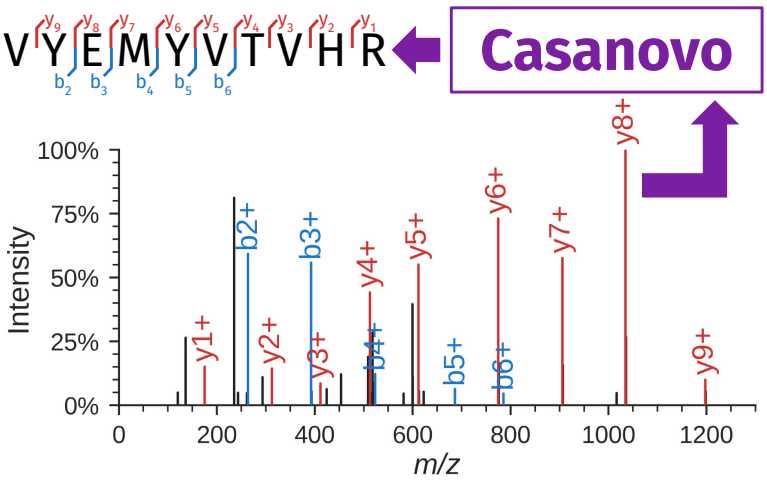


Figure 2.1: **Casanovo performs *de novo* peptide sequencing.** Casanovo takes as input an observed spectrum and produces the sequence of the generating peptide (e.g., VYEMYVTVHR). In the spectrum, peaks corresponding to b- and y-ions of the associated peptide are in color, and black peaks correspond to unexpected fragmentation events or noise. The spectrum annotation was created using spectrum_utils [8].

labels, through cross-species prediction. Our experiments show that Casanovo predicts peptide sequences with markedly higher precision relative to the state-of-the-art methods, DeepNovo and PointNovo, and does so using a model with fewer parameters and requiring much shorter inference time. Finally, we benchmark several variants of the model to demonstrate the robustness of some of our modeling choices.

2.2 Related Work

Early *de novo* methods used heuristic search (Lutefisk [61]) or dynamic programming (PEAKS [44] and SHERENGA [16]) to score peptide sequences against each observed spectrum. The PepNovo algorithm [26] uses a similar dynamic programming approach but employs a probabilistic score function that takes into account various chemical and physical rules governing peptide fragmentation. This model is closely related to the hidden Markov model that is, to our knowledge,

the first application of machine learning to the *de novo* peptide sequencing task [24]. A decade later, the Novor algorithm [43] achieved improved performance by using a decision tree as the score function in a dynamic programming algorithm.

The first deep neural network algorithm for *de novo* peptide sequencing, DeepNovo [63], combines two different network architectures—a convolutional neural network and a long short term memory (LSTM) network—each of which aims to predict the subsequent amino acid, given a spectrum and a peptide prefix. These two scores are combined in a dynamic programming procedure to yield the predicted peptide sequence. The recently described SMSNet algorithm [36] uses a network architecture similar to that of DeepNovo but also offers a post-processing step in which low-confidence amino acids are replaced by making use of a user-supplied peptide database. A competing method, pNovo 3 [72], works in three steps: (1) a traditional dynamic programming approach generates a set of candidate peptides for a given spectrum, (2) a previously described deep learning model, pDeep [77], predicts a theoretical spectrum for each candidate, and (3) a ranking support vector machine ranks the candidate peptides, based on features extracted by comparing the observed and theoretical spectra. Finally, PointNovo [54] is an improved version of DeepNovo which focuses specifically on handling high-resolution mass spectrometry data by using an order-invariant network architecture [53].

2.3 Methods

Transformers are highly capable of learning contextualized representations and modeling sequential data [65], with a variety of successful applications to biological sequences [56, 3]. In this context, *de novo* peptide sequencing can be formulated as a sequence-to-sequence learning problem where variable-length sequences of observed spectra peaks are translated into variable-length sequences of amino acids. The main contribution of this paper is to propose a transformer-based

de novo peptide sequencing framework, Casanovo, which provides a unified solution to *de novo* peptide sequencing sub-tasks such as learning latent representations for spectra, spectrum processing and peptide sequence processing, which existing methods tackle separately through more complex modeling schemes.

2.3.1 Casanovo

Casanovo consists of a transformer encoder and decoder stack as described in [65], which are respectively responsible for learning latent representations of the input spectrum peaks and decoding the amino acid sequence of the spectrum's generating peptide (Figure 2.2). The encoder takes d-dimensional spectrum peak embeddings as input and outputs d-dimensional latent representation vectors for each peak. Subsequently, the decoder takes as input these representations of prefix amino acids, coupled with a d-dimensional precursor embedding encapsulating precursor *m/z* and charge information, to predict the next amino acid in the peptide sequence. We discuss different aspects of our modeling strategy in detail below.

Input embeddings

Each spectrum $S = \{(m_j, I_j)\}_{j=1}^{N}$ is a bag of peaks, where each peak (m_j, I_j) is a 2-tuple representing the *m/z* value and intensity of the peak. The *m/z* value and intensity are embedded separately before being summed to yield the input peak embedding. We use a fixed, sinusoidal embedding [65] to project each *m/z* value to a d-dimensional vector, the *m/z* embedding f. Specifically, we create the *m/z* embedding from an equal number of sine and cosine waveforms spanning the

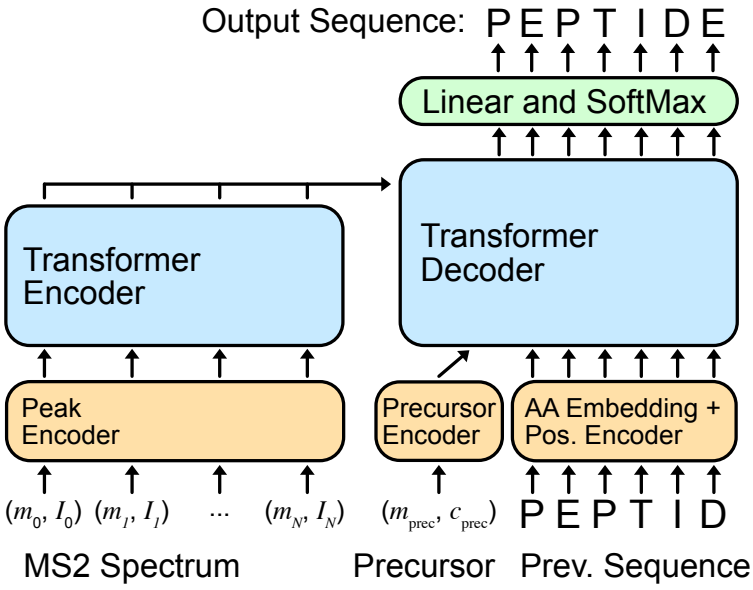


Figure 2.2: **Casanovo model architecture with inputs and outputs**

wavelengths from 0.001 to 10,000 *m/z*, where each feature in the embedding f_i is

$$f_i = \begin{cases} \sin(m_j / (\frac{\lambda_{\max}}{\lambda_{\min}} (\frac{\lambda_{\min}}{2\pi})^{2i/d})), & \text{for } i \leq d/2 \\ \cos(m_j / (\frac{\lambda_{\max}}{\lambda_{\min}} (\frac{\lambda_{\min}}{2\pi})^{2i/d})), & \text{for } i > d/2 \end{cases} \tag{2.1}$$

where $\lambda_{\max} = 10,000$ and $\lambda_{\min} = 0.001$. These input embeddings provide a granular representation of high-precision *m/z* information and, similar to relative positions in the original transformer model [65], may help the model attend to *m/z* differences between peaks, which are critical for identification of amino acids in the peptide sequence. The intensity, which is measured with lower precision than the *m/z* value, is embedded by projection to d dimensions through a linear layer, after which the *m/z* and intensity embeddings are summed to produce the input peak embedding. We also experiment in Section 2.4.3 with encoding intensity using a fixed, sinusoidal position embedding and concatenating it with the *m/z* embedding.

Precursor information, used as input to the decoder, consists of the total mass $m_{\text{prec}} \in R$

and charge state $c_{\text{prec}} \in \{1, ..., 10\}$ associated with the spectrum. We use the same sinusoidal position embedding as peak *m/z*'s for m_{prec}; c_{prec} is embedded using an embedding layer, and the embeddings are summed to obtain the input precursor embedding. Preceding amino acids in the peptide sequence, another decoder input, are also encoded as the sum of an amino acid embedding and a sinusoidal position embedding of their position in the sequence.

Training and inference strategy

Taking the previously described embeddings as input, the transformer outputs scores which are treated as a probability distribution over the amino acid vocabulary for the next position in the sequence at each decoding step. The amino acid vocabulary includes 20 canonical amino acids, post-translationally modified versions of three of them (oxidation of methionine and deamidation of asparagine or glutamine), plus a special `stop` token to signal the end of decoding, yielding a total of 24 tokens. During training, the decoder is fed the amino acid prefix for the ground truth peptide following the teacher forcing paradigm [69]. Cross-entropy between the model output probabilities and a binary matrix representing amino acid sequence of the ground truth peptide is minimized as the objective function. During inference, the highest scoring amino acid is predicted for each position in the sequence, and the decoder is fed its previous amino acid predictions at each decoding step. The decoding is finished either when the `stop` token is predicted or the pre-defined maximum peptide length of $\ell = 100$ amino acids is reached.

Model and training hyperparameters

We train models with nine layers, embedding size $d = 512$, and eight attention heads, yielding a total of ~47M model parameters. A batch size of 32 spectra and 10^{-5} weight decay is used during training, with a peak learning rate of 5×10^{-4}. The learning rate is linearly increased from zero to its peak value in 100k warm-up steps, followed by a cosine shaped decay. Models

are trained on 2 RTX 2080 GPUs for 30 epochs, which takes approximately two days, and model weights from the epoch with the lowest validation loss were selected for testing. These model hyperparameters—number of layers, embedding size, number of attention heads, and learning rate schedule—are used for all downstream experiments unless otherwise specified.

Precursor *m/z* filtering

A critical constraint in *de novo* peptide sequencing requires the relative difference between total mass of the predicted peptide m_{pred} and the observed precursor mass m_{prec} of the spectrum to be less than a threshold value ϵ (specified in ppm) for the predicted sequence to be plausible: $\Delta m_{ppm} = \frac{|m_{\text{prec}} - m_{\text{pred}}| \times 10^6}{m_{\text{prec}}} < \epsilon$ In addition to providing precursor information as an input for the model to learn from, we filter out peptide predictions that do not satisfy the above constraint. The threshold value ϵ is a property of the mass spectrometer that the data is collected with, and hence is known at inference time. Accordingly, we choose ϵ based on the precursor mass error tolerance used in the database search to obtain ground truth peptide sequences for the test data.

Casanovo's source code and trained model weights are available as open-source under the Apache 2.0 license at `https://github.com/Noble-Lab/casanovo`.

2.3.2 Data set

To evaluate the performance of Casanovo and compare it with state-of-the-art *de novo* peptide sequencing methods, we use the nine-species benchmark data set and evaluation framework first introduced by [63] and used in several subsequent studies [36, 54]. This data set combines a total of about 1.5 million mass spectra from nine different experiments, each using the same instrument to analyze peptides from a different species. Based on database search identification using the standard false discovery rate (FDR) of 1%, each spectrum comes with an assigned peptide sequence

which is treated as ground truth to train and evaluate the methods. With approximately 300,000 unique peptide sequences in the data set, each sequence has around five spectrum instances on average, but around 40% of all peptide sequences have a single spectrum associated with them. Following [63], we employ a leave-one-out cross validation framework where we train a model on eight species and test on the held-out species for each of the nine species in the data set. In each case, we split the training set 90/10 for training and validation. This cross-species evaluation framework allows for testing the model on never-before-seen peptide samples, because the peptides in the training set are almost completely disjoint from the peptides of the held-out species. To illustrate this point, among the ~26,000 unique peptide labels associated with the human spectra in the test data, only 7% overlap with the ~250,000 unique peptide labels associated with spectra from the other eight species. Cross-species testing is particularly important for *de novo* sequencing models because most practical applications of *de novo* sequencing require models to perform well on spectra with never-before-seen peptide sequences.

2.3.3 Evaluation metrics

We use precision calculated at the amino acid and peptide levels [44, 26, 63] as a function of coverage over the test set as performance measures to evaluate the quality of a given model's predictions. In each case, for each spectrum we compare the predicted sequence to the ground truth peptide from the database search. Following [63], for the amino acid-level measures we first calculate the number N^a_{match} of matched amino acid predictions, defined as all predicted amino acids which (1) differ by <0.1 Da in mass from the corresponding ground truth amino acid, and (2) have either a prefix or suffix that differs by no more than 0.5 Da in mass from the corresponding amino acid sequence in the ground truth peptide. We then define amino acid-level precision as $N^a_{\text{match}}/N^a_{\text{pred}}$, where N^a_{pred} is the number of predicted amino acids. For peptide

Table 2.2: **Empirical comparison of Casanovo, DeepNovo and PointNovo.** The table lists the peptide-level and amino acid-level precision of three competing models and coverage of Casanovo with precursor *m/z* filtering on all nine benchmark cross-validation folds. Each fold's test set contains spectra from a single species, with nearly disjoint sets of peptides between species. For cross-validation folds corresponding to mouse and human, five models were trained with different random initializations. For these species, we report standard deviation of the performance measures.

	Peptide-level performance					Amino acid-level performance			
	DeepNovo	PointNovo	Casanovo			DeepNovo	PointNovo	Casanovo	
Species	Prec.	Prec.	Prec.	Cov.	Prec. at Cov.=1	Prec.	Prec.	Prec.	Prec at Cov.=1
Mouse	0.286	0.355	0.665±0.015	0.666±0.013	0.443±0.019	0.623	0.626	0.899±0.018	0.562±0.021
Human	0.293	0.351	0.683±0.014	0.537±0.015	0.367±0.017	0.610	0.606	0.898±0.015	0.424±0.019
Yeast	0.462	0.534	0.824	0.681	0.561	0.750	0.779	0.952	0.591
M. mazei	0.422	0.478	0.771	0.630	0.486	0.694	0.712	0.935	0.518
Honeybee	0.330	0.396	0.732	0.557	0.408	0.630	0.644	0.920	0.461
Tomato	0.454	0.513	0.771	0.557	0.460	0.731	0.733	0.929	0.471
Rice bean	0.436	0.511	0.798	0.547	0.437	0.679	0.730	0.920	0.442
Bacillus	0.449	0.518	0.805	0.671	0.540	0.742	0.768	0.943	0.573
Clam bacteria	0.253	0.298	0.695	0.534	0.371	0.602	0.589	0.908	0.405

predictions, a predicted peptide is considered a correct match if all of its amino acids are matched. Among a collection of N^p_{orig} spectra, if our model makes predictions on a subset of N^p_{pred} and correctly predicts N^p_{match} peptides, we define coverage as $N^p_{\text{pred}}/N^p_{\text{orig}}$ and peptide-level precision as $N^p_{\text{match}}/N^p_{\text{pred}}$. To plot a precision-coverage curve, we sort predictions by the confidence score provided by the model. Amino acid-level confidence scores are obtained by applying a softmax to the output of the transformer decoder, which is a proxy for the probability of each predicted amino acid to occur in the given position along the peptide sequence. Casanovo directly outputs amino acid-level confidence scores, and we use the mean score over all amino acids as a peptide-level confidence score.

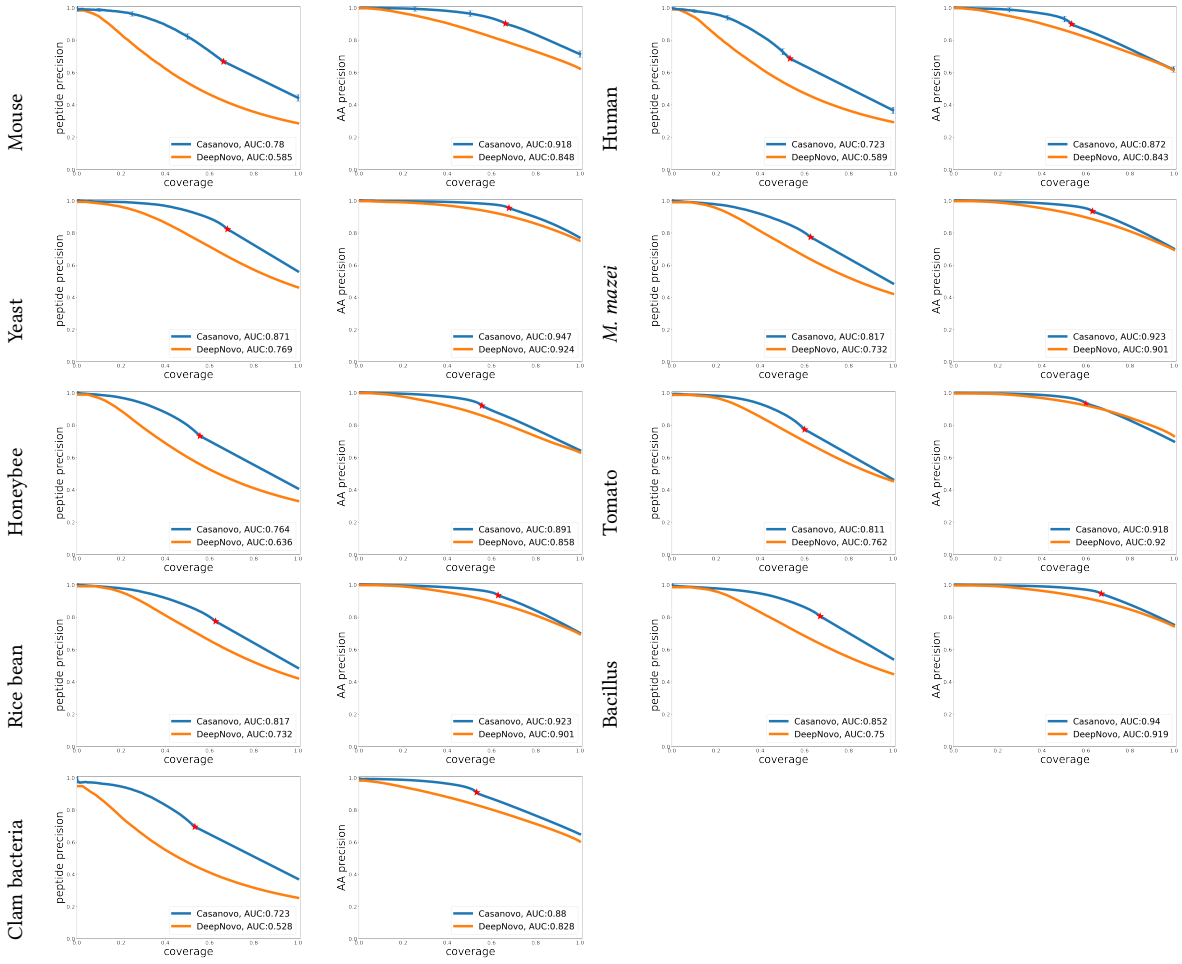


Figure 2.3: **Precision-coverage curves for Casanovo and DeepNovo.** Curves are shown for each species (two per row) at the peptide level (left sub-column) and amino acid level (right sub-column). Curves are computed by sorting predicted peptides according to their confidence scores. For the amino acid level curves, all amino acids within a given peptide receive equal scores. For Casanovo at both amino acid and peptide level, all peptides that pass the precursor *m/z* filtering are ranked above peptides that do not pass the filter, and similarly for all amino acids from peptides that pass the precursor *m/z* filtering versus those that do not pass the filter. The boundary between unfiltered and filtered entries is indicated by a red star on each curve. Error bars are provided for human and mouse species, for which five models were trained with different random initializations.

2.4 Results

2.4.1 Casanovo outperforms state-of-the-art methods

We begin by using the previously described experimental setup and evaluation metrics (Sections 2.3.2–2.3.3) to evaluate Casanovo's performance relative to two state-of-the-art neural network-based methods, DeepNovo [63] and PointNovo [54]. In this comparison, peptide-level performance measures are the primary quantifier of the sequencing model's practical utility, since the goal is to assign a complete peptide sequence to each observed spectrum. To characterize the performance of DeepNovo and PointNovo, we rely on the pre-trained weights of the former and the published results of the latter [54], since neither PointNovo's pre-trained weights nor its predictions for the benchmark data set are available.

At the peptide level, Casanovo substantially outperforms both previous methods across all species, with a mean improvement of 0.373 and 0.310 in precision relative to DeepNovo and PointNovo, respectively, at a mean coverage of 0.60 (Table 2.2). Even when precursor *m/z* filtering is turned off and the model is forced to make a prediction for all spectra, i.e. coverage is 1.0, Casanovo shows a mean improvement of 0.076 and 0.013 relative to DeepNovo and PointNovo, respectively. Indeed, the peptide-level precision-coverage curves (Figure 2.3) show that Casanovo consistently outperforms DeepNovo over a range of peptide confidence thresholds. This trend is also reflected by the area under the curve (AUC) metric (Figure 2.3), with Casanovo outperforming DeepNovo by 0.131 on average.

Similarly, at the amino acid level, Casanovo outperforms DeepNovo and PointNovo, particularly in the high-precision portions of the curves. As expected, the precursor *m/z* filtering, which prioritizes predicting full peptide sequences with high precision over partially correct peptide predictions, yields better overall precision at the cost of reduced precision at full coverage. In all

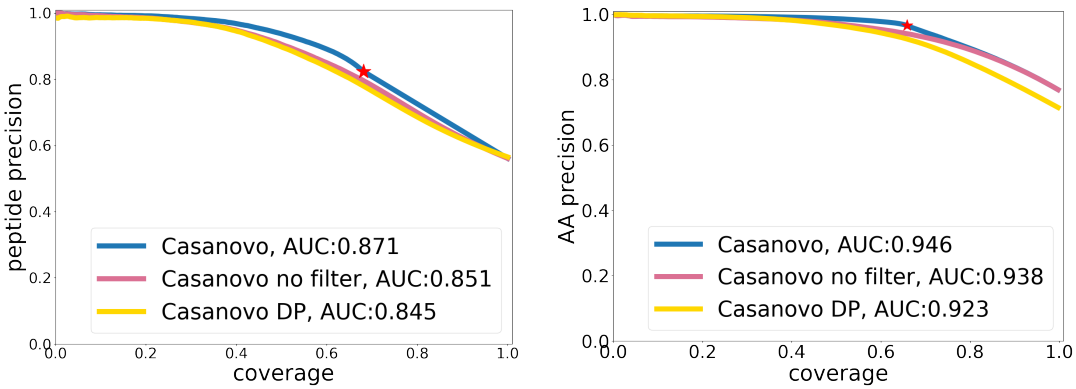


Figure 2.4: **Precision-coverage curves for Casanovo models with different post-processors** Standard Casanovo model with simple filter outperforms both no filter and dynamic programming post-processor on the yeast test set, where we see that the effect of the filter is to boost precision along the entire curve.

nine species, the point on the Casanovo curve corresponding to the filter lies above the DeepNovo precision-coverage curve, and in eight of the nine species Casanovo's AUC exceeds DeepNovo's. We further discuss the effects of precursor *m/z* filtering in Section 2.4.2.

