A Foundation Model for Mass Spectrometry Proteomics

Justin Sanders [*1] **Melih Yilmaz** [*1] **Jacob H. Russell** [2] **Wout Bittremieux** [3] **William E. Fondrie** [4]
Nicholas M. Riley [2] **Sewoong Oh** [1] **William Stafford Noble** [1 5]

Abstract

Mass spectrometry is the dominant technology in the field of proteomics, enabling high-throughput analysis of the protein content of complex biological samples. Due to the complexity of the instrumentation and resulting data, sophisticated computational methods are required for the processing and interpretation of acquired mass spectra. Here, we propose unifying various spectrum prediction tasks under a single foundation model. To this end, we pre-train a spectrum encoder using *de novo* sequencing as a pre-training task. We then show that using these pre-trained spectrum representations improves our performance on the four downstream tasks of spectrum quality prediction, chimericity prediction, phosphorylation prediction, and glycosylation status prediction, demonstrating that our foundation model has learned generalizable representations of mass spectra.

1. Introduction

In recent years, foundation models have emerged as a powerful machine learning paradigm for various problem domains [26, 27, 6]. These models are trained to learn rich latent representations of input modalities from large datasets of unlabeled or weakly labeled data using pre-training 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, with a relatively small amount of supervised fine-tuning a foundation model will outperform its peers trained only with supervision.

Motivated by the success of foundation models in language, vision, and multi-modal tasks, we develop a foundation model for tandem mass spectrometry proteomics. For many prediction tasks, insufficient training data and noisy training labels make it challenging to learn a rich understanding of mass spectra in isolation for each task. We hypothesized that learned spectrum embeddings, pre-trained on a large dataset of high-confidence spectrum annotations, may prove a valuable starting point for various downstream tasks.

2. Background and related work

Currently, tandem mass spectrometry is the only high-throughput method for systematically analyzing the full protein content of biological samples [21]. In a standard tandem mass spectrometry experiment, proteins are digested into short peptides, ionized, and fragmented. The mass-to-charge ratios (*m/z*) of the resulting fragment ions are then measured very precisely by the instrument. This process yields a list of "peaks," each representing the *m/z* of a specific ion along with an intensity value corresponding to its abundance. Together, this list of peaks is called an "MS/MS spectrum," which serves as a fingerprint of the specific analyte being measured. In a typical mass spectrometry run, the instrument will collect on the order of 100,000 such spectra, each corresponding to a distinct analyte. Canonically, these spectra are then processed by a database search algorithm, with the goal of assigning to each spectrum its generating peptide. However, there are a variety of other important downstream tasks involving tandem mass spectra.

De novo sequencing An alternative strategy to database search for spectrum annotation is *de novo* peptide sequencing. *De novo* sequencing aims to predict the generating peptide for a given spectrum without relying on prior knowledge, making it a valuable tool for identifying peptides not present in a pre-defined protein database. Algorithms for solving this problem were introduced in the late 1990's [34] and it was first solved using machine learning in 2015 [19] and deep learning in 2017 [35]. More recently, Casanovo [40] employed a transformer architecture to frame *de novo* sequencing as a sequence-to-sequence translation task.

[1] Paul G. Allen School of Computer Science, University of Washington [2] Department of Chemistry, University of Washington [3] Adrem Data Lab, University of Antwerp [4] Talus Bioscience [5] Department of Genome Sciences, University of Washington. Correspondence to: Justin Sanders <jsander1@cs.washington.edu>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025. Copyright 2025 by the author(s).

Downstream tasks. We study four downstream tasks that take tandem mass spectra as input: predicting spectrum quality, chimericity, phosphorylation, and glycosylation status.

In spectrum quality prediction, the model is asked to predict whether an observed spectrum is identifiable, meaning that it shows strong and clear signal for a peptide. This problem has been addressed with a variety of classical machine learning techniques [25, 30, 38, 20] and more recently using a convolutional neural network [12].

In mass spectrometry experiments, acquired spectra often inadvertently contain signal from multiple peptides. Such spectra are called chimeras, and they can be hard to analyze due to the mixture of signals from each peptide. To our knowledge, prediction of chimeric spectra has not previously been solved using machine learning methods. However, many existing methods generalize the database search procedure to allow for chimeric matches [42, 11].