Complementing its improved *de novo* peptide sequencing performance, Casanovo achieves these results with fewer model parameters (47 M) than DeepNovo (86 M). (The number of parameters and model dimensions were not reported for PointNovo.) Casanovo also runs inference at a faster rate of 119 spectra/s on an RTX 2080 compared to DeepNovo's 36 spectra/s and PointNovo's reported 20 spectra/s on an RTX 2080 Ti (a comparatively faster GPU) [54].

2.4.2 Precursor *m/z* post-processing

One of the key components of DeepNovo, and its successor PointNovo, is a post-processor that uses the knapsack dynamic programming algorithm to ensure that the mass of the predicted peptide is close to the observed precursor mass. Ablation experiments in the DeepNovo experiment paper show that removing this component leads to a decrease in peptide-level precision of 12.4% (averaged

over test sets) [63]. Accordingly, we tested three variants of Casanovo on the yeast species test set: no post-processor, the knapsack post-processor, and our simple *m/z* filter (Figure 2.4).

At the peptide level, we observe a much smaller benefit from the knapsack algorithm—an increase in the peptide-level precision from 0.561 to 0.565 when models are compared at full coverage—than was reported in the DeepNovo paper. On the other hand, a comparison of precision-coverage curves indicates that the knapsack algorithm hurts the AUC metric and precision at most coverage values.

At the amino acid level, the effects of these two post-processors is different. Relative to Casanovo with no post-processing, applying the knapsack algorithm yields a small decrease in precision at full coverage (0.769 $\rightarrow$ 0.726), whereas adding the precursor *m/z* filtering yields much higher precision (0.769 $\rightarrow$ 0.965) which decreases substantially when the coverage is extended to all spectra (0.769 $\rightarrow$ 0.576). Similar to the peptide level, the standard Casanovo model with the simple filter consistently yields the highest precision for different values of coverage as well as the largest AUC.

To better understand these results, we performed a qualitative review of the predictions from the three models. This analysis suggests that incorrect amino acid predictions in earlier decoding steps cause the post-processor to discard correct amino acids from among options in later decoding steps, leading to a drop in amino acid-level performance. This observation is supported by the plot of the precision-coverage curve with and without the precursor *m/z* filter (Figure 2.4), where we see that the effect of the filter is to boost precision along the entire curve.

2.4.3 Peak embeddings and loss function

Finally, we train three additional variants of Casanovo, none of which provides any performance improvements over the standard model (Table 2.3). The first variant uses a focal loss function,

Model variant	Peptide Prec.	Peptide Prec. at Cov.=1	Amino acid Prec.	Amino acid Prec. at Cov.=1
Standard Casanovo	0.851	0.561	0.965	0.576
Focal loss	0.802	0.532	0.938	0.543
I x m/z embedding	0.736	0.446	0.841	0.463
Sinusoidal I embedding	0.817	0.538	0.943	0.552

Table 2.3: **Performance comparison of different Casanovo variants.** All results are for the yeast test set.

adopted from [54], instead of cross entropy. We also investigated two alternate methods of peak embedding. The first, also adopted from PointNovo, replaces summation of m/z and I embeddings with direct multiplication of the I value and the m/z embedding. The second peak embedding strategy implements a sinusoidal encoding for I, similar to the m/z embedding, although using only 32 dimensions, and concatenates I with the m/z embeddings instead of summing the two.

2.5 Discussion

Prior work in *de novo* peptide sequencing has used deep learning models that combine separate neural network architectures followed by complex post-processing steps. Our approach, Casanovo, leverages the transformer architecture to produce a unified solution to translate mass spectra directly into peptide sequences, without resorting to discretization of the spectrum m/z axis and without complex post-processing. We find that Casanovo achieves state-of-the-art performance on the standard benchmark data set, with fewer model parameters compared to existing methods.

Casanovo's inference speed is fast enough to allow real time *de novo* sequencing, i.e., sequencing at the speed that the mass spectrometer generates spectra, raising the possibility of helping guide mass spectrometry experiments as they are being conducted [43]. In practice, real-time search results can be useful for making decisions about peptide elution order [4], improving the accuracy of stable isotope labeling [5], post-translational modification site localization [5], or deciding

whether to trigger an MS3 (secondary fragmentation) scan [57].

Casanovo improves substantially over the previous state of the art in terms of peptide-level precision, but this leaves a significant portion of the test spectra without plausible predictions. Clearly, exploring methods to find good predictions for these spectra is an avenue for future research. To explore the potential benefit of such an approach, we combined Casanovo and DeepNovo predictions by inserting DeepNovo predictions whenever the *m/z* filter eliminates a Casanovo prediction. The resulting model achieves up to 10% higher peptide precision than Casanovo and exceeds the previous state-of-the-art method, PointNovo, on all evaluation metrics across species. This observation suggests that precursor information should be included as a stronger prior in modeling mass spectra. A straightforward approach might involve choosing among a larger set of peptide candidates generated by beam search during inference, with a constraint on the predicted mass. Alternatively, Casanovo's loss function could be modified to penalize peptide predictions which do not match the precursor mass.

Chapter 3

Large scale training and *de novo* peptide sequencing applications with Casanovo

This chapter is adapted with minimal modification from:
Melih Yilmaz, William Fondrie, Wout Bittremieux, Carlo F Melendez, Rowan Nelson, Varun Ananth, Sewoong Oh, and William S Noble Sequence-to-sequence translation from mass spectra to peptides with a transformer model. *Nature communications* 15.1:6427, 2024.

3.1 Introduction

Mass spectrometry is currently the most popular analytical technique to characterize the proteome, by identifying and quantifying proteins present in complex biological systems [1]. During a bottom-up mass spectrometry proteomics experiment, proteins from a biological sample are enzymatically digested into peptides, their intact mass and charge are measured, and they are fragmented using tandem mass spectrometry. The fundamental challenge of mass spectrometry proteomics is then to determine the amino acid sequence of the resulting tandem mass (MS/MS) spectra. The standard

approach to solve this spectrum identification problem is sequence database searching, during which peptides are simulated *in silico* using known digestion rules from a database of protein sequences potentially present in the biological samples, typically from a reference proteome. Next, each observed MS/MS spectrum is scored against a list of candidate peptides based on simplified peptide fragmentation rules, and the best-scoring peptide-spectrum match (PSM) is reported. Pioneered by the SEQUEST algorithm [21], dozens of database search engines have been subsequently developed and are very widely deployed [22].

However, a fundamental requirement for sequence database searching is that the set of proteins that may be present in the sample is known in advance. While this is often the case for samples generated from species with well-characterized genomes, relying on a database prevents the detection of unexpected peptides. Such unexpected peptides include not just peptides derived from contaminant proteins or that arise due to variability in sample processing [11], but also biologically and clinically relevant peptides, such as peptides that deviate from the reference proteome due to genetic variation, peptides with unexpected post-translational modifications (PTMs), and peptides originating from foreign sources, such as microbes or consumed foods. Furthermore, there are tasks where generating a peptide database can prove impractical or even impossible. For example, the antigenic peptides presented by major histocompatibility complex (MHC) proteins—the "immunopeptidome"—are often generated from their parent proteins in an unpredictable manner, requiring at minimum every possible protein subsequence to be considered [59, 47, 30].

Similarly, constructing a peptide database for antibody sequencing is nearly impossible, due to the sequence variants created by V(D)J recombination [63]. Finally, creating an accurate database for mixtures of many organisms—metaproteomics—such as from microbiome or environmental samples, is often not feasible [48].

Such applications require the ability to sequence peptides directly from the acquired MS/MS

spectra *de novo*. Early *de novo* peptide sequencing algorithms used heuristic search [61] or dynamic programming [44, 16] to propose peptides for the observed MS/MS spectra. In addition to dynamic programming, the PepNovo algorithm [26] attempted to account for rules governing peptide fragmentation in its probabilistic score function and is closely related to the hidden Markov model employed by Fischer *et al.* [24] In 2015, the Novor algorithm [43] improved the state of the art by using a decision tree as the score function for its dynamic programming algorithm.

More recently, as in many other fields, deep learning has become the preferred solution for *de novo* peptide sequencing. DeepNovo [63] combines a convolutional neural network and a recurrent neural network to predict the subsequent amino acid when provided an MS/MS spectrum and a peptide prefix. SMSNet [36] uses a similar network architecture but reconciles the predicted sequences against a user-supplied peptide database. PointNovo [54], the successor to DeepNovo, leverages an order-invariant network architecture to model high-resolution MS/MS spectra [53]. Finally, pNovo 3 [72] first generates candidate peptides for each MS/MS spectrum using dynamic programming, after which a final score is determined by matching the spectrum against a theoretical spectrum for each candidate peptide, simulated using the pDeep [77] learning-to-rank framework.

Despite the advances of these deep learning-based methods for *de novo* peptide sequencing, they still suffer from several limitations. *De novo* tools typically can only annotate a minority of MS/MS spectra compared to sequence database searching, they struggle with natively encoding high-resolution MS/MS data, and they employ complex neural network architectures and post-processing steps.

To address these issues, here we describe Casanovo, which reframes the *de novo* peptide sequencing task as a machine translation problem: like translating a sequence of words in a sentence from one language to another, Casanovo translates a sequence of peaks in an MS/MS spectrum into a sequence of amino acids of the generating peptide. To do so, we leverage the

state-of-the-art architecture for modeling natural language—the transformer [65]. The transformer architecture allows Casanovo to directly use the *m/z* and intensity value pairs that comprise an MS/MS spectrum without discretization of the *m/z* axis and to directly output a predicted peptide sequence without a complicated dynamic programming step. We have previously[76] trained Casanovo on a limited collection of mass spectra from the multi-species benchmark used by Tran *et al.* [63] In this work, we present significant improvements to Casanovo and demonstrate its effectiveness at tackling common challenges with *de novo* peptide sequencing. We expand our training set to use 30 million confident PSMs from the MassIVE-KB spectral library [66], and we add a beam search decoding procedure to predict the best peptide for each MS/MS spectrum. We demonstrate that, together, these updates significantly improve Casanovo's already state-of-the-art performance. Additionally, we fine-tune a non-enzymatic version of Casanovo for tasks such as immunopeptidomics. We demonstrate how high-performance *de novo* peptide sequencing using Casanovo enables fast and effective immunopeptidome analysis, bolsters the characterization of metaproteomes, and sheds light on the dark proteome.

3.2 Results

3.2.1 A transformer architecture enables processing of raw mass spectra

Casanovo uses a transformer architecture to perform a sequence-to-sequence modeling task, from MS/MS spectrum to the generating peptide (Figure 3.1). Transformers are built upon the attention function [65], which allows transformer models to contextualize the elements of a sequence; transformer models thus learn the relationships of sequence elements to one another and how their interaction should be interpreted. As such, the transformer architecture has found success in not only natural language processing, but also applications to biological sequences [56, 3].

In Casanovo, each peak in an observed MS/MS spectrum is considered as an element in a variable length sequence. The *m/z* and intensity values of each peak are encoded using, respectively, a collection of sinusoidal functions and a learned linear layer, and these encodings are summed. The encoded peaks are then input into the transformer encoder, where context is learned between pairs of peaks in the MS/MS spectrum. The contextualized peak encodings are then used as input to the transformer decoder for predicting the peptide sequence.

The process of decoding proceeds in an iterative, autoregressive manner. We begin by providing the mass and charge of the observed precursor. The transformer decoder uses the contextualized peak encodings and the precursor information to begin predicting amino acids of the peptide. With the first predicted amino acid, we retain the top k residues, where k is a user-selected value for the number of beams in our beam search. In each subsequent iteration, amino acids are added to the decoded peptide sequence, retaining the top k sequences until the decoded sequences for all of the beams have terminated or exceeded the precursor mass. Finally, the sequence with the highest score is retained as the putative peptide that generated the provided MS/MS spectrum.

In generating its predictions, Casanovo will inevitably fail to generate plausible peptides for some MS/MS spectra. For example, some MS/MS spectra contain too few fragment ions to be sequenced reliably, or the true generating peptide may bear a modification that is unknown to Casanovo. We therefore refine the PSMs proposed by Casanovo using a simple precursor mass filter: any PSMs for which the *m/z* of the peptide falls outside the specified tolerance of the observed precursor, including potential isotopes, is discarded. This filter eliminates many poorly scored peptides from consideration. In our evaluations, PSMs that would normally be removed by this filter were retained and ranked last among all PSMs assigned by Casanovo.

3.2.2 Casanovo outperforms state-of-the-art methods

To evaluate Casanovo, we first used the nine-species benchmark dataset originally created by Tran *et al.* [63] to compare the performance of four *de novo* peptide sequencing algorithms: Novor, DeepNovo, PointNovo, and Casanovo. For these comparisons, we used the publicly available, pre-trained version of Novor to sequence the MS/MS spectra in the benchmark dataset. DeepNovo, PointNovo and Casanovo were trained in a cross-validated fashion, systematically training on eight species and testing on the remaining species. For DeepNovo, we used the models trained and provided by Tran *et al.* [63] for each of the cross-validation splits. For PointNovo, we cross-validated nine models from scratch using the code and settings provided by Qiao *et al.* [54] . This benchmark version of Casanovo, Casanovo$_{bm}$, employs a simple greedy decoding algorithm, rather than beam-search decoding. The results (Figure 3.2A) revealed that Casanovo$_{bm}$ substantially improved peptide-level sequencing performance over Novor, DeepNovo and PointNovo, with an average precision of 0.81 for Casaonovo$_{bm}$ compared to 0.58, 0.70 and 0.74 for Novor, DeepNovo and PointNovo, respectively. These results are consistent across all nine species in the benchmark dataset (Supplementary Figure A.1).

We hypothesized that Casanovo could achieve even better performance if provided with a larger training set of higher quality PSMs; hence, we turned to the MassIVE-KB spectral library of human MS/MS proteomics data [66]. MassIVE-KB provided us with a set of 30 million high confidence PSMs, which we previously collected for training our GLEAMS embedding model [9]. This dataset contains not only a greater diversity of peptides and MS/MS spectra generated from multiple instruments, but also additional types of post-translational modifications. We therefore created a new version of the nine-species benchmark dataset using the same nine PRIDE datasets but including seven different types of variable modifications (methionine oxidation, asparagine deamidation, glutamine deamidation, N-terminal acetylation, N-terminal carbamylation,

N-terminal NH_3 loss, and the combination of N-terminal carbamylation and NH_3 loss). In the process, we also fixed several problems that we uncovered in the previous benchmark, including adding consideration of isotope errors and eliminating peptides that occur in multiple species (see Methods for details). The final, revised benchmark dataset consists of 2.8 million PSMs drawn from 343 RAW files.

The results from evaluating with respect to this revised benchmark demonstrate the value of training from a much larger collection of higher quality PSMs (Figure 3.2B). When trained on the MassIVE-KB dataset, the average precision of Casanovo increases from 0.83 to 0.95. Furthermore, Casanovo succeeds in making a larger proportion of predictions with *m/z* values that fall within 30 ppm of the observed precursor (signified by the location of the diamonds in Figure 3.2B), increasing from 70% to 97%. Additionally, an analysis of spectrum identifications for all *de novo* sequencing tools on the nine-species benchmark dataset shows that correct Casanovo PSMs include almost all correct identifications of the competing *de novo* sequencing methods, as well as approximately 50% more correct PSMs that are unique to Casanovo (Supplementary Figure A.8). This version of Casanovo incorporates beam-search decoding, which improves both average precision and coverage compared to greedy decoding for the same model (Supplementary Figure A.3).

In one sense, this comparison is unfair because some of the spectra in the new version of the benchmark contain PTMs that cannot be identified by some of the competing methods. We therefore eliminated these spectra from each test set and then re-computed the precision-coverage curve. The results (Supplementary Figure A.4) are largely unchanged, suggesting that the PTMs contribute little to the observed overall differences in performance.

To better understand why the model trained on MassIVE-KB outperforms the one trained on the 9-species benchmark, we performed two follow-up experiments. First, we trained a series of Casanovo models on randomly sampled nested subsets of MassIVE-KB, ranging from 250,000

spectra to the full dataset of 28 million spectra. Each model was then evaluated with respect to the revised 9-species benchmark. The resulting learning curve (Supplementary Figure A.5) shows that the test set performance depends strongly on the size of the training set, though with diminishing returns after a million or so PSMs. Second, we directly compared a Casanovo model trained from a downsampled MassIVE-KB dataset to Casanovo$_{bm}$ which averages 9 models cross-validated on the 9-species benchmark, where the training sets contain approximately the same number of peptides (239,697 for Massive-KB and 246,713 for Casanovo$_{bm}$). We then evaluated both models using the revised 9-species benchmark. The results (Supplementary Figure A.6) show that the model trained from MassIVE-KB substantially outperforms Casanovo$_{bm}$, with the average precision increasing from 0.83 to 0.90. Thus, these results suggest that the improved performance of the MassIVE-KB model stems primarily from the improved quality of the data rather than the size of the data set.

In addition to evaluating Casanovo's ability to correctly predict whole peptides, we also evaluated Casanovo's ability to the predict individual amino acids of each peptide. We did so by ranking amino acids by their associated confidence score and then plotting a precision-coverage curve. We compared two versions of Casanovo (trained from the first benchmark and from MassIVE-KB) with DeepNovo and PointNovo on the revised nine-species benchmark with new modifications eliminated (Figure 3.2C). The amino acid-level performance was consistent with the trends we observed in peptide-level performance, with Casanovo outperforming Novor, DeepNovo and PointNovo: Casanovo trained on MassIVE-KB achieves a remarkable average precision of 0.98.

Finally, to characterize the improved *de novo* sequencing performance of Casanovo across generating peptides of different lengths and precursor charge states, we compare all methods on subsets of the revised nine-species benchmark dataset. First, we divide spectra into 3 groups by charge state where groups contain precursors with 2+, 3+ and 4+ or higher charge states each, and plot peptide precision-coverage curves for each group (Supplementary Figure A.9).

As expected, average precision is lower across all methods for groups with higher precursor charge states since those spectra tend to have more complex fragmentation patterns. However, the drop in performance is only 12% for Casanovo in precursors with 4+ or higher charge states versus precursors with 2+ charge states, thanks to the diversity of precursor charge states in its training data where 11% of precursors have 4+ or higher, whereas average precision for all competing methods decreases by more than 60%. Second, we bin spectra according to the length of their generating peptides into groups of short (fewer than 13 amino acids), medium (between 13 and 18 amino acids), and long (greater than 18 amino acids) peptides, and compare *de novo* sequencing performance in each group (Supplementary Figure A.10). Performance degrades for longer peptides because incorrect amino acid predictions tend to accumulate during decoding, but the observed decrease in average precision for Casanovo is much smaller relative to other methods, highlighting Casanovo's ability to accurately sequence long peptides as a key contributor to its improved performance.

3.2.3 Casanovo unravels the immunopeptidome

One important application of *de novo* peptide sequencing is the characterization of the peptides presented by major histocompabilitility complexes (MHCs), which is commonly referred to as the "immunopeptidome." These antigen peptides are presented on the extracellular surface and serve as targets for immune cell recognition. However, because these antigen peptides are generated through lysosomal or proteasomal degradation, they do not exhibit the characteristic tryptic termini from most proteomics experiments. Consequently, the peptide search space is exponentially larger than considering only tryptic peptides—every peptide subsequence in a protein within a defined peptide length must be considered. Furthermore, mutations in these peptides are of particular interest, because these mutation-containing neoantigens may serve as

tumor-specific markers to activate T cells and initiate antitumor immune responses. Unfortunately, expanding the search space to consider all possible mutated peptides is prohibitive both in terms of search speed and statistical power for traditional proteomics search engines.

Although immunopeptidomics is a prime application for *de novo* sequencing, naively applying Casanovo directly to immunopeptidomics data is problematic: the standard Casanovo model is heavily biased to predict tryptic peptides due to their overrepresentation in MassIVE-KB. To demonstrate this effect, we analyzed five mass spectrometry runs generated from MHC class I peptides isolated from MDA-MB-231 breast cancer cells [59] in two different ways: first using Casanovo and second by searching against a non-enzymatic digestion of the human proteome (see Methods). Among the peptides accepted at 1% FDR by the database search procedure, we observed a low proportion of "tryptic" peptides, i.e., peptides with C-terminal amino acids of K (1.12%) or R (0.80%). In contrast, among the top-scoring 10% of the Casanovo predictions, we observed a greater than six-fold increase in the rate of tryptic peptide predictions (5.87% K and 6.76% R).

We hypothesized that we could reduce this tryptic bias and produce a version of Casanovo that is better suited to immunopeptidomics data by fine tuning our existing model using data that lacks a tryptic bias. To create such a dataset, we combined data from two sources. First, we segregated PSMs from MassIVE-KB according to their C-terminal amino acid and then uniformly sampled up to 50,000 peptides within each group. For most amino acids, MassIVE-KB contained fewer than 50,000 PSMs, so for these we supplemented by randomly extracting additional PSMs from the PROSPECT collection [58] (Supplementary Table A.1). We then split this new collection of 1 million PSMs into training, validation, and testing sets. We then fine-tuned our existing Casanovo model by training it until convergence on this non-enzymatic training set.

The resulting model, Casanovo$_{ne}$, performs markedly better than the original Casanovo model at predicting peptides in our held-out, non-enzymatic test set. On the held-out test set of 100,000 non-enzymatic PSMs, Casanovo$_{ne}$ achieves an average precision of 0.83, compared with 0.60 for

the original Casanovo model on the same data (Figure 3.3A). The predicted C-terminal amino acid frequencies are also much closer to the true frequencies, with K and R dropping to 1.81% and 1.79%, respectively (Figure 3.3B).