Predicting whether a post-translational modification is present in a given spectrum is another key task. PhoStar uses a random forest to predict whether a given spectrum was generated by a phosphorylated peptide based on a set of hand-designed features [10]. AHLF improves on this using a convolutional model which takes as input a full spectrum [2]. For predicting whether a spectrum contains a peptide which is N- or O-glycosylated, current methods rely on hand-designed rules based on specific fragment ions [33].

Learning representations of spectra. Prior work has investigated learning spectrum representations, but have primarily focused on dimensionality reduction for clustering spectra and improving peptide identification. GLEAMS learns low-dimensional spectrum representations such that spectra from the same peptide cluster together [5]. Similarly, yHydra co-embeds peptides and spectra such that spectra are close to their generating peptides in embedding space [1].

Finally, prior work has investigated foundation models for tandem mass spectra in the metabolomics space. The methods LSM1-MS2 [4], PRISM [14], and DreaMS [7] use unsupervised masked-peak modeling to learn representations of metabolomics mass spectra, demonstrating that these representations improve performance on downstream chemical property prediction tasks.

3. *De novo* peptide sequencing as a pre-training task

To accurately perform *de novo* sequencing, a model needs to capture the fundamental relationships between the analyte present in the instrument (i.e., the peptide) and the observed signal measured by the mass spectrometer. This in turn requires a rich understanding of the physics and chemistry governing peptide chromatography, ionization, and fragmentation. We hypothesize that this prior understanding of mass spectra, acquired through pre-training on the *de novo* sequencing task, will generalize to other tasks involving mass spectra for which less training data is available.

Typically, foundation models are trained in an unsupervised manner, so as to benefit from massive datasets of unlabeled training examples. However, unlike the settings of natural language processing and computer vision, where there are orders of magnitude more unlabeled training examples than labeled samples, typically 40–60% of acquired spectra can be annotated with their generating peptide in a given mass spectrometry run. Additionally, this labeling is fully automated and high-throughput, with no need for costly human annotations. Thus, here we consider making use of these labels to explore the supervised task of *de novo* peptide sequencing as pre-training for a foundation model.

In this work we perform experiments with a state-of-the art, transformer-based *de novo* sequencing model, Casanovo [40, 41]. Casanovo is trained on a dataset of 30 million high-quality labeled tandem mass spectra from the MassIVE-KB spectral library [36]. We use Casanovo's pre-trained spectrum encoder off the shelf as a foundation model for mass spectrometry proteomics.

4. Downstream tasks

We use our foundation model as a starting point to address a series of downstream tasks. In each task, we compare the frozen Casanovo encoder coupled with a small task-specific dense predictor head ("Casanovo Foundation") against at least two baselines. First, we bin spectrum peaks along the *m/z* axis to obtain spectrum embeddings and then train a gradient boosted decision tree classifier directly on those embeddings ("binned embedding"). Second, we train a transformer spectrum encoder, which has the same architecture as Casanovo, along with a dense classifier head from scratch to learn the downstream tasks end-to-end ("end-to-end transformer").

4.1. Spectrum quality prediction

The first downstream task we consider is spectrum quality prediction. For this task, the goal is to predict whether a given observed MS/MS spectrum will be successfully annotated by database search. To create a labeled dataset for this task, we run databsase search on 20 high-resolution human mass spectrometry runs from MassIVE. Spectra that are matched to a peptide under 1% false discovery rate (FDR) are labeled as high quality, whereas spectra that failed to be matched are annotated as low quality for the binary classification task.

Applying Casanovo Foundation to this task, we achieve an AUROC of 0.820, outperforming our task-specific end-

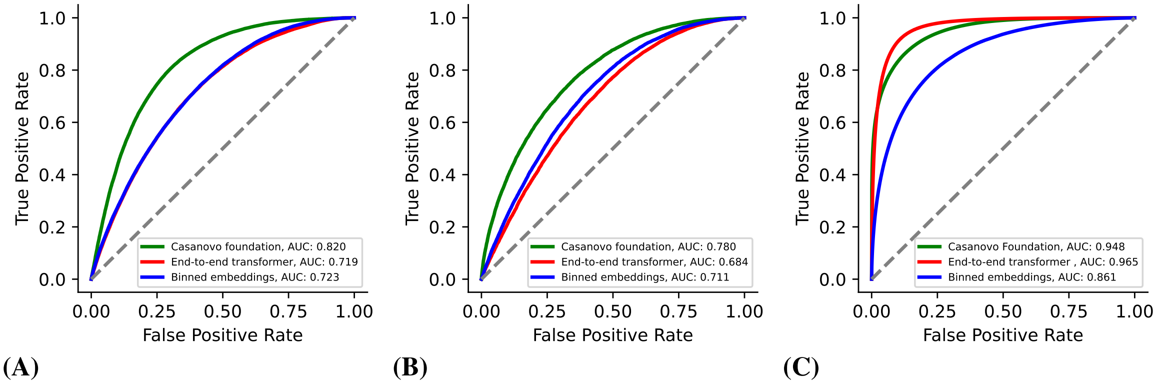

Figure 1. **Downstream tasks.** ROC curves and the area under the curve (AUC) reported for Casanovo foundation and baseline methods on the tasks of (A) spectrum quality prediction, (B) chimericity prediction, and (C) phosphorylation prediction.

to-end transformer and the binned embedding baselines (AUROC of 0.719 and 0.723, respectively) (Figure 1A). This result suggests that the pre-trained spectrum representations from Casanovo capture properties that are hard to learn from the quality prediction task alone. This is not too surprising, given that the *de novo* sequencing pre-training task is both an inherently richer task and took advantage of more data.

4.2. Spectrum chimericity prediction

Tandem mass spectrometry experiments are designed to attempt to isolate individual peptide species for fragmentation. 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. Accordingly, our second downstream task involves detecting when more than one peptide species is responsible for generating a given MS/MS spectrum, i.e., predicting whether it is chimeric or not. Many existing methods generalize the database search procedure to allow chimeric matches [42, 11]; however, prediction of chimeric spectra has not previously been solved using machine learning methods. Such a predictor would be useful in deciding which spectra to provide as input to one of the tools above or in adjusting the settings of an instrument to avoid unwanted chimeras.

To train a chimericity predictor, we use spectra from human, mouse, and yeast samples for training, validation, and test, respectively. Database search is performed using the wide-window setting in FragPipe [23], which allows spectra to be assigned multiple peptides. For the binary classification task, spectra assigned more than one peptide are labeled chimeric and spectra annotated with a single peptide are labeled non-chimeric. Comparing Casanovo Foundation to the baseline methods, we again see that it achieves improved performance (AUROC 0.780) compared to the two baselines (AUROC of 0.684 and 0.711) (Figure 1B).

4.3. Post-translational modification detection

The final type of downstream task we consider is the detection of spectra generated by peptides containing PTMs. A PTM is a molecular group that attaches to the side-chain of one of the amino acids in a peptide. Successfully identifying peptides carrying a PTM requires specific considerations in both how the mass spectrometry experiment is conducted and how the resulting data is analyzed. Thus, a model capable of identifying which PTMs are associated with a given MS/MS spectrum would be valuable in guiding both data collection and analysis. In fact, simple methods for solving this task are regularly employed in practice to improve the sensitivity and quantitative accuracy of experiments targeting peptides carrying a specific PTM [33, 18]. Here, we train classifiers to recognize two common types of PTMs.

Phosphorylation detection We first consider the detection of spectra from phosphorylated peptides. Protein phosphorylation is arguably one of the most important and well studied PTMs, and is responsible for driving key physiological activities such as energy metabolism, cell proliferation and growth, apoptosis, and signal transduction [3]. We frame phosphorylation prediction as a binary classification task, predicting whether or not a given spectrum derives from a phosphorylated peptide.

To train a classifier, we use 19.2 million labeled spectra from the human phosphoproteome dataset [22] which were used to train AHLF [2], a state-of-the-art phosphorylation predictor. Comparing the ROC curves for Casanovo Foundation (AUROC 0.948) to our baselines for this task, we observe that it performs better than the binned spectrum baseline (AUROC 0.861). Additionally, it outperforms the reported performances of AHLF (AUROC 0.921) and PhoStar (AUROC 0.917) on the dataset as a whole [2] and when broken down by cell type (Supplementary Figure S1). However, it performs worse than the end-to-end transformer model (AU-

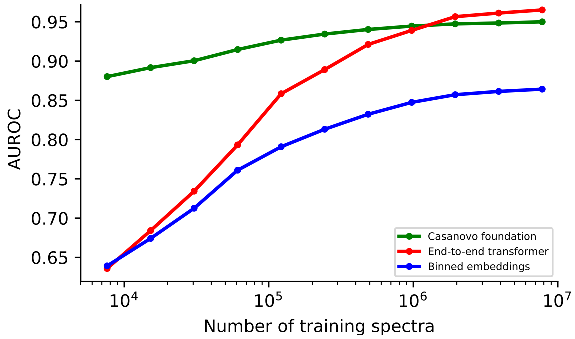

Figure 2. **Learning Curve**. Learning curve showing the performance of Casanovo Foundation and baselines on the phosphorylation prediction task when trained on datasets of varying size.