We next used Casanovo$_{ne}$ to sequence the immunopeptidome of MDA-MB-231 breast cancer cells [59]. For each peptide predicted by Casanovo, we investigated whether it (1) occurs anywhere within the human proteome, and (2) occurs within the set of peptides detected using a database search procedure. We first searched the data against a non-enzymatic digestion of the human proteome using the Tide search engine [18] followed by Percolator post-processing [35], using settings similar to those in the original study [59] (see Methods). Out of 26,377 unique peptides predicted by Casanovo, 2459 match to the human proteome, and a majority of these overlap with the 1544 unique peptides identified by Tide at 1% FDR (Supplementary Figure A.11). Notably, these overlapping peptides are predicted with high confidence by Casanovo, almost all within the first 10,000 Casanovo predictions (Figure 3.3C). Casanovo predicts an additional 1148 peptides that match to the human proteome but are not identified by Tide at 1% FDR, and further analysis shows that 751 (65.4%) of these peptides correspond to Tide hits that were not accepted at the 1% FDR threshold.

To further investigate the plausibility of Casanovo predicted peptides as MHC antigens, we used NetMHCpan-4.1 [55] to predict MHC binding affinity for these peptides. First, we compared peptides that were identified by both Casanovo and database search with peptides that were predicted only by Casanovo and match to the human proteome. These two groups exhibit similar distributions of predicted binding affinity profiles, with 87% of peptides identified by Casanovo alone and 86% of those identified by both methods predicted to be MHC binders at 500 nM (Figure 3.3D). In contrast, when we evaluate peptides that are identified by database search but not by Casanovo, the proportion of predicted MHC binders drops substantially to 50%. Overall, these results suggest that Casanovo not only identifies more peptides matching to the human

proteome than the standard database search procedure, but the peptides Casanovo predicts are also more likely to bind MHC antigens.

We also explored an alternative method for comparing Casanovo and Tide results, which does not rely on mapping Casanovo predictions to the reference proteome. For this analysis, we consider the 1497 peptides identified by Tide at 1% FDR which yielded valid binding affinity predictions from NetMHCpan alongside the top 1497 highest confidence Casanovo predictions. The results (Figure 3.3E) agree with those in Figure 3.3D: the 960 peptides in common between the two sets of peptides achieve the highest proportion of MHC binders (86%), the Casanovo-only predictions achieve a slightly lower percentage (80%), and the Tide-only predictions have the lowest percentage of MHC binders (70%).

3.2.4 Casanovo accurately sequences peptides from complex metaproteomes

Proteomics applications extend far beyond the analysis of single model organisms or well-characterized biological systems. Indeed, there is growing interest in using mass spectrometry proteomics methods to investigate the dynamics of complex biological ecosystems—whether microbiomes or environmental specimens—for which the identities of its members cannot be known *a priori*. Due to the unknown complexity of the sample and even the lack of reliable reference proteomes for the likely species in the sample, these metaproteomics experiments are difficult to analyze. One solution to these problems is to search the spectra against a large database, such as one containing all the microbial sequences in public databases for a sample that is likely dominated by unknown microbes. This "big database" approach is widely used but suffers from a significant loss in statistical power due to the implicit multiple hypothesis testing correction that must be made to account for the size of the database. An alternative solution involves first

subjecting the sample to genome sequencing, and then using the inferred peptide sequences as the basis for a "metapeptide" database. This approach yields better power to detect peptides [46] but requires the availability of a matched DNA sample and the additional cost associated with DNA sequencing.

We hypothesized that Casanovo's improved *de novo* sequencing capabilities would be useful in both scenarios—either in the presence or absence of a metapeptide database. To test this hypothesis, we applied Casanovo to data from six previously published ocean metaproteomics samples, three from the Bering Strait and three from the Chukchi Sea [46]. Critically, these samples were also subjected to DNA sequencing; hence, in addition to the non-redundant environmental database, we also have a metapeptide database for each sampling location.

We began by measuring the extent to which peptides detected by Casanovo occur within the corresponding metapeptide database or within the larger, non-redundant protein database. Because these samples were digested using trypsin, we used the standard Casanovo model, trained from the tryptic MassIVE-KB dataset. To control the error rate for the matching of Casanovo predictions to these databases, we employed a procedure similar to target-decoy competition used in the false discovery rate estimation for database search (see Methods for details) and only considered as correct Casanovo peptides found in the corresponding database that fall within the 1% random matching threshold. Using this logic, we obtain much better power to detect peptides using Casanovo than using a standard database search procedure against the metapeptide database (Figure 3.4A–B). In particular, when we search the data against the metapeptide database using Tide [18] followed by Percolator [35], we detect 5623 peptides at 1% FDR in the Bering Strait data and 2460 peptides in the Chukchi Sea data. In contrast, if we run Casanovo and accept as correct only peptides that appear in the metapeptide database (subject to our 1% random matching criterion), then we detect 8277 and 3532 peptides, respectively, in the two datasets, representing increases of 47% and 44%. Casanovo also outperforms database search when we consider the

non-redundant protein database rather than the sample-specific metapeptide database. We detect 1364 peptides in the Bering Strait data and 682 peptides in the Chukchi Sea data at 1% FDR by searching the non-redundant environmental database using Tide and Percolator. In comparison, Casanovo predictions, filtered at 1% error rate using the environmental database, detect 3425 and 1612 peptides, respectively, representing increases of 151% and 136%, respectively.

When using both metapeptide databases or the non-redundant environment database, Casanovo detects most of the peptides identified by Tide database search and Percolator, where it respectively detects 71% and 75% of Tide identifications on metapeptide and non-redundant databases, while also detecting a substantial number of additional unique peptides (Supplementary Figure A.12).

To validate the peptides that were detected by Casanovo but not the database search, we used the Prosit machine learning tool [27] to predict spectrum peak intensities and retention times for peptide identifications. First, we compared the cosine similarities between the observed and predicted MS/MS spectrum peak intensities across three groups of peptides: peptides only predicted by Casanovo that matched to the database with 1% error, peptides detected both by Casanovo and by Tide and Percolator at 1% FDR, and a control group of peptides detected by Tide and Percolator with >10% FDR. The control group was randomly sampled to be the same size as the Casanovo-only group. The results (Figure 3.4C) indicate that the Casanovo-only identifications have a high concentration of high cosine similarity peptides, similar to the overlapping identifications between Casanovo and database search. This stands in contrast with the control group, which exhibits a much broader distribution of cosine similarities.

Second, we compared the observed retention times with the predicted retention times from Prosit for the same three groups of peptides. For each group, we calibrated the predicted retentions times to the observed retention times using linear regression (Supplementary Figure A.13). We observed that the peptides detected only by Casanovo and those detected by Casanovo and Tide had a similar slope and resulted in similar residual distributions (Figure 3.4D). When compared

against the control group, the residual distributions for peptides only detected by Casanovo and those detected by Casanovo and Tide are close to zero.

Ultimately, Casanovo does not yet allow us to achieve as much power with the non-redundant database as with the metapeptide database. For example, for the Bering Strait data, the union of the 3425 peptides detected using Casanovo and the 1364 peptides detected using database search is 3750, which is fewer than the 5623 peptides detected using the metapeptide database. (The corresponding numbers for the Chukchi Sea data are 1798 and 2460.) This difference is perhaps not surprising, because the environmental non-redundant database is incomplete: 3715 of the 5623 peptides found by the database search procedure in the Bering Strait metapeptide database are not even present in the environmental database. Thus, a rigorous FDR control procedure for *de novo* peptide sequencing is needed in order to rescue the many peptides that are correctly detected by Casanovo but cannot be validated by matching to a database.

3.2.5 Casanovo shines a light on the dark proteome

The "dark matter" of mass spectrometry-based proteomics consists of MS/MS spectra that are observed repeatedly across experiments but consistently fail to be identified. In many cases, these MS/MS spectra may have been generated by peptides that are not in the canonical human proteome, because they represent contaminant peptides, result from non-standard enzymatic cleavage, or contain sequence variants. We hypothesized that Casanovo could shed light on some of this dark matter.

Accordingly, we applied Casanovo to a collection of MS/MS spectra drawn from a previous analysis,[9] in which 511 million human spectra from MassIVE were grouped into 60 million clusters, and the clusters were systematically analyzed using targeted open modification searching of representative spectra. The analysis yielded a collection of 39 million unidentified clusters,

containing a total of 207 million MS/MS spectra. For our analysis, we selected 3.4 million of these unidentified, clustered MS/MS spectra from eight randomly selected MassIVE datasets. These MS/MS spectra belong to 573,597 distinct clusters. Because we were investigating spectra that had already failed to be identified using a standard, tryptic pipeline, we opted to use the non-enzymatic Casanovo model (Casanovo$_{ne}$) to assign a peptide to each selected MS/MS spectrum, eliminating peptides for which the predicted *m/z* falls outside the associated mass range. This analysis yielded a total of 1.3 million predicted peptides.

We sought to ascertain how well Casanovo had assigned peptide sequences to these dark matter clusters by addressing this question in two complementary ways. First, we identified all clusters in which a plurality (and at least two) of the spectra were assigned to the same peptide sequence, and then we mapped those peptides to the human reference proteome, allowing at most one amino acid mismatch. The first step of this procedure assigns peptides to 89,250 (16%) of the clusters, of which 65% could be matched to the human proteome. The clusters identified in this fashion vary in size, ranging from 2 to 542 spectra per cluster, but when we limited the above analysis only to clusters larger than a certain size, we observed that the shares of identified clusters more than doubled (Supplementary Figure A.14). Second, we performed a complementary analysis, first eliminating all predicted peptides that do not occur within the human proteome (again, allowing one mismatch) and then finding clusters with two or more spectra assigned the same sequence and no other spectrum assigned to a different sequence. This procedure assigns peptides to 52,523 clusters, corresponding to 9% of all previously unidentified clusters. The overlap between the two approaches—plurality vote followed by proteome matching or vice versa—is high: 98% of the 52,523 clusters overlapped with the clusters from the previous analysis. Overall, Casanovo is able to assign peptides to 196,724 of the 3.4 million unidentified MS/MS spectra using the combination of these two strategies.

One potential reason for an MS/MS spectrum to remain unidentified is the presence in the

generating peptide sequence of a genetic variant that does not appear in the reference proteome. To investigate whether Casanovo is identifying such sequences, we looked more closely at the subset of Casanovo cluster assignments that match to the human proteome with a single amino acid mismatch, focusing on the 51,555 assignments that agree between the two methods described above. Two pieces of evidence suggests that these peptides are indeed enriched for genetic variants. First, we observe an enrichment for amino acid substitutions that can be explained by a corresponding single-nucleotide substitution. Among the Casanovo predictions, 83.4% correspond to a potential single-nucleotide substitution, compared with only 38.6% of all possible amino acid substitutions that fit this criterion. Second, we see a strong enrichment for substitutions with positive BLOSUM62 scores [29]. The BLOSUM score is an integerized log-odds score indicating the empirical substitutability of one amino acid for another. In the BLOSUM62 matrix, only 11% of the 380 non-diagonal entries are positive. However, if we rank the Casanovo-predicted substitutions by frequency, we find that the top five substitutions have BLOSUM scores of 1 or 2 (Supplementary Table A.3). This observation strongly suggests that the Casanovo is predicting substitutions that are biochemically plausible.

3.3 Discussion

Casanovo's excellent performance derives from two sources: the availability of a large, high-quality set of training data, and the use of a machine learning architecture that can make use of that data. Our experiments suggest that the carefully curated MassIVE-KB collection provides particularly good training data. This is likely because the dataset was derived from a massive collection of 669 million spectra, in combination with extremely stringent FDR control. In particular, the data were searched at 1 % FDR, after which only the top 100 PSMs for each unique precursor were retained, corresponding to 30 million high-quality PSMs (uniformly 0 % FDR from the original searches).

The transformer architecture is uniquely suited to contextualize the elements of variable length sequences and has therefore proven immensely successful in modeling natural language. In comparison to recurrent neural networks, the transformer architecture is able to learn long-range dependencies between the elements of a sequence and can be parallelized for efficient training. By encoding the peaks of a mass spectrum as a sequence, similar to tokenizing the words of a sentence, Casanovo leverages the strengths of the transformer architecture and the rapid advances pioneered in large language models to improve *de novo* peptide sequencing from MS/MS spectra [17]. One important open question, which we leave for future work, is how the number of model parameters impacts *de novo* sequencing performance.

Casanovo's utility extends beyond the applications we have demonstrated here. Most obviously, any application in which a peptide database is unavailable, incomplete, or extremely large may benefit from *de novo* sequencing, such as paleoproteomics, forensics, or astrobiology [34]. However, even in the analysis of human or model organism data, Casanovo can assist in the detection of "foreign" spectra, i.e., spectra generated by peptides that are not present in the database. Such foreign spectra might correspond to contaminants introduced during the experiment itself, but they can also represent microbial species, genetic variation, or trans-spliced peptides. In general, we envision applying Casanovo as a post-processor for spectra that fail to be assigned a peptide during a standard database search procedure, akin to the last stage of a cascade search [37].

One important application of *de novo* sequencing that we have not yet explored is antibody sequencing. However, a recent publication carried out a systematic comparison of six *de novo* sequencing tools, including Casanovo, on the problem of antibody sequencing [7]. The results show that Casanovo strongly outperforms the competing methods by all of the measures that the authors considered. Notably, this comparison employed a version of Casanovo that used greedy decoding and was trained on only 2 million spectra. Hence, our results (Figure 3.2B) suggest that the version of Casanovo trained from 30 million spectra will yield even better antibody sequencing

performance.

We anticipate many opportunities for fine-tuning the Casanovo model for particular applications. Our analysis with the non-enzymatic model suggests that Casanovo's enzymatic bias can be adjusted by using a relatively small amount of training data. Thus, in the short term, we plan to train variants of Casanovo that are appropriate for a variety of different cleavage enzymes. The Casanovo software makes such fine-tuning straightforward, so any user interested in adapting the model to a particular experimental setting should be able to do so. Longer term, an ideal model would take as input a spectrum along with relevant metadata, such as the digestion enzyme, collision energy, and instrument type, and predict accurately for many different types of experimental settings.

The potential for deep learning methods to improve our ability to perform *de novo* sequencing has now been widely recognized. While this paper was under review, at least six additional deep learning *de novo* sequencing methods have been published, including GraphNovo [45], PepNet [42], Denovo-GCN [70], Spectralis [39], π-HelixNovo [73], and NovoB [41]. Clearly, the field would benefit from an exhaustive and rigorous benchmark comparison of this growing field of tools.

On a related note, at this stage one of the key bottlenecks in the field is the absence of a rigorous method for confidence estimation for *de novo* sequencing. In our metaproteomics analysis, we have matched Casanovo predictions to a target and corresponding decoy peptide database, but such an approach misses out on the power of *de novo* sequencing to assign peptides to foreign spectra. Thus, an open question is whether, for a given data-dependent acquisition dataset, Casanovo outperforms a standard database search procedure in terms of statistical power to detect peptides. Trained from sufficiently large training sets, we may be approaching the end of database searching as the go-to method for analysis of DDA tandem mass spectrometry data.

3.4 Methods

3.4.1 Casanovo

Casanovo consists of a transformer encoder and decoder stack as described by Vaswani *et al.* [65], which are respectively responsible for learning latent representations of the input spectrum peaks and decoding the amino acid sequence of the spectrum's generating peptide. The encoder takes d-dimensional spectrum peak embeddings as input and outputs d-dimensional latent representation vectors for each peak. Subsequently, the decoder takes as input these representations of prefix amino acids, coupled with a d-dimensional precursor embedding encapsulating precursor m/z and charge information, to predict the next amino acid in the peptide sequence. We discuss different aspects of our modeling strategy in detail below.

Spectrum preprocessing

We preprocess each mass spectrum by removing noise peaks and scaling the peak intensities before they are transformed into input embeddings for Casanovo. First, we discard any peaks outside the range 50–2500 m/z, as well as any peaks within 2 Da of the observed precursor mass. We then remove any peaks with an intensity value lower than 0.01% of the most intense peak's intensity, and we retain up to 150 of the most intense peaks in the spectrum. Finally, peaks intensities are square-root transformed and then normalized by dividing by the sum of the square-root intensities.

Input embeddings

Each spectrum $S = \{(m_j, I_j)\}_{j=1}^{N}$ is a bag of peaks, where each peak (m_j, I_j) is a 2-tuple representing the m/z value and intensity of the peak. For the task of *de novo* peptide sequencing, the most important relationships for our model to learn are how the spacing of m/z values between each pair of peaks corresponds to the peptide ions that may have generated them. Secondarily to the

spacing of *m/z* values, the intensity of each peak also contains information about the generating ion; for example, y-ions are generally more intense than b-ions for some fragmentation methods. Given this prior knowledge, we chose embedding methods that would enable Casanovo to learn from the spacing of *m/z* values and that would emphasize the relative importance of these peak attributes for the *de novo* sequencing task.

We use a fixed, sinusoidal embedding [65] to project each *m/z* value to a d-dimensional vector, the *m/z* embedding f. Specifically, we create the *m/z* embedding from an equal number of sine and cosine waveforms spanning the wavelengths from 0.001 to 10,000 *m/z*, where each feature in the embedding f_i is a value from one waveform (Supplementary Figure A.15A). Let λ_{max} be the maximum wavelength, λ_{min} be the minimum wavelength, i be the index of the feature (zero-based), and d be the number of features. We begin by defining the number of features that are to be represented by sine and cosine waveforms as d_{sin} and d_{cos}, respectively:

$$d_{\text{sin}} = \lceil \frac{d}{2} \rceil \tag{3.1}$$

$$d_{\text{cos}} = d - d_{\text{sin}} \tag{3.2}$$

The encoded features are then calculated as:

$$f_i(m_j) = \begin{cases} \sin(m_j / (\frac{\lambda_{\text{min}}}{2\pi} (\frac{\lambda_{\text{max}}}{\lambda_{min}})^{i/(d_{\text{sin}}-1)})), & \text{for } i \leq d/2 \\ \cos(m_j / (\frac{\lambda_{\text{min}}}{2\pi} (\frac{\lambda_{\text{max}}}{\lambda_{min}})^{(i-d_{\text{sin}})/(d_{\text{cos}}-1)})), & \text{for } i > d/2 \end{cases} \tag{3.3}$$

where $\lambda_{\text{max}} = 10,000$ and $\lambda_{\text{min}} = 0.001$ in Casanovo.

These input embeddings provide a granular representation of high-precision *m/z* information and, similar to relative positions in the original transformer model [65], may help the model attend to *m/z* differences between peaks, which are critical for identification of amino acids in the

peptide sequence. In the *m/z* embeddings, we chose high-frequency waveforms to capture the fine structure present in a mass spectrum, such as that introduced by isotopes and near isobaric species. The waveforms then capture more distant relationships as they progress to lower frequencies; thus, if one were to subtract the *m/z* of one peak from another, the features that are activated depend on the scale of their relationship. While consecutive b- and y-ions may activate one set of features in f, complementary b- and y-ions would likely activate a later set due to their larger *m/z* difference. Furthermore, the cosine similarity between pairs of *m/z* embeddings are negatively correlated with their *m/z* (Supplementary Figure A.15B). We postulate that this property preserves information about *m/z* distances—which are critical for *de novo* peptide sequencing—in a manner that is readily accessible to the subsequent transformer layers.

The intensity, which is measured with lower precision than the *m/z* value, is embedded by projection to d dimensions through a linear layer, after which the *m/z* and intensity embeddings are summed to produce the input peak embedding. However, in developing Casanovo, we found that using a sinusoidal encoding for intensity as well as *m/z* performs similarly.

Precursor information, used as input to the decoder, consists of the total mass $m_{\text{prec}} \in R$ and charge state $c_{\text{prec}} \in \{1, ..., 10\}$ associated with the spectrum. We use the same sinusoidal position embedding as peak *m/z*'s for m_{prec}: c_{prec} is embedded using an embedding layer, and the embeddings are summed to obtain the input precursor embedding.

Modeling *de novo* peptide sequencing as a sequence-to-sequence task

The transformer architecture in Casanovo follows the standard encoder-decoder design of Vaswani *et al.* [65]. The process begins by embedding the peaks of a mass spectrum to obtain input embeddings f, which are then contextualized using the transformer encoder stack. Thus, a full mass spectrum consisting of k peaks is represented as an unordered sequence of peak embeddings $g \in \mathbb{R}^{k \times d}$. The self-attention mechanism of the transformer encoder learns relationships between

these peaks and outputs a contextualized embedding of each peak in the mass spectrum $\hat{g} \in \mathbb{R}^{k \times d}$.

The decoding process begins by feeding both the precursor embedding and the contextualized spectrum embedding $\hat{g}$ into the transformer decoder stack. Decoding proceeds in an autoregressive manner; up to a maximum number of iterations, the decoder stack will attempt to predict the next amino acid in the generating peptide from c-terminus to n-terminus. During the decoding phase, a learned representation of the previously predicted amino acid is concatenated to the input into the decoder for each iteration and summed with a sinusoidal positional embedding of its position in the sequence. The output of the decoder are scores $s \in \mathbb{R}^{p \times v}$ representing how confident Casanovo is about each amino acid it has predicted for a peptide sequence of length p and an amino acid vocabulary of size v.

Training and inference strategy

Taking the previously described embeddings as input, the transformer outputs scores which are treated as a probability distribution over the amino acid vocabulary for the next position in the sequence at each decoding step. The amino acid vocabulary includes 20 canonical amino acids (with cysteine considered to be carbamidomethylated), post-translationally modified versions of three of them (oxidation of methionine and deamidation of asparagine or glutamine), N-terminal modifications (acetylation, carbamylation, loss of ammonia, and the combination of loss of ammonia and carbamylation), plus a special `stop` token to signal the end of decoding, yielding a total of 28 tokens. During training, the decoder is fed the amino acid prefix for the ground truth peptide following the teacher forcing paradigm [69]. Cross-entropy between the model output probabilities and a binary matrix representing the amino acid sequence of the ground truth peptide is minimized as the objective function. During inference, beam search is used to find the highest-scoring predicted peptide sequence, with k a user-specified value for the number of beams. At each prediction step, for every peptide prefix considered, the k top-scoring amino

acids are selected, after which the k top-ranked amino acid sequences are used for the subsequent decoding step. Beams are terminated when the `stop` token is predicted, the predicted peptide mass is similar to (given the precursor mass tolerance) or exceeds the precursor mass, or the pre-defined maximum peptide length of $\ell = 100$ amino acids is reached. As final prediction, the top-scoring peptide sequence that fits the precursor mass tolerance (optionally accounting for isotope offsets) is selected. If no peptide prediction fits the precursor mass tolerance, the top-scoring peptide sequence with a non-matching peptide mass is selected.