ROC 0.965) (Figure 1C). This result is not too surprising, because foundation modeling is not expected to provide a major advantage on tasks with very large amounts of high-quality labeled data available for training.

For PTMs other than phosphorylation, such a large training set is unavailable and foundation modeling approach may prove more valuable. Accordingly, to investigate the relationship between the number of training samples and the performance of each model, we create a series of 10 nested subsets of the phosphoproteomics training data ranging in size from 7,700 to 7.7 million training spectra. We find that for datasets with fewer than ∼1 million spectra, the relative performance of our Casanovo Foundation model and the end-to-end transformer baseline cross over (Figure 2). The difference in performance grows as the size of the training set decreases. Strikingly, Casanovo Foundation achieves an AUROC of 0.881 when trained on a dataset of just 7,615 spectra, compared to 0.635 for the end-to-end transformer.

Glycosylation determination. To further explore the task of PTM prediction, we turn to another important modification for which training data is less readily available. Protein glycosylation is a complex PTM where various combinations of mono- and oligosaccharides are attached to specific residues. Here, we consider the task of predicting the glycosylation class of a peptide from its spectrum. The two most common classes of glycosylation are N-glycosylation, where glycans are attached to asparagine residues, and O-glycosylation, where glycans are attached to serine or threonine residues [13, 31, 17]. Recognizing whether a given spectrum represents a glycosylated peptide is straightforward due to the presence of characteristic oxonium ions [39, 29, 32, 43]. However, distinguishing N-glycosylation from O-glycosylation is more difficult. Nevertheless, effectively classifying N- vs O-glycopeptides during data acquisition is critical for the sensitivity and throughput of glycoproteomics experiments (Supplementary Note S3.5).

To train a model to distinguish N- versus O-glycsolyation, we use a publicly available dataset of the mouse brain glycoproteome [24]. This dataset contains 252,970 total glycopeptide identifications, of which 25,757 (10.2%) are O-glycosylated. In addition to our two standard baselines, for this task we also consider two domain-specific baselines. The first looks at the ratio in intensity between the oxonium ion at 138 *m/z* to that at 144 *m/z*. This ratio between the abundances of expected product ions from N- and O-glycans is currently used in practice for real-time prediction in glycoproteomics experiments [33]. The second baseline is a slightly more sophisticated version of the prior approach, which trains an XGBoost classifier on the abundance of 54 oxonium ions, extracted by GlyCounter [16], that are known to be characteristic of glycosylation status.

Given the significant class imbalance in the data, we evaluate the performance of each method based on the area under the precision-recall curve (AUPR). We find that the domain-specific baselines are already reasonably good, with an AUPR of 0.753 for the 138/144 ratio and 0.860 for GlyCounter+XGBoost. The binned embedding baseline and end-to-end transformer baselines are not much better, achieving AUPRs of 0.811 and 0.867, respectively. However, we again find that Casanovo Foundation offers the best results, achieving an AUPR of 0.914 (Figure 3).

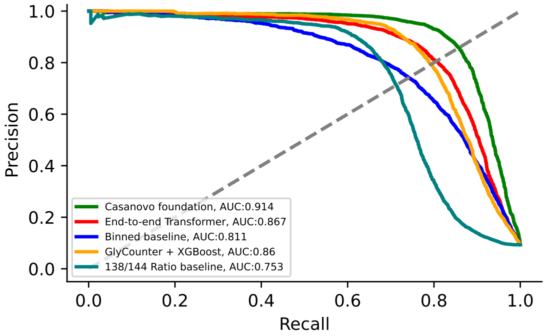

Figure 3. **Glyco precision-recall curve**. Precision-recall curve for each model on the glycosylation status prediction task.