Model and training hyperparameters

We train models with nine layers, embedding size $d = 512$, and eight attention heads, yielding a total of ~47M model parameters. A batch size of 32 spectra and 10^{-5} weight decay is used during training, with a peak learning rate of 5×10^{-4}. The learning rate is linearly increased from zero to its peak value in 100,000 warm-up steps, followed by a cosine shaped decay. The MassIVE-KB model was trained for a single epoch on 30 million PSMs from the MassIVE-KB dataset, which took approximately 8 days on 4 RTX 2080 Ti GPUs, while evaluating the performance on the validation set after every 50,000 iterations. The final model weights were taken from the snapshot with the lowest validation loss. This model was fine-tuned on the non-enzymatic training dataset using a peak learning rate of 5×10^{-5}. We selected the fine-tuned model with the minimal validation loss, which occurred after 5 epochs. These model hyperparameters—number of layers, embedding size, number of attention heads, and learning rate schedule—are used for all downstream experiments unless otherwise specified.

Precursor *m/z* filtering

A critical constraint in *de novo* peptide sequencing requires the difference between the total mass of the predicted peptide m_{pred} and the observed precursor mass m_{prec} of the spectrum to

be smaller than a threshold value ϵ (specified in ppm) for the predicted sequence to be plausible: $\Delta m_{ppm} = \frac{|m_{\text{prec}} - m_{\text{pred}}| \times 10^6}{m_{\text{prec}}} < \epsilon$. Therefore, in addition to providing precursor information as an input for the model to learn from, we filter out peptide predictions that do not satisfy the above constraint. The threshold value ϵ is a property of the mass spectrometer that the data is collected with, and hence is known at inference time. Accordingly, we choose ϵ based on the precursor mass error tolerance used in the database search to obtain ground truth peptide sequences for the test data.

Code availability

The transformer model was implemented using PyTorch [52] and PyTorch Lightning [23]. Additionally, NumPy (`https://doi.org/10.1038/s41586-020-2649-2`), Pandas (`10.25080/Majora-92bf1922-00a`), and Scikit-Learn (`https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html`) were used for scientific data processing; and spectrum_utils [8], Pyteomics [52], and ppx [25] were used to process MS/MS data. Matplotlib [31] and Seaborn (`https://doi.org/10.21105/joss.03021`) were used for visualization purposes.

Casanovo's source code and trained model weights are available under the Apache 2.0 license at `https://github.com/Noble-Lab/casanovo`. Casanovo, Casanovo$_{bm}$ and Casanovo$_{ne}$ are all trained using version 4.0.1.

3.4.2 Datasets

MassIVE-KB dataset

A large-scale, heterogeneous dataset derived from the MassIVE knowledge base (MassIVE-KB; v.2018-06-15) was used to develop Casanovo [66]. The MassIVE-KB data set consists of 31 TB of human data from 227 public proteomics datasets, containing over 669 million MS/MS spectra.

MassIVE-KB contains a designated subset of 30,506,973 "high quality" PSMs, identified by applying a strict (~0%) PSM-level FDR filter and then selecting at most 100 PSMs for each combination of peptide sequence and charge. These 30 million PSMs were randomly split so that the training, validation and test sets are disjoint at the peptide-level and consist of approximately 28 million training PSMs, 1 million validation PSMs, and 1 million test PSMs.

Nine-species benchmark dataset

We created a new version of the nine-species benchmark originally described by Tran *et al.* [63] To do so, we downloaded the RAW files from the same nine PRIDE projects (Supplementary Table A.2) and converted them to MGF format using the ThermoRawFileParser v1.3.4. We also downloaded the corresponding nine Uniprot reference proteomes and constructed a Tide index for each one, using Crux version 4.1. For one species (*Vigna mungo*), no reference proteome is available, so we used the proteome of the closely related species *Vigna radiata*. We specified Cys carbamidomethylation as a static modification and allowed for the following variable modifications: Met oxidation, Asn deamidation, Gln deamidation, N-term acetylation, N-term carbamylation, N-term NH_3 loss, and the combination of N-term carbamylation and NH_3 loss by using the tide-index options `-mods-spec 1M+15.994915,1N+0.984016,1Q+0.984016 -nterm-peptide-mods-spec 1X+42.010565,1X+43.005814,1X-17.026549,1X+25.980265 -max-mods 3`. Note that one of the nine experiments (*Mus musculus*) was performed using SILAC labeling, but we searched without SILAC modifications and hence include in the benchmark only PSMs from unlabeled peptides. Each index also contains a shuffled decoy peptide corresponding to each target peptide. Each MGF file was searched against the corresponding index using the precursor window size and fragment bin tolerance specified in the original study (Supplementary Table A.2). We used XCorr scoring with Tailor calibration [60], and we allowed for 1 isotope error in the selection of candidate peptides. All search results were then analyzed jointly per species using the Crux

implementation of Percolator, with default parameters. For the benchmark, we retained all PSMs with Percolator q value < 0.01. We identified 13 MGF files with fewer than 100 accepted PSMs, and we eliminated all of these PSMs from the benchmark. We then post-processed the PSMs to eliminate peptides that are shared between species. Among the 229,984 unique peptides, we identified 3797 (1.7%) that occur in more than one species. For each such peptide, we selected one of the associated species at random and then eliminated all PSMs containing that peptide in other species. Note that when identifying shared peptides between species, we considered all modified forms of a given peptide sequence to be the same. Hence, if a given peptide appears in more than one species, then that peptide, including all its modified forms, is randomly assigned to a single species and eliminated from the others. The final benchmark dataset consists of 2.8 million PSMs drawn from 343 RAW files. The revised nine-species benchmark is available on MassIVE at `https://doi.org/doi:10.25345/C52V2CK8J`.

3.4.3 Evaluation metrics

We use precision calculated at the amino acid and peptide levels [44, 26, 63] as a function of coverage over the test set as performance measures to evaluate the quality of a given model's predictions. In each case, for each spectrum we compare the predicted sequence to the ground truth peptide from the database search. Following Tran *et al.* [63], for the amino acid-level measures we first calculate the number N^a_{match} of matched amino acid predictions, defined as all predicted amino acids which (1) differ by <0.1 Da in mass from the corresponding ground truth amino acid, and (2) have either a prefix or suffix that differs by no more than 0.5 Da in mass from the corresponding amino acid sequence in the ground truth peptide. We then define amino acid-level precision as $N^a_{\text{match}}/N^a_{\text{pred}}$, where N^a_{pred} is the number of predicted amino acids. For peptide predictions, a predicted peptide is considered a correct match if all of its amino acids are matched.

Among a collection of N^p_{orig} spectra, if our model makes predictions on a subset of N^p_{pred} and correctly predicts N^p_{match} peptides, we define coverage as $N^p_{\text{pred}}/N^p_{\text{orig}}$ and peptide-level precision as $N^p_{\text{match}}/N^p_{\text{pred}}$. To plot a precision-coverage curve, we sort predictions by the confidence score provided by the model. Amino acid-level confidence scores are obtained by applying a softmax to the output of the transformer decoder, which is a proxy for the probability of each predicted amino acid to occur in the given position along the peptide sequence. Casanovo directly outputs amino acid-level confidence scores, and we use the mean score over all amino acids as a peptide-level confidence score.

3.4.4 Competing methods

We downloaded DeepNovo weights from `https://github.com/nh2tran/DeepNovo/tree/PNAS` on Sep 6, 2022. Similar to Casanovo$_{bm}$, DeepNovo and PointNovo were trained in a cross-validated fashion using the original nine-species benchmark, systematically training on eight species and testing on the remaining species. Accordingly, nine different sets of pre-trained DeepNovo weights were available, and the corresponding set of weights were used for testing on each species data set. In the absence of pre-trained PointNovo weights, we cross-validated nine models from scratch by training on eight species and testing on the remaining species. We downloaded the PointNovo code provided by Qiao *et al.* [54] from `https://zenodo.org/records/3960823` on Mar 27, 2023.

We downloaded Novor v1.05.0573 from `https://github.com/compomics/searchgui/tree/master/resources/Novor` on Dec 3, 2022.

3.4.5 Creating a non-enzymatic dataset

To create a dataset of PSMs that does not exhibit tryptic bias, we selected PSMs with a uniform distribution of amino acids at the C-terminal peptide positions from two datasets: MassIVE-KB

[66] and PROSPECT [58]. The MassIVE-KB dataset contains 30 million PSMs and consists entirely of data generated using trypsin; hence, only a small proportion of the MassIVE-KB peptides do not end in K or R, corresponding to those that appear at the C-terminus of the corresponding protein. The PROSPECT dataset is a collection of 61 million PSMs generated from synthetic peptides. We performed three filtering steps on this dataset: (1) removed duplicate peaks with identical *m/z* values from each spectrum, (2) eliminated spectra with fewer than 20 peaks, and (3) eliminated spectra with Andromeda score less than 70, selecting the highest-scoring peptide for each spectrum. To avoid over-representation of some peptides in this dataset, we randomly selected at most 100 PSMs per unique peptide, similar to the processing that was done during the creation of the MassIVE-KB dataset. This pre-selection step reduced the size of the PROSPECT dataset to 12.6 million PSMs. Finally, to create a non-enzymatic dataset containing 1 million peptides, we selected 50,000 PSMs for each canonical amino acid. These PSMs were selected at random from MassIVE-KB, then supplemented as necessary from PROSPECT to obtain the desired count (Supplementary Table A.1). This dataset contained PSMs from 247,859 unique peptides, which were randomly split into training, validation and test sets in the ratio 80/10/10. The non-enzymatic dataset with the training, validation and test splits is available on MassIVE at `https://doi.org/doi:10.25345/C5KS6JG0W`.

3.4.6 Immunopeptidome analysis

For the analysis of the MHC class I peptides, we used the Tide search engine and adopted search settings from the original publication [59]. To create the peptide index, we ran tide-index allowing M oxidation or phosphorylation of S/T/Y, with a maximum of one modification per peptide. We set the digestion to be "non-specific-digest," allowed zero missed cleavages, and specified a peptide length range of 8–15 amino acids. Using the canonical human reference proteome

downloaded from Uniprot on July 17, 2022, the resulting index contains 286,319,284 peptides and an equal number of shuffled decoy peptides. We searched the data using the tide-index command, specifying a precursor window size of 30 ppm and using Tailor calibration. The resulting sets of PSMs from all five runs were analyzed jointly using Percolator with default settings. All of the above commands were implemented within Crux [51] version 4.1-2fab3c91-2022-11-09. For comparative analysis between Casanovo predictions and database search results, only sequences within a peptide length range of 9–15 amino acids were considered. NetMHCpan-4.1 was used to predict MHC binding affinities for peptide sequences [55]. Binding affinity was predicted for 9-mer amino acid motifs in reference to the HLA-A02:01, HLA-A02:17, HLA-B41:01, HLA-B40:02, HLA-C02:02, and HLA-C17:01 alleles of the MHC molecule.

3.4.7 Metaproteomics analysis

We analyzed data from six mass spectrometry runs, three replicates each from the Bering Strait (BSt) and the Chukchi Sea (CS) [46], downloaded in mzXML format from `https://noble.gs.washington.edu/proj/metapeptide`. From the same URL, we also downloaded the two corresponding metapeptide databases, and we downloaded the environmental non-redundant database (env_nr) from NCBI on Nov 12, 2022. We used Tide [18] to build three peptide indices from the two metapeptide databases and from env_nr with default parameters, except we allowed three methionine oxidations per peptide and up to 1 missed cleavage. The resulting indices contained 41,665,963 peptides (CS), 34,116,884 peptides (BSt), and 310,021,565 peptides (env_nr), respectively, as well as an equal number of shuffled decoy peptides. Each mzXML file was searched against the relevant metapeptide database and against env_nr using the tide-search command, specifying a precursor window of 10 ppm, allowing one isotope error, and using Tailor calibration. Finally, we used Percolator [35] with default parameters to jointly analyze the search results from each

set of three runs against the same peptide index. All of the above commands were implemented within Crux [51] version 4.1-2fab3c91-2022-11-09. Metaproteomics spectrum files and peptides detected via database search are available on MassIVE with the dataset identifier MSV000094709 at `https://doi.org/doi:10.25345/C5ST7F78Z`.

To estimate the error rate for the matching of Casanovo predicted peptides against a protein database, we developed a procedure akin to the target-decoy competition used in false discovery rate estimation for mass spectrometry database search [19]. To do so, we created a decoy protein database by randomly shuffling each protein sequence. We then asked whether each Casanovo prediction appears in the target (i.e., unshuffled) or decoy protein list, marking each predicted peptide as matching to the target, decoy, or neither. Peptides assigned to both were randomly assigned to either the target or decoy. We then segregated the peptides by length, and for each length we sorted the Casanovo predictions by confidence score. At each position k in the ranked list, we estimated the rate of random matching among the target matches as D_k/T_k, where D_k (respectively, T_k) is the number of decoys (respectively, targets) with rank smaller than k. We then selected the largest value of k such that $D_k/T_k < \alpha$, for some specified random matching rate α. In this work, we used $\alpha = 0.01$.

We used the online version of Prosit [27] at `https://www.proteomicsdb.org/prosit/` to predict peak intensities and retention times using the Prosit_2020_intensity_hcd and Prosit_2019_irt models, respectively. A fixed collision energy of 27 was used for all peptides based on metadata from the spectrum files. Spectrum peaks predicted by Prosit were matched to observed peaks using a 0.05 Da fragment *m/z* tolerance to calculate the cosine similarity between the predicted and experimental spectra.

3.5 Data availability

The revised nine-species benchmark is available on MassIVE with the dataset identifier MSV000090982 at https://doi.org/doi:10.25345/C52V2CK8J. The non-enzymatic dataset with training, validation and test splits is available on MassIVE with the dataset identifier MSV000094014 at https://doi.org/doi:10.25345/C5KS6JG0W. Metaproteomics spectrum files and peptides detected via database search are available on MassIVE with the dataset identifier MSV000094709 at https://doi.org/doi:10.25345/C5ST7F78Z. The spectrum identifications for Casanovo and other evaluated tools on the revised nine-species benchmark are available on MassIVE with the dataset identifier MSV000094434 at https://doi.org/doi:10.25345/C5B56DG2T. Similarly, Casanovo-only spectrum identifications for the immunopeptidomics, metaproteomics, and dark matter analyses are available on MassIVE with the dataset identifier MSV000093980 at https://doi.org/doi:10.25345/C5VD6PG45. Source data are provided with this paper.

3.6 Code availability

Casanovo's source code and trained model weights from the MassIVE-KB training and non-enzymatic fine tuning are available under the Apache 2.0 license at https://github.com/Noble-Lab/casanovo [74]. Model weights from the nine-species benchmark training are available at https://doi.org/10.5281/zenodo.10694984.

3.7 Figure legends

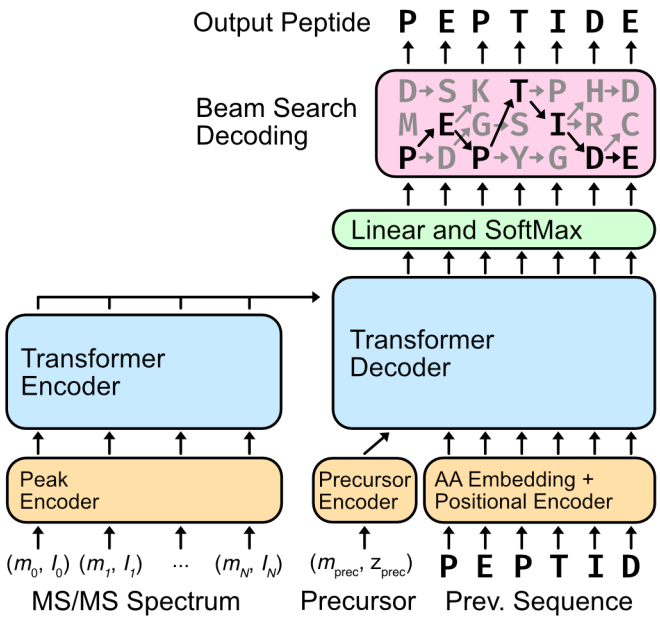


Figure 3.1: **Casanovo performs *de novo* peptide sequencing using a transformer architecture.** The peaks from each MS/MS spectrum are contextualized by the transformer encoder. The resulting peak encodings are then fed into the transformer decoder along with the observed precursor mass and charge to iteratively decode the peptide sequence. Casanovo uses a beam search decoding strategy, following the most promising sequence predictions until they terminate or exceed the precursor mass. The highest scoring peptide sequence is returned as the putative peptide that generated the MS/MS spectrum.

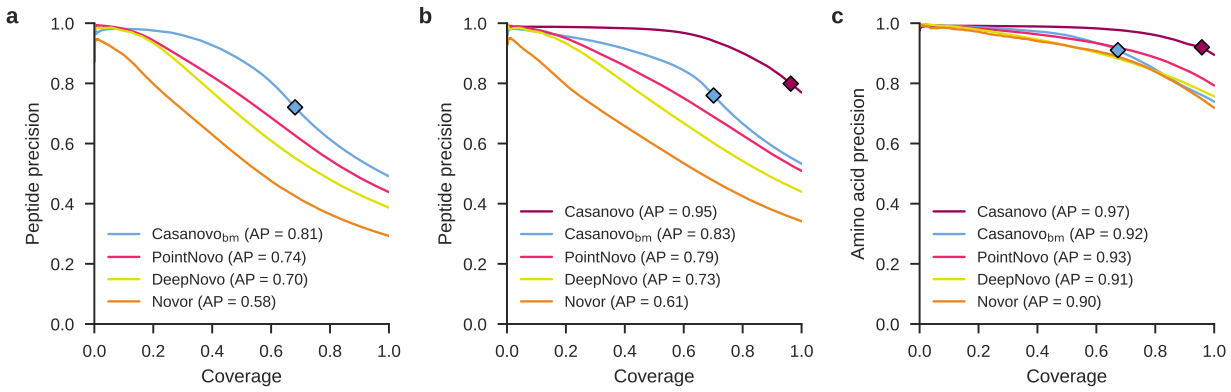


Figure 3.2: **Casanovo outperforms PointNovo, DeepNovo, and Novor on a nine-species benchmark. (a)** Casanovo maintains high peptide-level precision (the proportion of correctly predicted peptides) across all values of coverage (the proportion of spectra for which a prediction is made). Each curve is computed by sorting predicted peptides for all nine species according to their peptide-level confidence scores. For Casanovo, all peptides that pass the precursor *m/z* filter are ranked above peptides that do not pass the filter, and the boundary is indicated by a diamond on each curve. Average precision (AP) corresponds to the area under the precision–coverage curve. **(b)** Same as panel (a), but using the revised benchmark and including a version of Casanovo trained on MassIVE-KB. **(c)** Casanovo's amino acid-level precision is greatly improved by the expanded training data provided by MassIVE-KB. The test set is the revised nine-species benchmark, with PSMs only containing modifications considered by both DeepNovo and Casanovo.

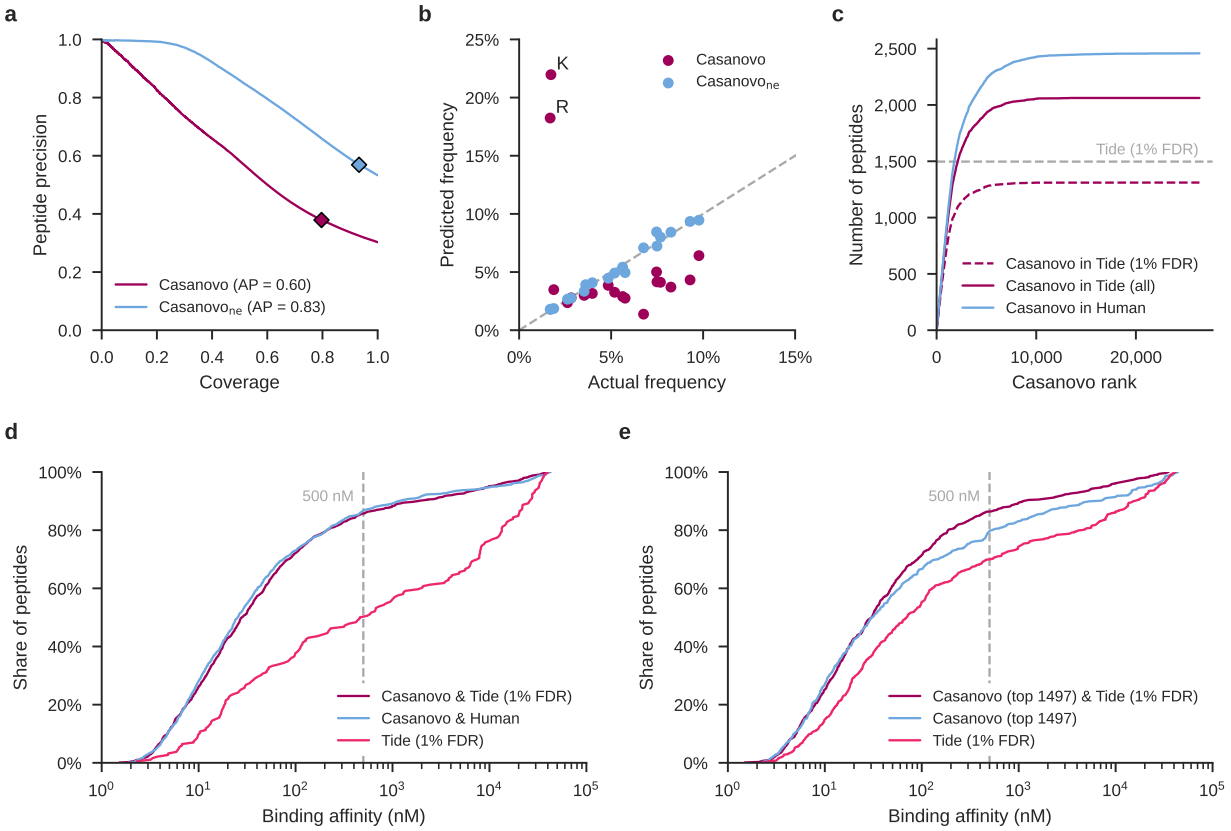


Figure 3.3: **Fine-tuning reduces Casanovo's bias for tryptic peptides.** **(a)** Fine-tuning Casanovo (Casanovo$_{ne}$) improves peptide-level precision on sequencing MS/MS spectra generated by non-tryptic peptides. **(b)** Casanovo$_{ne}$ predicts non-tryptic C-terminal peptides more readily than the standard Casanovo model, improving performance on the non-enzymatic validation set. **(c)** Casanovo detects many peptides that are present in the human proteome but are not detected via database search. The dashed dark pink line only includes peptides detected by database search within the 1% FDR threshold, whereas the solid dark pink line includes all peptides from the database search, irrespective of FDR threshold. **(d)** The peptides proposed by Casanovo generally have higher predicted binding affinities for the MHC class I receptor, matching the performance of a Tide database search. The vertical bar corresponds to the 500 nM binding affinity below which peptides are predicted to be MHC binders. **(e)** Similar to panel (d), but considering only the 1497 peptides that are accepted at 1% FDR by Tide which yielded valid binding affinity predictions from NetMHCpan and a corresponding set of 1497 highest-confidence Casanovo peptides.

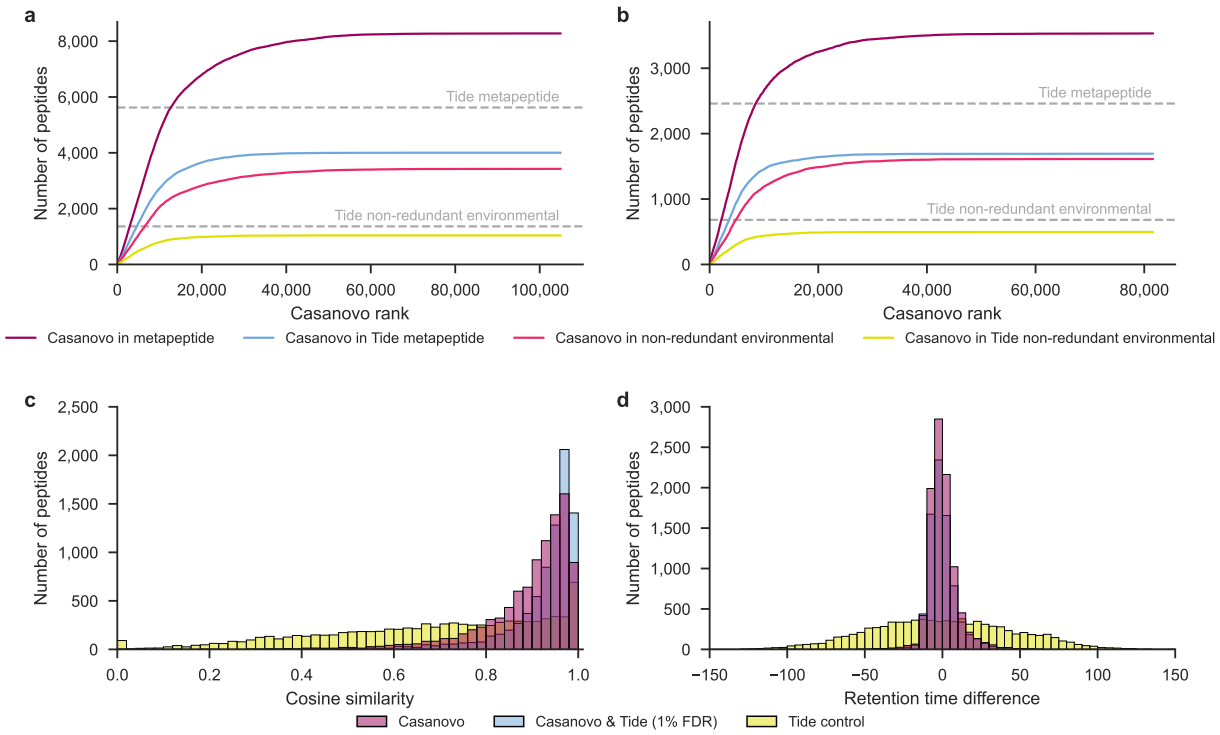


Figure 3.4: **Casanovo improves power to detect peptides from metaproteomics samples.** Casanovo assigns more peptides matching the metapeptide database and the non-redundant environment database than Tide and Percolator at 1% FDR in seawater samples from **(a)** the Bering Sea and **(b)** the Chukchi Sea. Peptides are ranked according to the Casanovo confidence score, assigning each peptide the maximum score across all three runs from each sampling location. Horizontal lines indicate the total number of distinct peptides detected by Tide+Percolator, searching against two different databases. **(c)** The PSMs assigned by Casanovo at a 1% error rate and Tide and Percolator at 1% FDR have high cosine similarities to the predicted MS/MS spectra for the respective peptides from Prosit when compared to control PSMs sampled from the Tide search results with >10% FDR. Each group represents the aggregated results for Bering and Chukchi Sea data, as well as non-redundant environmental and metapeptide databases. **(d)** The PSMs assigned by Casanovo at a 1% error rate and Tide and Percolator at 1% FDR closely align with the predicted retention times from Prosit.

Chapter 4

Training a tandem mass spectrometry foundation model with *de novo* peptide sequencing

This chapter is work done with the following authors:
Melih Yilmaz, Justin Sanders, Wout Bittremieux, William Fondrie, Sewoong Oh, and William S Noble

4.1 Introduction

Foundation models have emerged as one of the most powerful machine learning paradigms for various problem domains in the past few years [12]. These models are trained to learn rich latent representations of an input modality from large amounts of unlabeled data (e.g., online text, protein sequences in public repositories) using self-supervised learning tasks such as masked language modeling. The trained model can subsequently be used to perform a variety of downstream

tasks, relying on the same input modality with little or no supervised fine-tuning for the specific task in question. In many cases, a foundation model will outperform its peers trained only with supervision and will achieve even better performance with a relatively small amount of supervised fine tuning.

Motivated by these results, we set out to leverage *de novo* peptide sequencing, i.e. predicting the generating peptide sequence given a spectrum, as a task to train a foundation model for proteomics tandem mass spectrometry (MS/MS) data. Our Casanovo model [75], which is trained with millions of peptide-spectrum matches on the *de novo* peptide sequencing task, already learns intermediate spectrum embeddings with its encoder component, and we hypothesize that Casanovo's encoder could serve as an MS/MS proteomics foundation model off the shelf, because accurate *de novo* sequencing would require the model to learn about several critical spectral properties that relate to the peptide sequence responsible for generating the spectrum. Accordingly, we hypothesized that our MS/MS foundation model, Casanovo, will provide a starting point for tackling different downstream tasks even without task-specific fine-tuning, including spectrum quality assessment, detection of "chimeric" spectra (i.e., spectra generated by more than one peptide species) and detection of spectra generated by peptides containing post-translational modifications (PTMs), e.g. phosphorylation.

4.2 *De novo* peptide sequencing as pre-training for a proteomics foundation model

Our Casanovo model, which is trained with the *de novo* peptide sequencing task, already learns intermediate spectrum embeddings with its encoder component. We hypothesize that Casanovo's encoder could serve as a mass spectrometry proteomics foundation model off the shelf. For each

downstream task, we train a task-specific predictor head that takes frozen spectrum embeddings from Casanovo as input and we experiment with an XGBoost model as the predictor head.

In the *de novo* sequencing task, the input is an observed spectrum S, and the output is a peptide P, which consists of a series of amino acids. A correct prediction occurs when the output peptide P was actually present in the mass spectrometer and was responsible for generating spectrum S.

Labeled training data for this task is generated by running MS/MS spectra through a database search engine and selecting peptide-spectrum matches (PSMs) subject to an estimated false discovery rate (FDR) threshold. The search engine assigns to each spectrum a peptide drawn from a given peptide database and a corresponding score. Critically, the database contains a mixture of real ("target") peptides, which are known to occur in the organism being analyzed, and "decoy" spectra created by shuffling the targets. The peptide-spectrum matches (PSMs) are ranked by search engine score, including some spectra whose top-scoring peptide is a target and others that randomly matched to a decoy. We can rigorously control the false discovery rate (FDR) among PSMs scoring greater than a given threshold τ [28] by computing the ratio of the number of decoy PSMs (plus 1) over the number of target PSMs. In practice, we typically use an FDR threshold of 1% to create training data for *de novo* sequencing.

Performance on *de novo* sequencing is evaluated by applying the trained model to labeled test data, including novel spectra assigned to novel peptide sequences. The model assigns a score to each predicted peptide, and we rank the predictions by score to produce a curve that plots peptide precision (i.e., proportion of correctly predicted peptides) as a function of coverage (i.e., proportion of spectra with a predicted peptide). The primary performance measure is the average precision along this curve. If the model produces scores for individual amino acids, we can also produce an analogous curve at the amino acid level.

4.3 Downstream tasks

We use our foundation model as a starting point to address three downstream tasks.

4.3.1 Spectrum quality

In the spectrum quality prediction task, the goal is to predict whether a given observed spectrum will be identified with high confidence using the database search procedure described above. The motivation for this task is two-fold. First, if we can quickly identify low quality spectra, then we can save time and potentially boost our statistical power by eliminating these spectra prior to the database search procedure. Second, spectra that are deemed to be high quality by the trained model but nonetheless fail to be identified during database search procedure are good candidates for more expensive computational analyses.

Because spectrum quality prediction is a binary classification task, we can use standard ROC or precision-recall curves to evaluate performance. In practice the task is roughly balanced—~50% of the spectra in a given dataset can be assigned a peptide with high confidence; therefore, we use the area under the ROC curve (AUROC) as the primary performance measure. The presence of foreign spectra make this task particularly challenging, because these may be high quality spectra that will never be confidently assigned a peptide by the database search procedure. As a result, we do not expect *a priori* to be able to achieve AUROC values close to 1.

To create a labeled dataset for this task, we first randomly sample 20 mzML files from the subset of experiments using high-resolution instruments with HCD fragmentation in MassIVE-KB repository and select 10/5/5 files to create training/validation/test splits, where each split contains approximately 450 thousand, 245 thousand and 295 thousand spectra, respectively. We perform a database search against the reference human proteome (UniProt ID UP000005640) using Sage (version 0.14.7) [40] with the default workflow. Spectra that are matched to a peptide under 1%

FDR are labeled as high quality, whereas spectra that failed to be matched are annotated as low quality for the binary classification task. In our dataset, we observe a 40%/60% distribution of high and low quality spectra.

4.3.2 Spectrum chimericity

Tandem mass spectrometry experiments are designed to attempt to isolate individual peptide species, by first separating them by hydrophobicity in the liquid chromatography step and then separating peptides by m/z in the first round of mass spectrometry analysis. Nonetheless, in many cases, two peptides with similar hydrophobicities and m/z values end up being fragmented simultaneously. The result is an MS/MS spectrum that contains peaks corresponding to both peptides. Such chimeric spectra are difficult to analyze. Most database search algorithms assign at most one peptide to each spectrum, and even assigning a single peptide to a chimeric spectrum is challenging due to the presence of unexplained peaks from the undetected peptide.

To our knowledge, detection of chimeric spectra has not previously been solved using machine learning methods. On the other hand, many existing methods generalize the database search procedure to allow chimeric matches [62]. Here, we propose to stop short of actually predicting the generating peptides and instead simply predict whether more than one peptide species is responsible for generating a given MS/MS spectrum. This could be useful, for example, in deciding which spectra to provide as input to one of the tools above.

To train a chimericity predictor, we use spectra from human, mouse and yeast samples for training, validation and test, respectively. The samples were prepared using the method described in [67] and analyzed using an Orbitrap Fusion Lumos mass spectrometer. Raw MS/MS data were converted to mzML format files using MSConvert with peak picking enabled in ProteoWizard (version 3.0.24031) [13]. The human, mouse and yeast MS/MS data were then searched against

a human (20597 proteins, 02/2024), mouse (21701, 02/2024) and yeast (6060, 02/2024) proteome database, respectively, using FragPipe (version 22.0) with the default workflow and "DDA+" mode (i.e., wide window database search). We perform a wide window database search using FragPipe, which allows spectra to be assigned multiple peptides. For the binary classification task, spectra assigned more than one peptide are labeled chimeric and spectra identified with a single peptide are labeled non-chimeric. Unidentified spectra are discarded. For each of the splits, we have 60,000 to 65,000 spectra identified with at least one peptide, and approximately 45% of these spectra are chimeric.

4.3.3 Post-translational modification detection

The final downstream task we consider is detection of spectra generated by peptides containing post-translational modifications (PTMs). A PTM is a molecular group that attaches to the side-chain of one of the amino acids in a peptide. Common PTMs include methylation and phosphorylation, but many more potential types of PTMs exist in nature, and some are quite rare. The peptide database used during database search of MS/MS data can be augmented to include PTMs, but because of the many potential types of PTMs and the fact that a single peptide can harbor multiple PTMs, accounting for all possible PTMs is not computationally or statistically feasible. A model capable of identifying which PTMs are associated with a given MS/MS spectrum would thus be very valuable.

For this task, we consider the detection of spectra from phosphorylated peptides, also framed as a binary classification task. To train a phosphorylation classifier, we use 19.2 million labeled spectra from the human phosphoproteome dataset [50] used to train AHLF [2], a state-of-the-art phosphorylation predictor. The human phosphoproteome dataset consists of 112 individual PRIDE repositories, containing 101 human cell or tissue types, where each dataset was collected with

phospho-enrichment assays. To create labeled data for training the AHLF, human phosphoproteome data was subjected to database search, and a binary label was assigned to PSMs indicating phosphorylated or unphosphorylated peptides (see [2] for details). Of the resulting 19.2 million spectra 54% are labeled phosphorylated. Following the cross-validation setup described in [2], we use train, validation and test splits of the AHLF-α model. For phosphorylation prediction, we use the F1 score in addition to the AUROC metric to account for class imbalances at the level of individual datasets within the test set.

4.3.4 Competing methods

In each task, we compare Casanovo coupled with a task-specific predictor head against at least two baselines: 1) we bin spectrum peaks along the *m/z* axis to obtain spectrum embeddings and then train an XGBoost classifier directly on those embeddings, 2) we train a Casanovo spectrum encoder, which has the same architecture as Casanovo, and a linear classifier head from scratch to learn the downstream tasks end-to-end. For the PTM detection task, we also benchmark against a task-specific, state-of-the-art classifier [2].

For the Casanovo pipeline, we use Casanovo version 3.3.1 to get 512-dimensional spectrum embeddings and train an XGBoost classifier with the default hyperparameters [14]. For the binned embeddings, we experimented with different binning resolutions to obtain the spectrum embeddings and settled on using 100-bin, i.e. 100-dimensional, embeddings (Supplementary Figure A.16). Peaks outside the range 140 *m/z* to 2,000 *m/z* are filtered out, and the remaining peak intensities are binned into 100 bins at 18.6 *m/z*. For the non-pre-trained Casanovo pipeline, we use a smaller Casanovo encoder (1 layer versus 9 layers in the pre-trained model) which takes precursor *m/z* and charge alongside spectrum peaks as input but is otherwise identical to the pre-trained model.

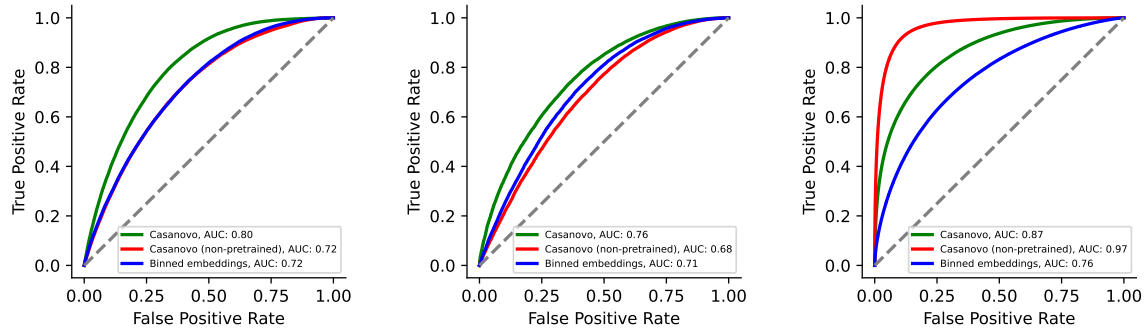


Figure 4.1: **Comparison of Casanovo and two baselines on three downstream tasks** ROC curves and the area under the curve (AUC) reported for: (A) Spectrum quality prediction, (B) Chimericity prediction, (C) PTM detection.

4.4 Results

Benchmarking results indicate that *de novo* sequencing pre-training could boost performance on downstream tasks. The results on the spectrum quality prediction task show that Casanovo embeddings with an XGBoost predictor achieve 0.80 AUROC on the test set, outperforming the binned embeddings and non-pre-trained Casanovo-encoder, with 0.72 AUROC for both (Figure 4.1A). Similarly for the spectrum chimericity prediction Casanovo achieves 0.76 AUROC on the test set, surpassing binned embeddings and the non-pre-trained Casanovo-encoder, with 0.71 and 0.68 AUROC, respectively (Figure 4.1B). For the PTM detection task, Casanovo outperforms the binned embeddings, 0.87 versus 0.70 AUROC, on the combined test set but the non-pre-trained Casanovo-encode surpasses both with 0.97 AUROC (Figure 4.1C).

We hypothesized that Casanovo's lack of access to precursor *m/z* and charge information, unlike the non-pre-trained Casanovo encoder which takes precursor information alongside spectrum peaks as input, could explain its relatively weaker performance in PTM detection. To test this, we perform a simple experiment by concatenating precursor *m/z* and charge as two additional dimensions to the Casanovo spectrum embeddings that are input to the XGBoost classifier during

training and inference. The results show only a small increase in AUROC for the precursor added Casanovo pipeline, suggesting that the availability of precursor information is unlikely to account for the performance difference alone (Supplementary Figure A.17).

Finally, we compare Casanovo with the state-of-the-art phosphorylation predictor AHLF. On the 26 tissue samples that make up the test set, Casanovo embeddings outperform AHLF in 6 datasets by at least 0.05 in F-1 score, while AHLF outperforms Casanovo by the same margin in 9 datasets (Table 4.1). By the same measure, non-pre-trained Casanovo encoder surpasses AHLF in 14 datasets and it is only outperformed by AHLF in 2 datasets (Supplementary Table A.4).

Dataset	n_non-phospho	n_phospho	Bacc-AHLF	Bacc-Casanovo	F1-AHLF	F1-Casanovo	AUROC-AHLF	AUROC-Casanovo
OVAS	90936	37720	0.95	0.85	0.92	0.75	0.99	0.93
TOV-21-Primary	62350	26978	0.94	0.84	0.92	0.73	0.99	0.92
ES2-Primary	16297	6667	0.94	0.84	0.91	0.65	0.99	0.92
Daudi	150915	210916	0.89	0.83	0.90	0.85	0.96	0.91
U2OS	92329	205353	0.77	0.75	0.90	0.87	0.95	0.90
HaCaT	19216	113775	0.78	0.80	0.95	0.92	0.93	0.90
HT-29	1625	27531	0.72	0.82	0.97	0.94	0.92	0.91
HeLa	1469194	2949614	0.83	0.77	0.89	0.85	0.92	0.86
HEPG2	426	45416	0.76	0.80	0.98	0.94	0.92	0.90
A549	4068	172792	0.82	0.81	0.92	0.91	0.91	0.90
Colon	8359	28798	0.81	0.75	0.90	0.81	0.90	0.83
Primary-Gastro	22026	219767	0.72	0.66	0.94	0.91	0.88	0.79
LNCaP	53851	5200	0.78	0.66	0.45	0.24	0.87	0.73
RPMI-8226	1184	413	0.76	0.66	0.65	0.50	0.87	0.72
HEK293	322811	332690	0.77	0.68	0.73	0.68	0.86	0.75
Primary-Prostate	9223	100617	0.77	0.68	0.89	0.84	0.86	0.78
Primary-AML	494184	5893	0.72	0.74	0.29	0.37	0.85	0.83
Kasumi-1	2294	29470	0.75	0.62	0.88	0.84	0.85	0.68
HPAC	594	807	0.67	0.65	0.56	0.63	0.78	0.74
SU.86.86	909	984	0.65	0.65	0.53	0.60	0.77	0.74
CFPAC-1	999	780	0.64	0.63	0.52	0.55	0.76	0.72
PANC-05-04	1079	1426	0.65	0.67	0.55	0.66	0.74	0.74
PANC-02-03	273	815	0.65	0.62	0.56	0.68	0.72	0.67
OVSAYO	11515	28	0.56	0.61	0.02	0.01	0.69	0.71
HDMVEC	4320	2961	0.59	0.60	0.40	0.56	0.60	0.65

Table 4.1: **Comparison of Casanovo and AHLF across PTM detection datasets.** 26 datasets listed correspond to the holdout split *a* described in [2] and AHLF results are directly taken from the paper. The first two columns indicate the number of non-phosphorylated versus phosphorylated PSMs in each dataset. The performance metrics are balanced accuracy, F1 score and AUROC.

4.5 Discussion

Casanovo pre-trained with *de novo* peptide sequencing shows promise as a mass spectrometry proteomics foundation model, especially in data-constrained downstream applications as exemplified by the chimericity (~60 thousand training PSMs) and quality prediction (~450 thousand training PSMs) tasks in this paper. Results on the PTM detection task reflect a different scenario, where the advantage of large scale pre-training without any task-specific fine-tuning disappears in the presence of a proteome-level training set for the downstream task (over 10 million PSMs, the same order of magnitude as the pre-training data) and deep learning models such as the Casanovo encoder and AHLF to take advantage of it. We leave systematically investigating the impact of downstream data with a learning curve by training with nested subsets of the task datasets, as well as fine-tuning the Casanovo embeddings on downstream applications, for future work.

Other factors contributing to the weaker performance of Casanovo relative to the non-pre-trained version and AHLF on PTM detection could be the absence of phosphorylation in Casanovo's pre-training data and vocabulary, as well as the missing precursor information at Casanovo encoder's input during pre-training. Adding phosphorylation to the Casanovo *de novo* sequencing vocabulary or experimenting with the detection of another PTM Casanovo already learns to predict during pre-training, e.g. deamidation, could help explore the first factor. For the precursor information, removing it from the non-pre-trained Casanovo encoder's input or adding it to Casanovo encoder as input during *de novo* sequencing pre-training are the next steps of inquiry to determine precursor's impact on the PTM detection performance.

A *de novo* sequencing trained proteomics foundation model stands to benefit a wide array of applications that use tandem mass spectrometry data as input, but alternative training tasks could help improve performance on *de novo* sequencing itself. Contrastive learning with PSMs is a candidate which has been shown to help *de novo* peptide sequencing [33], as well as other tasks

[9], but still falls short of utilizing troves of unidentified, potentially lower quality spectra that typically present the hardest challenge in most applications. Self-supervised learning, e.g. masked language modeling with predicting held out peaks' *m/z* or intensity, could help foundation model training scale to billion-spectra datasets but remains to be explored in the context of proteomics data.

Chapter 5

Conclusion

Tandem mass spectrometry coupled with computational methods for data analysis has powered proteomics research for decades and greatly expanded our understanding of the protein machinery across the tree of life. Sequencing proteins at scale to determine their exact amino acid content remains a key computational challenge, even as successive generations of database search and *de novo* sequencing tools have made great progress in terms of accuracy and usability. Machine learning methods have been at the forefront of progress in the field for over a decade, and deep learning in particular has shown strong potential in realizing the goal of accurate *de novo* sequencing of peptides. The success of larger and more expressive models paired with ever-expanding datasets in diverse problem domains sets the course for method development in *de novo* sequencing, including our work, and it would not be surprising for these trends to remain influential in the decades to come.

This dissertation describes our contributions to the *de novo* peptide sequencing and deep learning for mass spectrometry proteomics literature that revolves around Casanovo.

Chapter 2 introduces our original formulation of *de novo* sequencing in analogy to neural machine translation and Casanovo as the first transformer-based *de novo* sequencing method.

We demonstrate its superiority to other deep learning-based techniques on an established multi-species benchmark and explore the implications of a number of design decisions that go into building and training the model.

Chapter 3 presents an improved version of Casanovo and illustrates the impact of a larger and higher quality training dataset on *de novo* sequencing performance. Casanovo is not only shown to outperform other *de novo* sequencing tools on database-search-labeled benchmarks but also pitted against database search itself in challenging application domains such as immunopeptidomics and metaproteomics with promising results.

Finally, Chapter 4 reframes *de novo* peptide sequencing as a pre-training task to build a tandem mass spectrometry foundation model for proteomics data. Casanovo encoder is used as a starting point for three downstream tasks without fine-tuning and showcases the utility of *de novo* sequencing-based pre-training in data-constrained settings.

Since its publication in 2022, Casanovo has been used by proteomics researchers for a range of applications from metagenomics to microbial proteomics [68, 38]. Antibody sequencing is the most popular application with both benchmarking studies [7, 15] and discovery-oriented work [6, 32] utilizing Casanovo. In addition to being applied to diverse problems in proteomics, Casanovo has also inspired a growing number of novel transformer-based *de novo* sequencing methods [10], which combined it with new machine learning techniques such as diffusion-based refinement, contrastive pre-training and bi-directional decoding [20, 33, 71].

Despite the advances made with deep learning in *de novo* peptide sequencing, the field still faces fundamental challenges. One such challenge is the accurate evaluation of existing tools, which commonly involves treating the database search results as ground truth for both training and validating the models. This risks inheriting the limitations of database search strategies as well as rendering misleading performance estimates, which are confined to the subset of spectra identifiable with database search. However, in the absence of database search results,

no statistically rigorous methods of controlling the false discovery rate for *de novo* sequences currently exist, making it difficult to set a confidence threshold for predictions. Overcoming these limitations in future work would greatly expand the utility of *de novo* sequencing tools and might even see them become an essential component of the standard spectrum identification procedure.

Beyond just the *de novo* peptide sequencing problem, there are many outstanding challenges in proteomics that large scale machine learning could help with, and for that reason, building a foundation model for tandem mass spectrometry is another promising avenue for future research. Despite the prohibitive computational costs and expertise associated with training such models, they stand to lower the bar for entry in downstream tasks and one-off use cases they will be applied to. For example, being able to sequence novel PTMs that Casanovo was not originally trained to predict currently requires retraining the model from scratch and finding sufficient training data for the infrequently identified PTMs. Having a foundation model that Casanovo would be built from with limited fine-tuning could dramatically decrease the amount of data and compute required for fine-tuning with novel PTMs.

To conclude this dissertation on a note of optimism for the future of our field, as machine learning techniques advance, publicly available datasets expand, and mass spectrometry technology improves, we believe *de novo* sequencing tools are likely to see wider adoption, facilitating analyses that were previously difficult or unattainable.

Bibliography

[1] R. Aebersold and M. Mann. Mass-spectrometric exploration of proteome structure and function. *Nature*, 537:347–355, 2016.

[2] Tom Altenburg, Sven H Giese, Shengbo Wang, Thilo Muth, and Bernhard Y Renard. Ad hoc learning of peptide fragmentation from mass spectra enables an interpretable detection of phosphorylated and cross-linked peptides. *Nature Machine Intelligence*, 4(4):378–388, 2022.

[3] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

[4] D. K. Bailey, M. T. McDevitt, M. C. Westphall, D. J. Pagliarini, and J. J. Coon. Intelligent data acquisition blends targeted and discovery methods. *Journal of Proteome Research*, 13:2152–2161, 2014.

[5] D. K. Bailey, C. M. Rose, G. C. McAlister, J. Brumbaugh, P. Yu, C. D. Wenger, M. C. Westphall, J. A. Thomson, and J. J. Coon. Instant spectral assignment for advanced decision tree-driven mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 109(22):8411–8416, 2012.

[6] Parker T Bassett, Binh A Nguyen, Virender Singh, Patrick Garrett, James J Moresco, Gurbakhash Kaur, Shumaila Afrin, Maja Pekala, and Lorena Saelices. Cryo-EM reveals that cardiac IGLV6-derived AL fibrils can be polymorphic. *bioRxiv*, pages 2024–12, 2024.

[7] 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 Bioinoformatics*, 12 2022. Advance online access.

[8] W. Bittremieux. spectrum_utils: A python package for mass spectrometry data processing and visualization. *Analytical Chemistry*, 92(1):659–661, 2020.

[9] W. Bittremieux, D. H. May, J. Bilmes, and W. S. Noble. A learned embedding for efficient joint analysis of millions of mass spectra. *Nature Methods*, 19(6):675–678, 2022.

[10] Wout Bittremieux, Varun Ananth, William E. Fondrie, Carlo Melendez, Marina Pominova, Justin Sanders, Bo Wen, Melih Yilmaz, and William S. Noble. Deep learning methods for de novo peptide sequencing. *Mass Spectrometry Reviews*, n/a(n/a), 2024.

[11] Wout Bittremieux, David L Tabb, Francis Impens, An Staes, Evy Timmerman, Lennart Martens, and Kris Laukens. Quality control in mass spectrometry-based proteomics. *Mass Spectrometry Reviews*, 37(5):697–711, 2018.

[12] Rishi Bommasani, Drew A Dai, Dani Yogatama, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[13] M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M. Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz, J. E. Katz, D. B. Agus, M. J. MacCoss, D. L. Tabb, and P. Mallick. A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*, 30(10):918–920, 2012.

[14] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[15] Maria Chernigovskaya, Khang Le Quy, Maria Stensland, Sachin Singh, Rowan Nelson, Melih Yilmaz, Konstantinos Kalogeropoulos, Pavel Sinitcyn, Anand Patel, Natalie Castellana, et al. Systematic benchmarking of mass spectrometry-based antibody sequencing reveals methodological biases. *bioRxiv*, pages 2024–11, 2024.

[16] V. Dancik, T.A. Addona, K.R. Clauser, J.E. Vath, and P.A. Pevzner. *De novo* peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 6(3-4):327–342, 1999.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[18] B. Diament and W. S. Noble. Faster SEQUEST searching for peptide identification from tandem mass spectra. *Journal of Proteome Research*, 10(9):3871–3879, 2011.

[19] J. E. Elias and S. P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3):207–214, 2007.

[20] K. Eloff, K. Kalogeropoulos, O. Morell, A. Mabona, J. B. Jespersen, W. Williams, S. P. B. van Beljouw, M. Skwark, A. H. Laustsen, S. J. J. Brouns, et al. De novo peptide sequencing

with InstaNovo: Accurate, database-free peptide identification for large scale proteomics experiments. *bioRxiv*, pages 2023–08, 2023.

[21] J. K. Eng, A. L. McCormack, and J. R. Yates, III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5:976–989, 1994.

[22] J. K. Eng, B. C. Searle, K. R. Clauser, and D. L. Tabb. A face in the crowd: recognizing peptides through database search. *Molecular and Cellular Proteomics*, 10(11), 2011.

[23] W Falcon and TPL Team. PyTorch Lightning the lightweight PyTorch wrapper for high-performance AI research. scale your models, not the boilerplate, 2019.

[24] Bernd Fischer, Volker Roth, Joachim M Buhmann, Jonas Grossmann, Sacha Baginsky, Wilhelm Gruissem, Franz Roos, and Peter Widmayer. A hidden Markov model for de novo peptide sequencing. *Advances in Neural Information Processing Systems*, 17:457–464, 2005.

[25] W. Fondrie, W. Bittremieux, and W. S. Noble. ppx: Programmatic access to proteomics data repositories. *Journal of Proteome Research*, 20(9):4621–4624, 2021.

[26] A. Frank and P. Pevzner. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Analytical Chemistry*, 77:964–973, 2005.

[27] S. Gessulat, T. Schmidt, D. P. Zolg, P. Samaras, K. Schnatbaum, J. Zerweck, T. Knaute, J. Rechenberger, B. Delanghed, A. Huhmer, U. Reimer, H. Ehrlich, S. Aiche, B. Kuster, and M. Wilhelm. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nature Methods*, 16(6):509, 2019.

[28] K. He, Y. Fu, W.-F. Zeng, L. Luo, H. Chi, C. Liu, L.-Y. Qing, R.-X. Sun, and S.-M. He. A theoretical foundation of the target-decoy search strategy for false discovery rate control in proteomics. *arXiv*, 2015. https://arxiv.org/abs/1501.00537.

[29] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89:10915–10919, 1992.

[30] D. F. Hunt, R. A. Henderson, J. Shabanowitz, K. Sakaguchi, H. Michel, N. Sevilir, A. L. Cox, E. Appella, and V. H. Engelhard. Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science (New York, N.Y.)*, 255(5049):1261–1263, March 1992.

[31] John D Hunter. Matplotlib: A 2D graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007.

[32] Yuming Jiang, Devasahayam Arokia Balaya Rex, Dina Schuster, Benjamin A Neely, Germán L Rosano, Norbert Volkmar, Amanda Momenzadeh, Trenton M Peters-Clarke, Susan B Egbert, Simion Kreimer, et al. Comprehensive overview of bottom-up proteomics using mass spectrometry. *ACS Measurement Science Au*, 2024.

[33] 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. *arXiv preprint arXiv:2312.11584*, 2023.

[34] Richard S Johnson, Brian C Searle, Brook L Nunn, Jason M Gilmore, Molly Phillips, Chris T Amemiya, Michelle Heck, and Michael J MacCoss. Assessing protein sequence database suitability using de novo sequencing. *Molecular & Cellular Proteomics*, 19(1):198–208, 2020.

[35] L. Käll, J. D. Canterbury, J. Weston, W. S. Noble, and Michael J MacCoss. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4(11):923–925, 2007.

[36] K. Karunratanakul, H.-Y. Tang, D. W. Speicher, E. Chuangsuwanich, and S. Sriswasdi. Uncovering thousands of new peptides with sequence-mask-search hybrid de novo peptide sequencing framework. *Molecular and Cellular Proteomics*, 18:2478–2491, 2019.

[37] A. Kertesz-Farkas, U. Keich, and W. S. Noble. Tandem mass spectrum identification via cascaded search. *Journal of Proteome Research*, 14(8):3027–3038, 2015.

[38] Simon Klaes, Christian White, Lisa Alvarez-Cohen, Lorenz Adrian, and Chang Ding. De novo assembled databases enable species-specific protein-based stable isotope probing of microbiomes without prior knowledge of the community composition. *bioRxiv*, pages 2024–11, 2024.

[39] Daniela Klaproth-Andrade, Johannes Hingerl, Yanik Bruns, Nicholas H Smith, Jakob Träuble, Mathias Wilhelm, and Julien Gagneur. Deep learning-driven fragment ion series classification enables highly precise and sensitive de novo peptide sequencing. *Nature Communications*, 15(1):151, 2024.

[40] M. R. Lazear. Sage: An open-source tool for fast proteomics searching and quantification at scale. *Journal of Proteome Research*, 22(11):3652–3659, 2023.

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

[42] K. Liu, Y Ye, S. Li, and H. Tang. Accurate de novo peptide sequencing using fully convolutional neural networks. *Nature Communications*, 14(1):7974, 2023.

[43] B. Ma. Novor: Real-time peptide de novo sequencing software. *Journal of the American Society for Mass Spectrometry*, 26:1885–1894, 2015.

[44] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie. PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 17(13):2337–2342, 2003.

[45] 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, 2023.

[46] Damon H. May, Emma Timmins-Schiffman, Molly P. Mikan, H. Rodger Harvey, Elhanan Borenstein, Brook L. Nunn, and William S. Noble. An alignment-free "metapeptide" strategy for metaproteomic characterization of microbiome samples using shotgun metagenomic sequencing. *Journal of Proteome Research*, 15(8):2697–2705, 2016.

[47] R. L. Mayer and F. Impens. Immunopeptidomics for next-generation bacterial vaccine development. *Trends in Microbiology*, 29(11):1034–1045, 2021.

[48] T. Muth, D. Benndorf, U. Reichl, E. Rapp, and L. Martens. Searching for a needle in a stack of needles: challenges in metaproteomics data analysis. *Molecular Biosystems*, 9(4):578–585, 2013.

[49] A. I. Nesvizhskii. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *Journal of Proteomics*, 73(11):2092 – 2123, 2010.

[50] David Ochoa, Andrew F Jarnuczak, Cristina Viéitez, Maja Gehre, Margaret Soucheray, André Mateus, Askar A Kleefeldt, Anthony Hill, Luz Garcia-Alonso, Frank Stein, et al. The functional landscape of the human phosphoproteome. *Nature biotechnology*, 38(3):365–373, 2020.

[51] C. Y. Park, A. A. Klammer, L. Käll, M. P. MacCoss, and W. S. Noble. Rapid and accurate peptide identification from tandem mass spectra. *Journal of Proteome Research*, 7(7):3022–3027, 2008.

[52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., Vancouver, Canada, 2019.

[53] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. PointNet: deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference On Computer Vision and Pattern Recognition*, pages 652–660, 2016.

[54] R. Qiao, N. H. Tran, L. Xin, X. Chen, M. Li, B. Shan, and A. Ghodsi. Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nature Machine Intelligence*, 3:420–425, 2021.

[55] Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Research*, 48(W1):W449–W454, 2020.

[56] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 118(15):e2016239118, 2021.

[57] D. K. Schweppe, J. K. Eng, D. Bailey, R. Rad, Q. Yu, J. Navarrete-Perea, E. L. Huttlin, B. K. Erickson, J. A. Paolo, and S. P. Gygi. Full-featured, real-time database searching platform enables fast and accurate multiplexed quantitative proteomics. *Journal of Proteome Research*, 19(5):2026–2034, 2020.

[58] Omar Shouman, Wassim Gabriel, Victor-George Giurcoiu, Vitor Sternlicht, and Mathias Wilhelm. Prospect: Labeled tandem mass spectrometry dataset for machine learning in proteomics. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[59] Lauren E Stopfer, Joshua M Mesfin, Brian A Joughin, Douglas A Lauffenburger, and Forest M White. Multiplexed relative and absolute quantitative immunopeptidomics reveals MHC I repertoire alterations induced by CDK4/6 inhibition. *Nature Communications*, 11(1):1–14, 2020.

[60] P. Sulimov and A. Kertész-Farkas. Tailor: A nonparametric and rapid score calibration method for database search-based peptide identification in shotgun proteomics. *Journal of Proteome Research*, 19(4):1481–1490, 2020.

[61] J. A. Taylor and R. S. Johnson. Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 11:1067–1075, 1997.

[62] Geoffrey C Teo, Daniel A Polasky, Fei Yu, and Alexey I Nesvizhskii. Fragpipe: Integrated and scalable pipeline for proteomics data analysis. *Nature Methods*, 18:828–830, 2021.

[63] N. H. Tran, X. Zhang, L. Xin, B. Shan, and M. Li. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, 31:8247–8252, 2017.

[64] M. M. VanDuijn, L. J. Dekker, W. F. van Ijcken, P. A. E. S. Smitt, and T. M. Luider. Immune repertoire after immunization as seen by next-generation sequencing and proteomics. *Frontiers in Immunology*, 8:1286, 2017.

[65] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[66] M. Wang, J. Wang, J. Carver, B. S. Pullman, S. W. Cha, and N. Bandeira. Assembling the community-scale discoverable human proteome. *Cell Systems*, 7:412–421.e5, 2018.

[67] Bo Wen, Chris Hsu, Wen-Feng Zeng, Michael Riffle, Alexis Chang, Miranda Mudge, Brook L Nunn, Matthew D Berg, Judit Villen, Michael J MacCoss, et al. Carafe enables high quality in silico spectral library generation for data-independent acquisition proteomics. *bioRxiv*, pages 2024–10, 2024.

[68] Jacob A West-Roberts, Luis E Valentin Alvarado, Susan Mullen, Rohan Sachdeva, Justin Smith, Laura A Hug, Daniel Gregoire, Wentso Liu, Tzu-Yu Lin, Gabriel Husain, et al. Giant genes are rare but implicated in cell wall degradation by predatory bacteria. *bioRxiv*, pages 2023–11, 2023.

[69] R. J. Williams and D. Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989.

[70] Ruitao Wu, Xiang Zhang, Runtao Wang, and Haipeng Wang. Denovo-GCN: De novo peptide sequencing by graph convolutional neural networks. *Applied Sciences*, 13(7), 2023.

[71] Siyu Wu, Zhongzhi Luan, Zhenxin Fu, Qunying Wang, and Tiannan Guo. BiATNovo: A self-attention based bidirectional peptide sequencing method. *bioRxiv*, pages 2023–05, 2023.

[72] H. Yang, H. Chi, W. Zeng, W. Zhou, and S. He. pNovo 3: precise de novo peptide sequencing using a learning-to-rank framework. *Bioinformatics*, 35(14):i83–i90, 2019.

[73] 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.

[74] M. Yilmaz. Noble-lab/casanovo, 2023.

[75] M. Yilmaz, W. E. Fondrie, W. Bittremieux, R. Nelson, V. Ananth, S. Oh, and W. S. Noble. Sequence-to-sequence translation from mass spectra to peptides with a transformer model. *Nature Communications*, 15(1):6427, 2024.

[76] M. Yilmaz, W. E. Fondrie, W. Bittremieux, S. Oh, and W. S. Noble. *De novo* mass spectrometry peptide sequencing with a transformer model. In *Proceedings of the International Conference on Machine Learning*, pages 25514–25522, 2022.

[77] X. Zhou, W. Zeng, H. Chi, C. Luo, C. Liu, J. Zhan, S. M. He, and Z. Zhang. pDeep: predicting MS/MS spectra of peptides with deep learning. *Analytical Chemistry*, 89(23):12690–12697, 2017.

Appendix A

Appendix

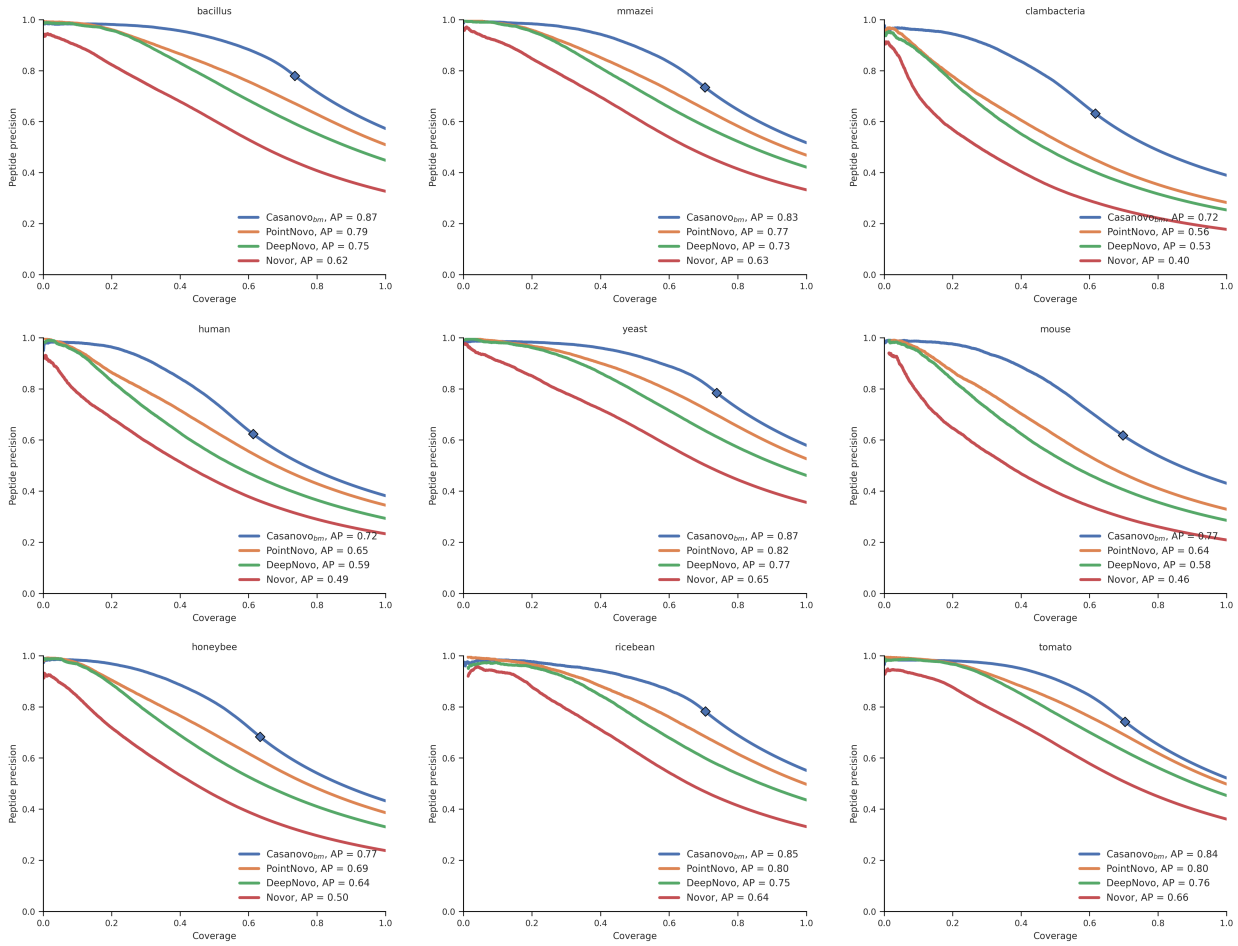


Figure A.1: **Casanovo$_{bm}$ outperforms Novor, DeepNovo, and PointNovo on the original nine species benchmark.** Each panel evaluates the peptide-level performance on the held-out species in the nine species benchmark. For Casanovo$_{bm}$ all peptides that pass the precursor *m/z* filter are ranked above peptides that do not pass the filter, and the boundary is indicated by a diamond on the curve.

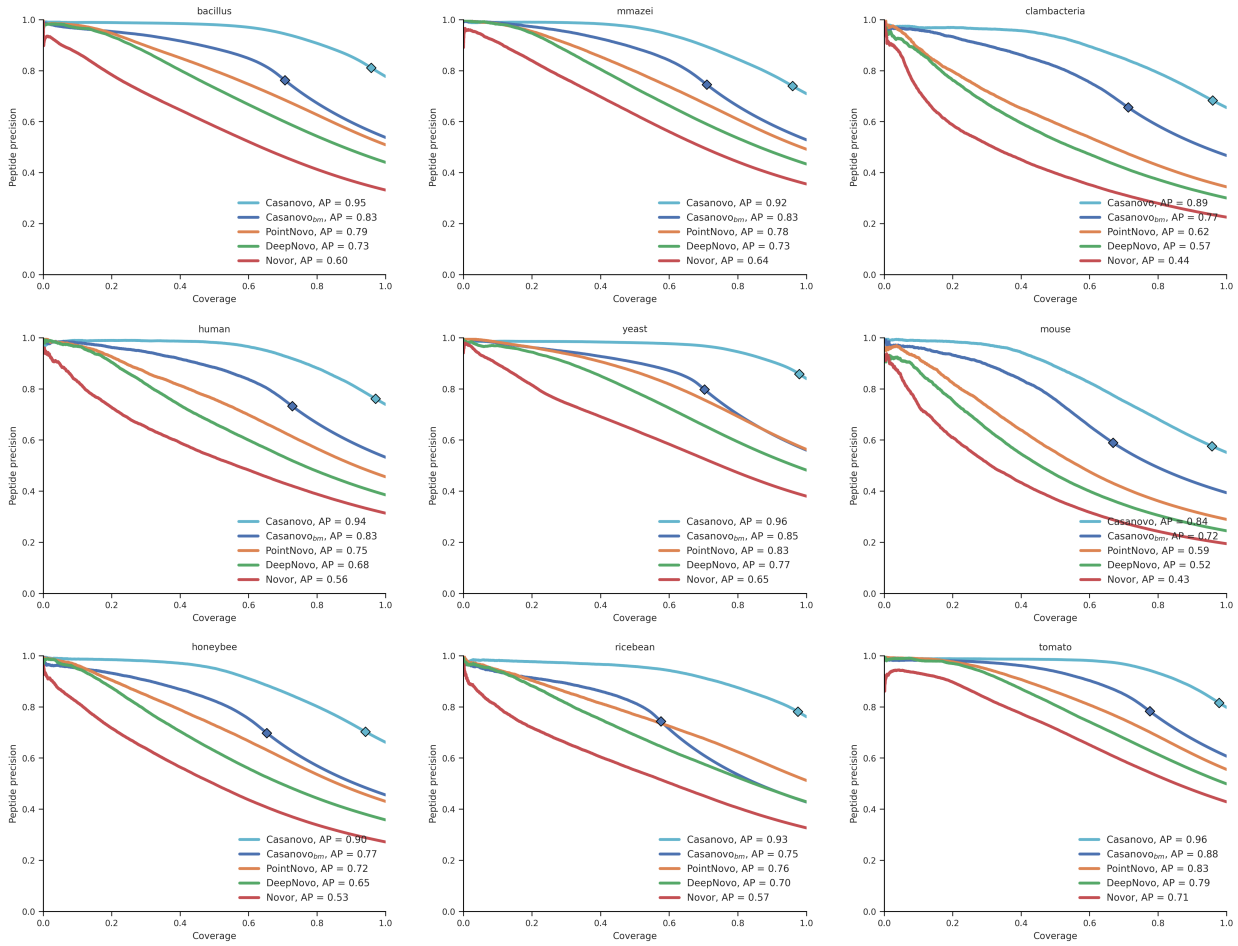
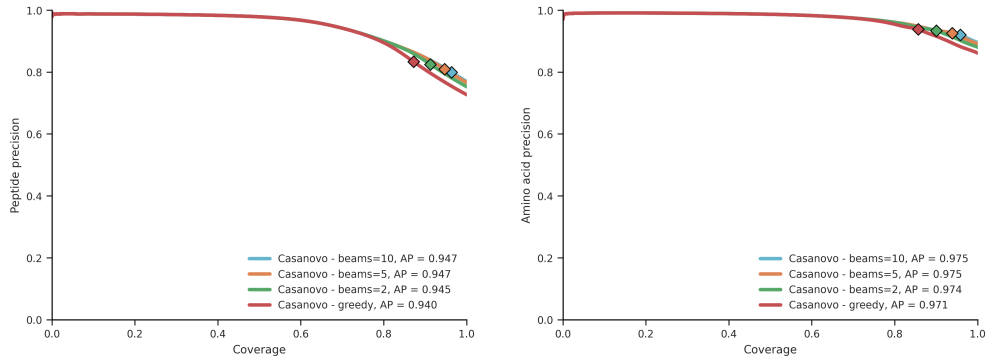


Figure A.2: **Casanovo outperforms Novor, DeepNovo, PointNovo, and Casanovo$_{bm}$ on the revised nine-species benchmark** Each panel evaluates the peptide-level performance on the held-out species in the nine species benchmark. For Casanovo and Casanovo$_{bm}$ all peptides that pass the precursor *m/z* filter are ranked above peptides that do not pass the filter, and the boundary is indicated by a diamond on each curve.



A B

Figure A.3: **Comparison of greedy and beam-search decoding.** (A) The plot shows precision-coverage curves for the Casanovo model, using either greedy decoding or beam-search decoding with different number of beams. The revised 9-species benchmark was used for this analysis. (B) Similar to panel (A), but showing amino acid-level precision and coverage for the same data set.

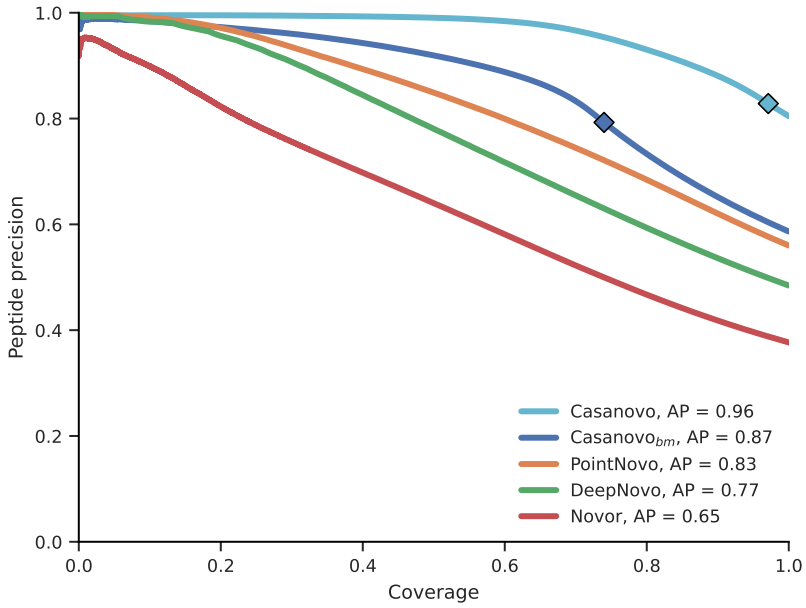


Figure A.4: **Casanovo outperforms Novor, DeepNovo, PointNovo, and Casanovo$_{bm}$ on the revised nine-species benchmark, even when PTMs unavailable to DeepNovo or Novor are eliminated.** The figure plots peptide-level precision as a function of coverage for all species in the nine-species benchmark for Casanovo, DeepNovo and Novor. Casanovo maintains higher peptide-level precision across the full coverage range over Novor, DeepNovo, PointNovo, and Casanovo$_{bm}$ on the aggregated, revised nine-species benchmark. MS/MS spectra associated with modifications that cannot be detected by DeepNovo and Novor are eliminated.

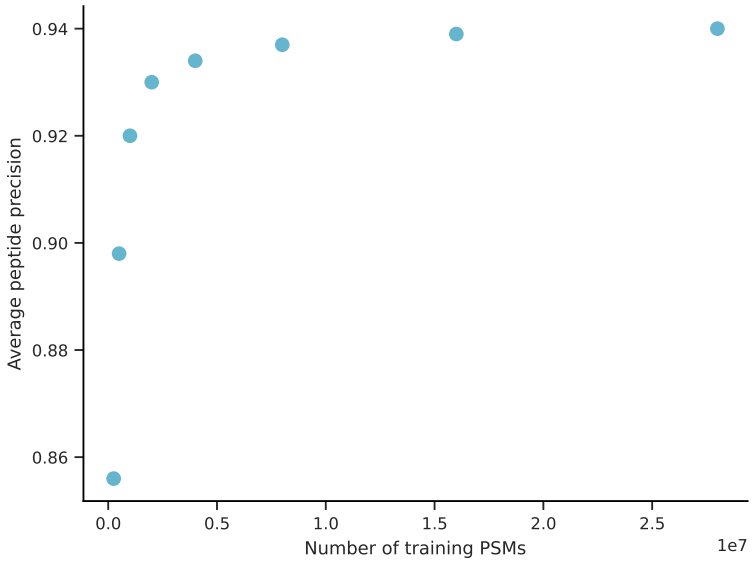


Figure A.5: **Casanovo performance on the 9-species benchmark improves with more training data.** Each point corresponds to a Casanovo model trained on one of the nested subsets of MassIVE-KB, ranging from 250,000 spectra to the full dataset of 28 million spectra. Average precision is reported on the revised 9-species benchmark.

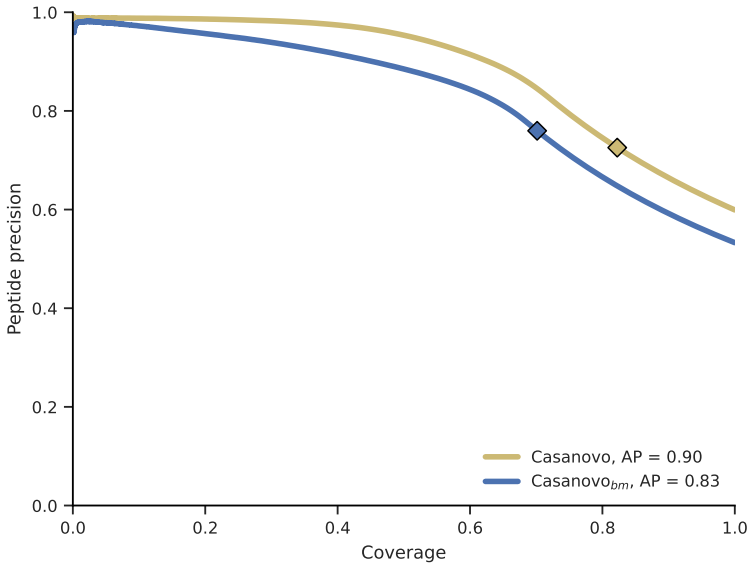


Figure A.6: **Comparison of Casanovo models trained on the 9-species benchmark and MassIVE-KB.** The figure plots, for each model, precision on the revised 9-species benchmark as a function of coverage. The training sets for Casanovo$_{bm}$ and MassIVE-KB trained Casanovo model contain 246,713 and 239,697 distinct peptides, respectively.

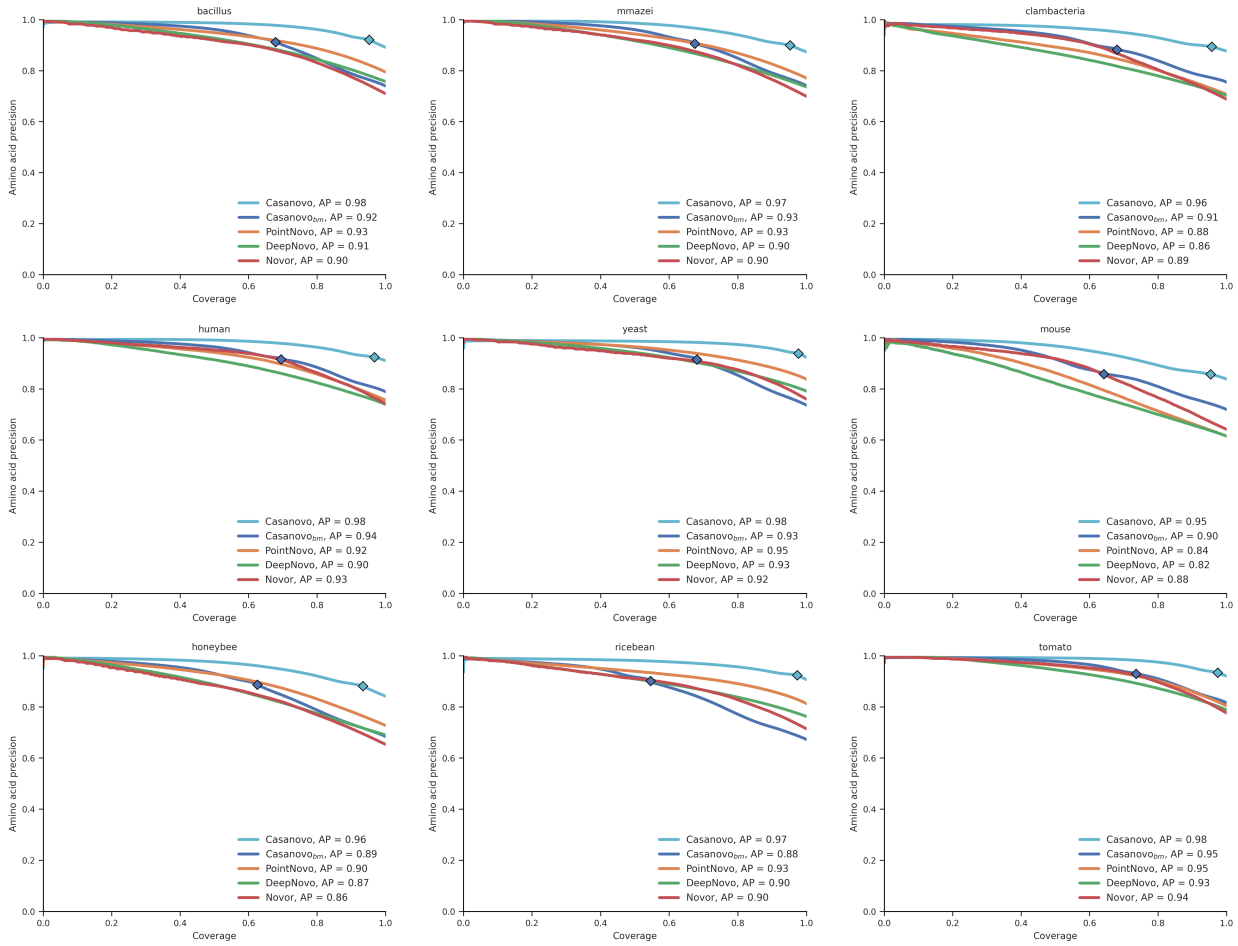


Figure A.7: **Casanovo outperforms Novor, DeepNovo, PointNovo, and Casanovo$_{bm}$ at the amino acid-level on the nine-species benchmark** Each panel evaluates the amino acid-level performance on the held-out species in the nine species benchmark. For Casanovo and Casanovo$_{bm}$ all peptides that pass the precursor m/z filter are ranked above peptides that do not pass the filter, and the boundary is indicated by a diamond on each curve.

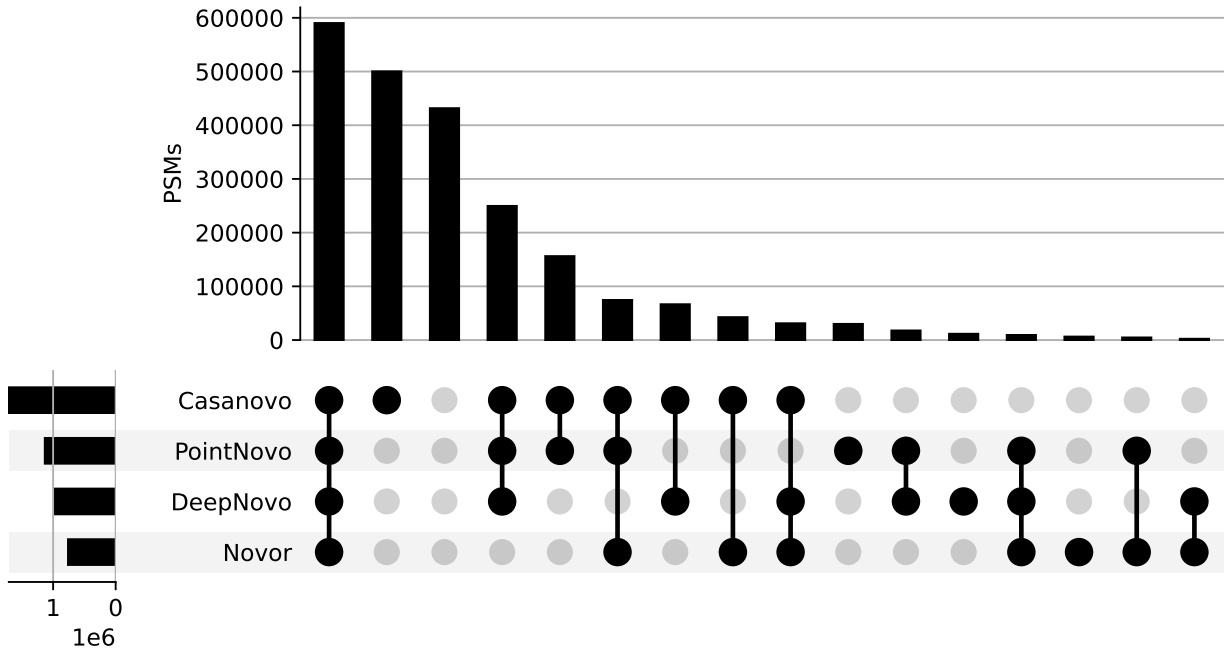


Figure A.8: **Casanovo expands the number of correct PSMs identified by its competitors on the nine-species benchmark.** The plot shows the overlap in peptide predictions between Casanovo and three competing *de novo* sequencing methods for the nine-species benchmark dataset. For each subset of PSMs, black circles denote whether the corresponding method is correct. Horizontal bars indicate the total number of correct PSMs for each method.

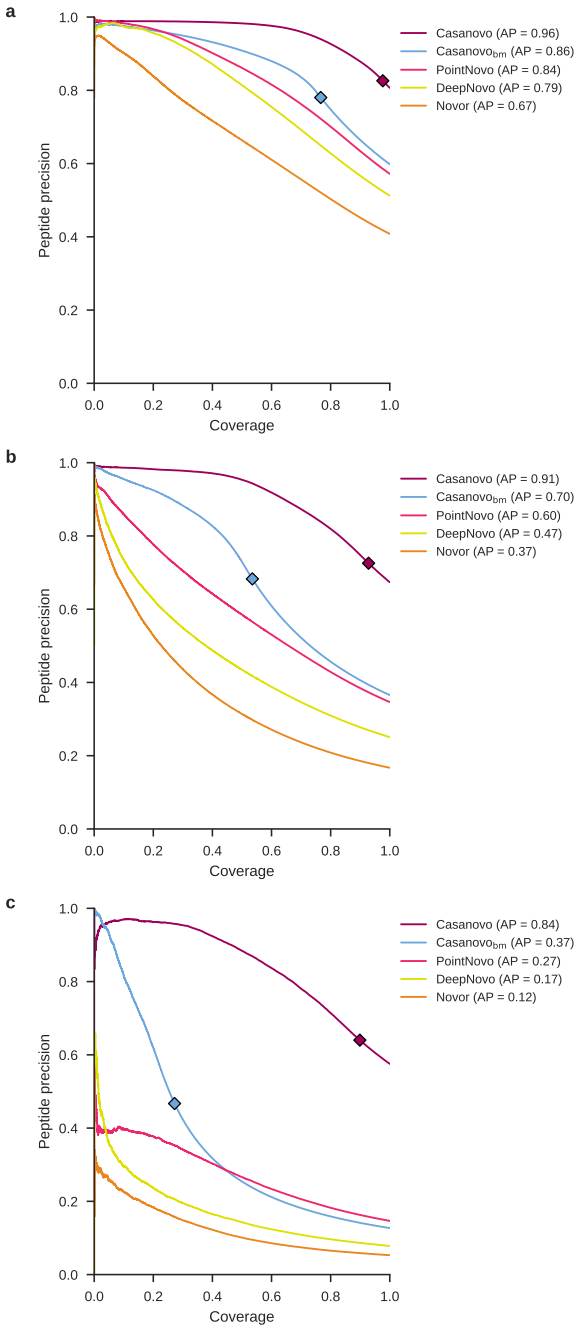


Figure A.9: **Breakdown of *de novo* sequencing performance by charge state.** Plots show peptide precision-coverage curves for subsets of the revised nine-species benchmark, grouped by charge state where panels correspond to spectra with (A) 2+ charge, (B) 3+ charge, (C) 4+ or higher charge.

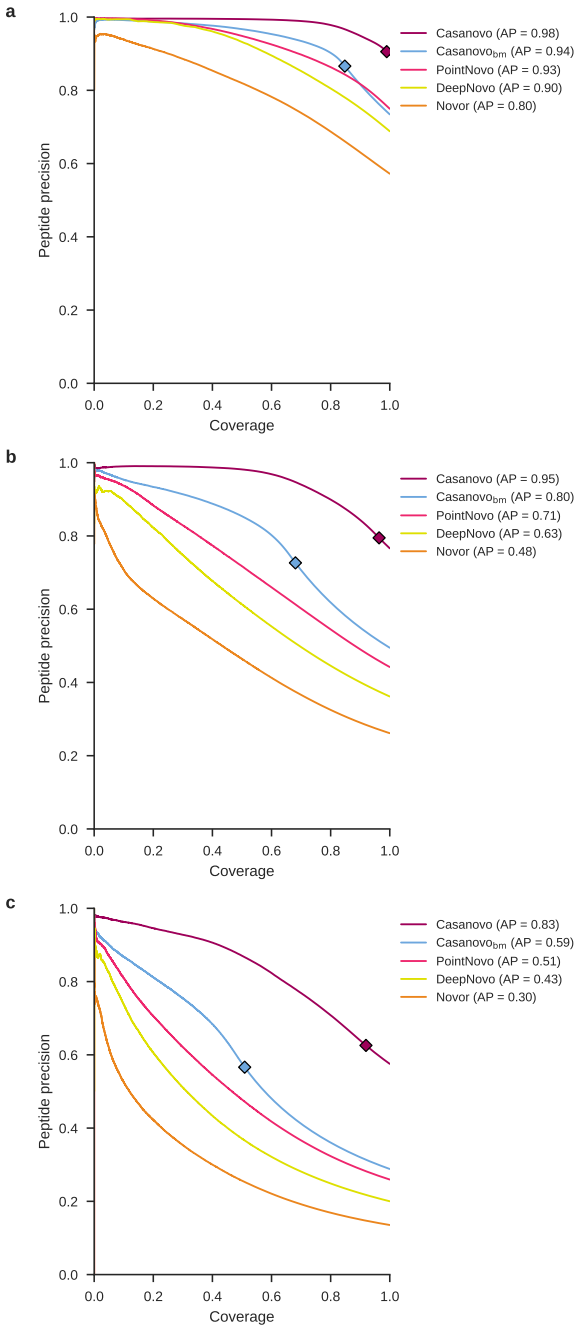


Figure A.10: **Breakdown of *de novo* sequencing performance by peptide length.** Plots show peptide precision-coverage curves for subsets of the revised nine-species benchmark grouped by the length of database search assigned peptides where panels correspond to peptides with (A) fewer than 13 amino acids (B) between 13 and 18 amino acids, (C) greater than 18 amino acids.

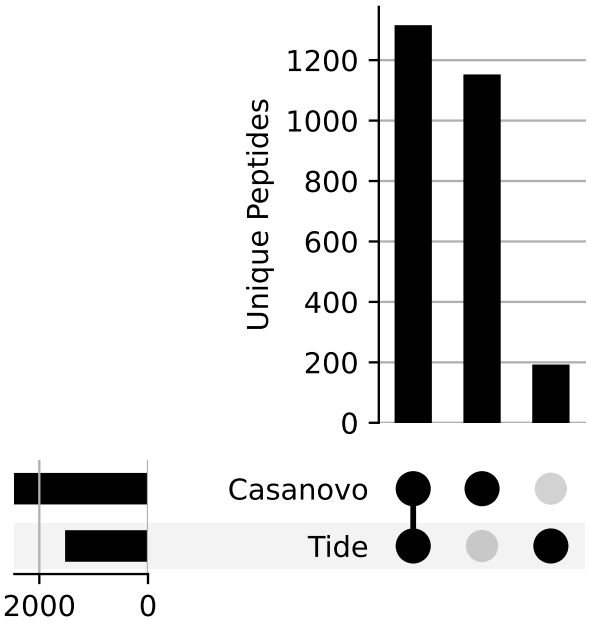


Figure A.11: **Casanovo identifies a greater number of immunopeptides than Tide database search.** The plot shows the overlap between unique peptides assigned by Casanovo that match to the human proteome and by Tide at 1% FDR for the immunopeptidomics dataset.

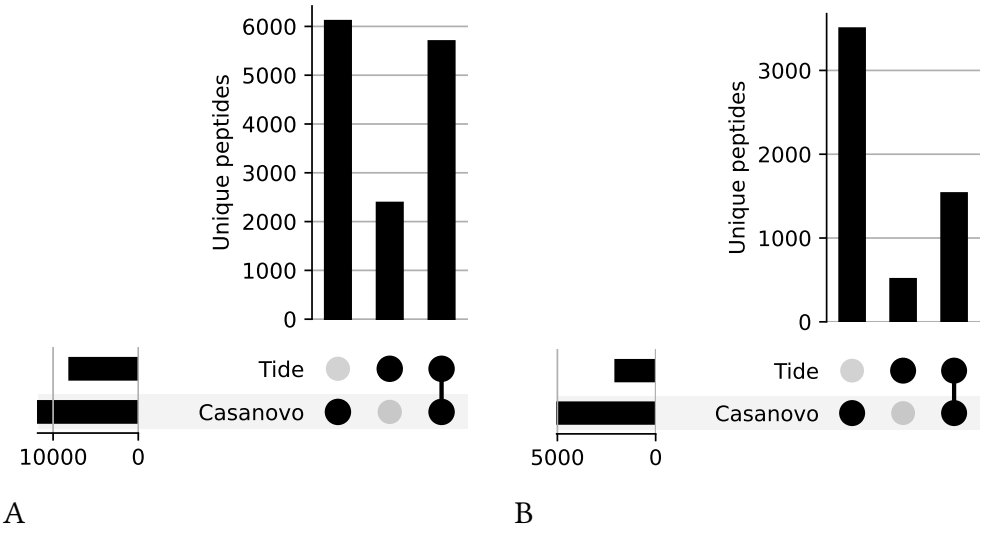


Figure A.12: **Casanovo detects a substantial number of additional peptides compared to Tide database search in metaproteomics samples.** (A) The plot shows the overlap between unique peptides assigned by Casanovo at 1% error rate and Tide at 1% FDR when respective metapeptide databases are used for the Bering Sea and the Chukchi Sea datasets. (B) Similar to panel (A), but the non-redundant environment database is used for searching and error rate control.

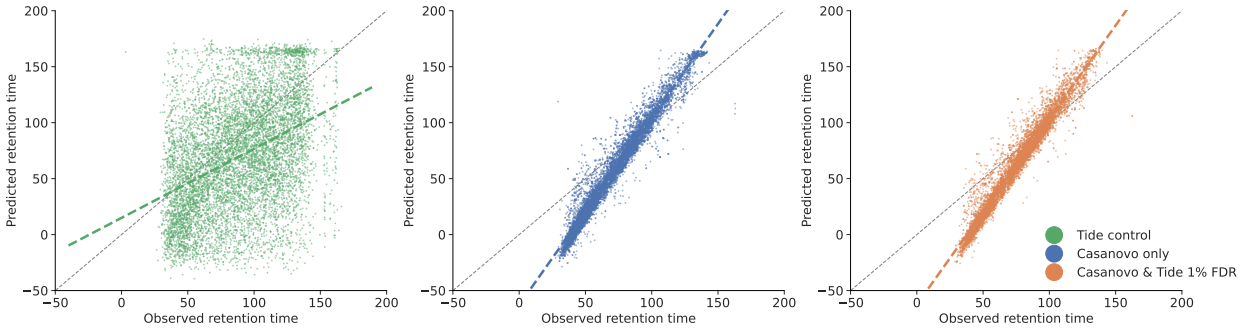


Figure A.13: **Confident Casanovo assignments have retention times that are highly correlated with Prosit predictions.** We compared the observed retention times against Prosit-predicted retention times for three groups of peptides: (1) peptides predicted only by Casanovo that match to the relevant database at 1% error rate, (2) peptides identified by both Casanovo and database search with a 1% FDR threshold, and (3) peptides identified by database search with FDR >10%. The dashed line in each plot represents the ordinary least squares linear regression.

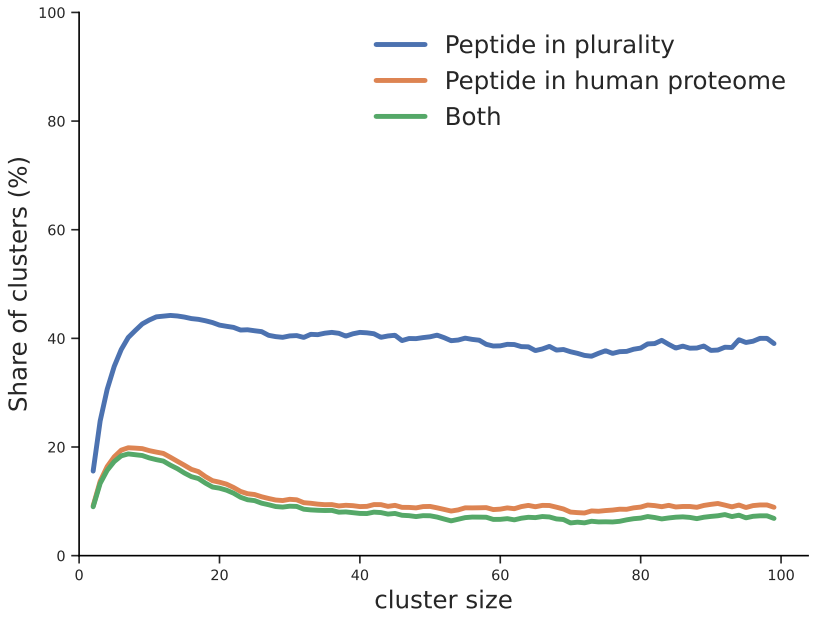


Figure A.14: **Casanovo assigns new peptides to dark matter clusters of various sizes.** We evaluated the fraction of clusters which have a Casanovo peptide in plurality, have a unique Casanovo peptide matching to human proteome or satisfy both criteria among all previously unidentified clusters larger than a given size.

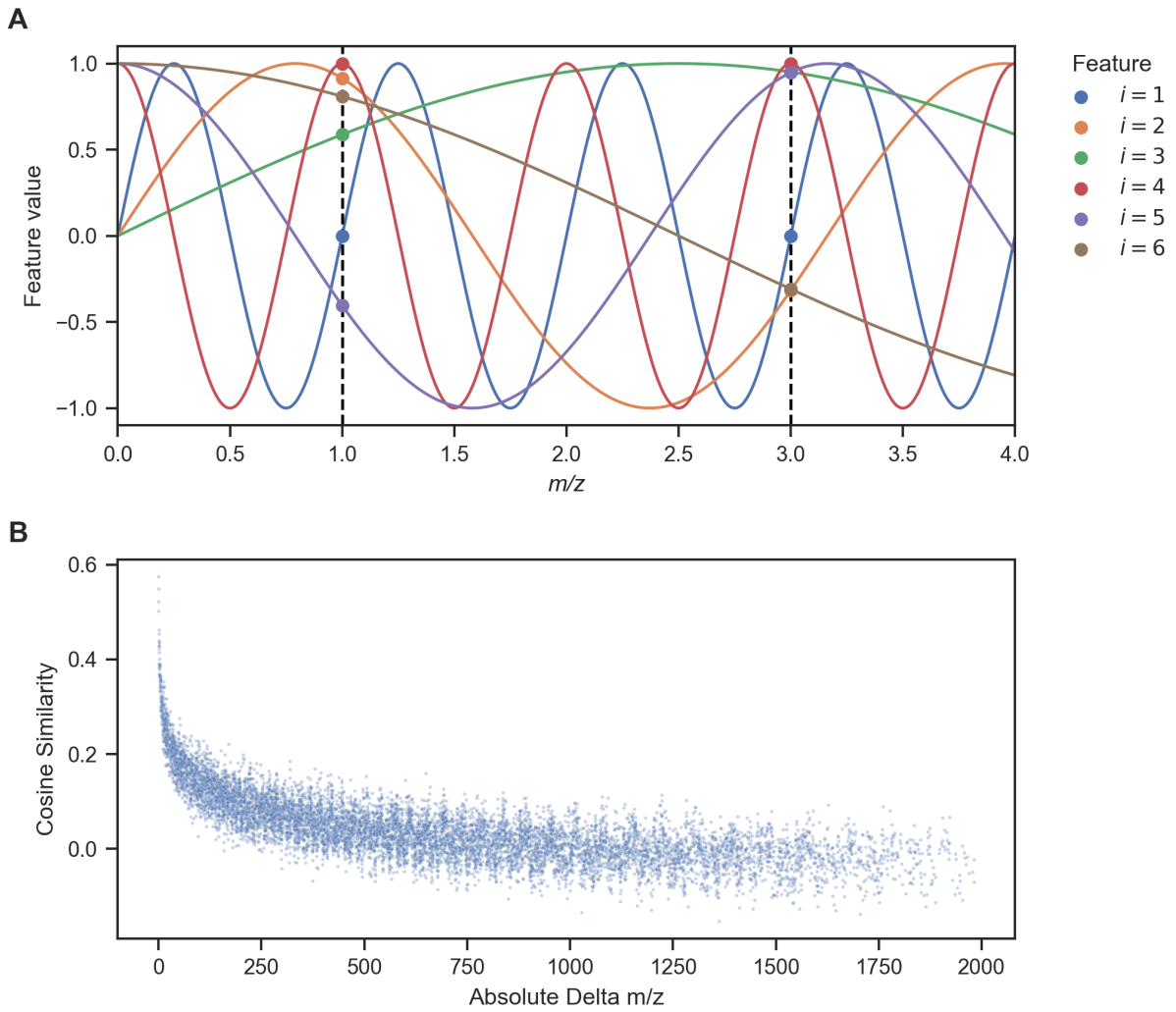


Figure A.15: **Sinusoidal encodings represent *m/z* distance between peaks in a mass spectrum. (A)** The *m/z* value of each peak is encoded from a progression of sinusoids defined by a minimum and maximum wavelength. In this example, a 6-dimensional embedding ($d = 6$) of *m/z* 1.0 and *m/z* 3 is created from sinusoids ranging from a wavelength of *m/z* 1 ($\lambda_{\text{min}} = 1$) to 10 ($\lambda_{\text{max}} = 10$)to demonstrate how the encoding is performed. **(B)** The Casanovo sinusoidal embeddings are 512-dimensional ($d = 512$) and created from sinusoids ranging from *m/z* 0.0001 ($\lambda_{\text{min}} = 0.0001$) to 10,000 ($\lambda_{\text{min}} = 10,000$). The utility of these embeddings lies in their preservation of *m/z* distance in their embedded space. Here, we sample 10,000 pairs of *m/z* values between *m/z* 0 and 2000. The cosine similarity between these embeddings is negatively correlated with the original distance between *m/z* values.

	MassIVE-KB count	MassIVE-KB selected	PROSPECT count	PROSPECT selected
A	24910	24910	540777	25090
C	11473	11473	90772	38527
D	44654	44654	267716	5346
E	42977	42977	603469	7023
F	38183	38183	884261	11817
G	21640	21640	344670	28360
H	166398	50000	357268	0
I	12984	12984	437698	37016
K	16289160	50000	960538	0
L	50002	50000	2264210	0
M	19405	19405	388750	30595
N	40867	40867	294550	9133
P	9685	9685	114504	40315
Q	33572	33572	542063	16428
R	13589301	50000	1135502	0
S	28389	28389	457519	21611
T	15845	15845	443535	34155
V	26011	26011	774624	23989
W	4261	4261	341456	45739
Y	35180	35180	1368308	14820
Total	30504897	610036	12612190	389964

Table A.1: **Creating a non-enzymatic dataset by sampling from PROSPECT and MassIVE-KB.** PROSPECT was first downsampled to include at most 100 PSMs per peptide sequence. MassIVE-KB and PROSPECT were then segregated by C-terminal amino acid, and we randomly selected from each category from MassIVE-KB, supplementing as necessary from PROSPECT to obtain 50,000 PSMs per terminal amino acid.

PRIDE	Species	Uniprot	Files	Spectra	PSMs	Peptides	precursor	fragment
PXD005025	*Vigna mungo*	UP000087766	24	932848	108514	12001	20	0.05
PXD004948	*Mus musculus*	UP000000589	13	306786	25541	5899	10	0.05
PXD004325	*Methanosarcina mazei*	UP000033058	72	3728183	267333	15925	10	0.05
PXD004565	*Bacillus subtilis*	UP000001570	106	4336428	1358337	30786	30	0.05
PXD004536	*Candidatus endoloripes*	UP000094849	11	2272023	82290	8392	20	0.05
PXD004947	*Solanum lycopersicum*	UP000004994	60	603506	178413	49745	15	0.05
PXD003868	*Saccharomyces-cerevisiae*	UP000002311	27	1477397	585593	19720	20	0.05
PXD004467	*Apis mellifera*	UP000005203	17	823169	194281	21559	20	0.05
PXD004424	*H. sapiens*	UP000005640	26	684821	44604	11289	20	0.02
Total			343	15,165,161	2,844,906	175,316		

Table A.2: **The nine-species benchmark.** The final two columns specify the precursor window size (in ppm) and fragment bin size (in Da) used in the database search step. No reference proteome is available for *Vigna mungo*, so the proteome for the closely related species *Vigna radiata* was used instead.

In human proteome	In Casanovo peptide	BLOSUM score	Count
L	V	1	749
V	L	1	650
E	Q	2	467
N	D	1	437
R	K	2	371
E	H	0	343
E	D	2	338
L	K	-2	314
L	M	2	304
L	F	0	301

Table A.3: **The most common amino acid swaps have positive BLOSUM scores.** We found that the top ten most common single amino acid substitutions that can be explanied with a single nucleotide polymorphism detected by Casanovo are enriched for positive BLOSUM scores.

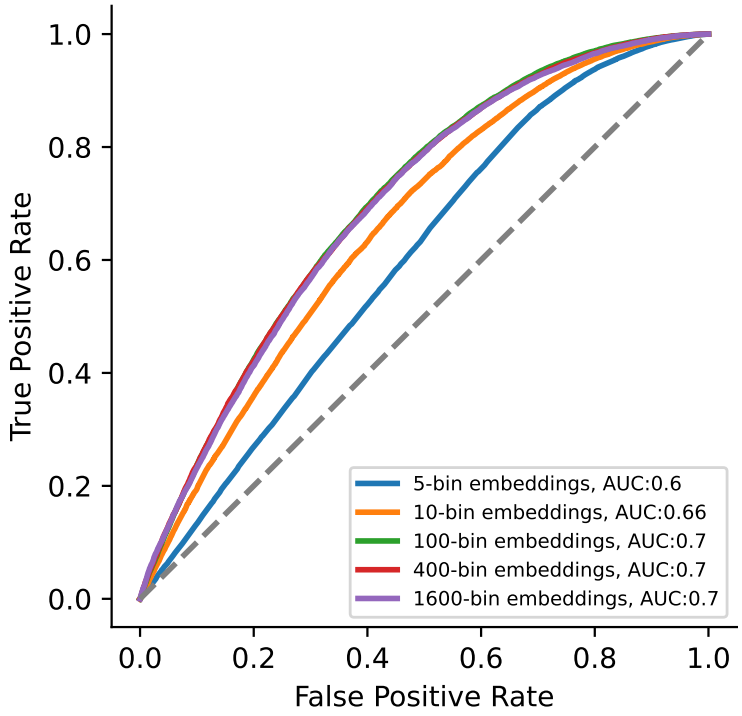


Figure A.16: **Comparison of binned embedding performance at different binning resolutions on the chimericity prediction.** ROC curves and the area under the curve (AUC) are reported by the number of bins used to represent peak intensities for the chimericity prediction validation set.

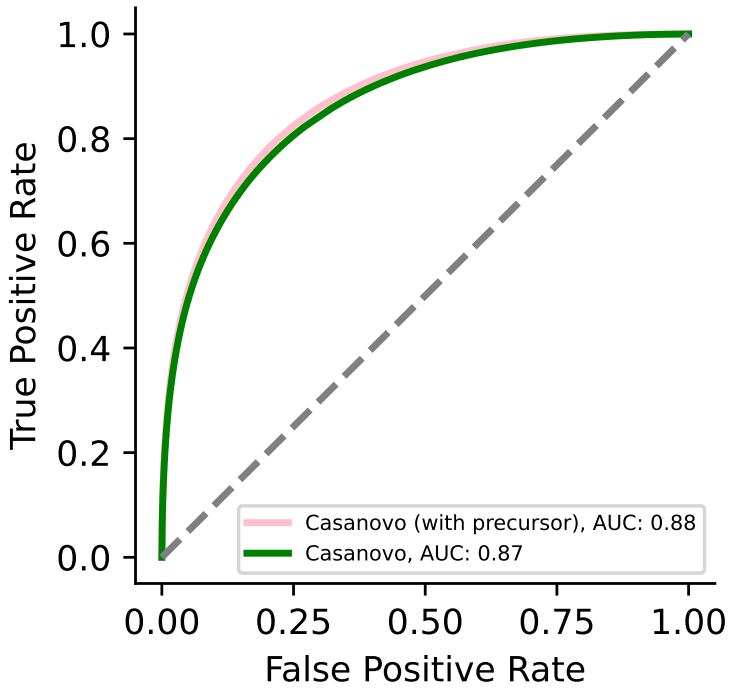


Figure A.17: **Comparison of the original and precursor information added Casanovo pipelines on the PTM detection task.** ROC curves and the area under the curve (AUC) reported for the original Casanovo embeddings and a modified version with precursor *m/z* and charge concatenated.

Dataset	n_non-phospho	n_phospho	Bacc-AHLF	Bacc-Casanovo$_{nopre}$	F1-AHLF	F1-Casanovo$_{nopre}$	AUROC-AHLF	AUROC-Casanovo$_{nopre}$
OVAS	90936	37720	0.95	0.93	0.92	0.88	0.99	0.98
TOV-21-Primary	62350	26978	0.94	0.93	0.92	0.87	0.99	0.98
ES2-Primary	16297	6667	0.94	0.93	0.91	0.82	0.99	0.98
Daudi	150915	210916	0.89	0.94	0.90	0.95	0.96	0.99
U2OS	92329	205353	0.77	0.85	0.90	0.91	0.95	0.94
HaCaT	19216	113775	0.78	0.87	0.95	0.94	0.93	0.95
HT-29	1625	27531	0.72	0.98	0.97	0.99	0.92	1.00
HeLa	1469194	2949614	0.83	0.88	0.89	0.92	0.92	0.95
HEPG2	426	45416	0.76	0.98	0.98	0.99	0.92	1.00
A549	4068	172792	0.82	0.94	0.92	0.98	0.91	0.99
Colon	8359	28798	0.81	0.91	0.90	0.94	0.90	0.96
Primary-Gastro	22026	219767	0.72	0.84	0.94	0.93	0.88	0.93
LNCaP	53851	5200	0.78	0.94	0.45	0.69	0.87	0.99
RPMI-8226	1184	413	0.76	0.99	0.65	0.99	0.87	1.00
HEK293	322811	332690	0.77	0.90	0.73	0.91	0.86	0.97
Primary-Prostate	9223	100617	0.77	0.94	0.89	0.95	0.86	0.98
Primary-AML	494184	5893	0.72	0.95	0.29	0.75	0.85	0.99
Kasumi-1	2294	29470	0.75	0.98	0.88	0.99	0.85	1.00
HPAC	594	807	0.67	0.96	0.56	0.96	0.78	1.00
SU.86.86	909	984	0.65	0.97	0.53	0.97	0.77	1.00
CFPAC-1	999	780	0.64	0.98	0.52	0.97	0.76	1.00
PANC-05-04	1079	1426	0.65	0.96	0.55	0.96	0.74	0.99
PANC-02-03	273	815	0.65	0.96	0.56	0.97	0.72	1.00
OVSAYO	11515	28	0.56	0.65	0.02	0.02	0.69	0.83
HDMVEC	4320	2961	0.59	0.90	0.40	0.87	0.60	0.97

Table A.4: **Comparison of non-pre-trained Casanovo encoder (Casanovo$_{nopre}$) and AHLF across PTM detection datasets.** 26 datasets listed correspond to the holdout split *a* described in [2] and AHLF results are directly taken from the paper. The first two columns indicate the number of non-phosphorylated versus phosphorylated PSMs in each dataset. The performance metrics are balanced accuracy, F1 score and AUROC.