5. Conclusion and future work

In this work, we demonstrate that the spectrum encoder learned by a model trained on the *de novo* sequencing task is generally applicable as a foundation model for tandem mass spectrometry data. Small models trained on frozen spectrum embeddings give good performance across a wide range of downstream tasks. These results demonstrate the utility of foundation models for mass spectrometry proteomics as a flexible starting point for solving novel tasks without the need for massive task-specific labeled datasets.

One promising avenue for future research is to replace or augment the *de novo* pre-training with an unsupervised pre-training task, as has been done in metabolomics [4, 14, 7]. Although this will not dramatically increase the training dataset size, it may lead to richer and more generalizable spectrum representations. Additionally, such an approach would allow the inclusion of more diverse spectra, including those not readily annotatable by database search.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

[1] Altenburg, T., Muth, T., and Renard, B. Y. yHydra: Deep Learning enables an Ultra Fast Open Search by Jointly Embedding MS/MS Spectra and Peptides of Mass Spectrometry-based Proteomics, December 2021.

[2] Altenburg, T., Giese, S. H., Wang, S., Muth, T., and Renard, B. Y. 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, April 2022. doi: 10.1038/s42256-022-00467-7. URL https://www.nature.com/articles/s42256-022-00467-7. Publisher: Nature Publishing Group.

[3] Ardito, F., Giuliani, M., Perrone, D., Troiano, G., and Muzio, L. L. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review). *International Journal of Molecular Medicine*, 40(2):271–280, August 2017. ISSN 1107-3756. doi: 10.3892/ijmm.2017.3036. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5500920/.

[4] Asher, G., Campbell, J. M., Geremia, J., and Kassis, T. LSM1-MS2: A Self-Supervised Foundation Model for Tandem Mass Spectrometry Applications, Encompassing Extensive Chemical Property Predictions and Spectral Matching, February 2024. URL https://chemrxiv.org/engage/chemrxiv/article-details/65cc155d66c1381729978dd9.

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

[6] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., and Brunskill, E. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[7] Bushuiev, R., Bushuiev, A., Samusevich, R., Brungs, C., Sivic, J., and Pluskal, T. Emergence of molecular structures from repository-scale self-supervised learning on tandem mass spectra. *chemRxiv*, 2024. doi:10.26434/chemrxiv-2023-kss3r-v2.

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

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

[10] Dorl, S., Winkler, S., Mechtler, K., and Dorfer, V. PhoStar: Identifying Tandem Mass Spectra of Phosphorylated Peptides before Database Search. *Journal of Proteome Research*, 17(1):290–295, January 2018. ISSN 1535-3893. doi: 10.1021/acs.jproteome.7b00563. URL https://doi.org/10.1021/acs.jproteome.7b00563. Publisher: American Chemical Society.

[11] Frejno, M., Berger, M. T., Tueshaus, J., Hogrebe, A., Seefried, F., Graber, M., Samaras, P., Fredj, S. B., Sukumar, V., Eljagh, L., Bronshtein, I., Mamisashvili, L., Schneider, M., Gessulat, S., Schmidt, T., Kuster, B., Zolg, D. P., and Wilhelm, M. Unifying the analysis of bottom-up proteomics data with chimerys. *bioRxiv*, 2024. doi: 10.1101/2024.05.27.596040. URL https://www.biorxiv.org/content/early/2024/05/27/2024.05.27.596040.

[12] Gholamizoj, S. and Ma, B. SPEQ: quality assessment of peptide tandem mass spectra with deep learning. *Bioinformatics*, 38(6):1568–1574, 2022.

[13] He, M., Zhou, X., and Wang, X. Glycosylation: mechanisms, biological functions and clinical implications. *Signal Transduction and Tar-*

geted Therapy, 9(1):1–33, August 2024. ISSN 2059-3635. doi: 10.1038/s41392-024-01886-1. URL https://www.nature.com/articles/s41392-024-01886-1. Publisher: Nature Publishing Group.

[14] Healey, D., Domingo-Fernández, D., Taylor, J., Krettler, C., Lightheart, R., Park, T., Kind, T., Allen, A., and Colluru, V. PRISM: A foundation model for life's chemistry. URL https://enveda.com/prism-a-foundation-model-for-lifes-chemistry/.

[15] Kingma, D. and Ba, J. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[16] Kothlow, K., Schramm, H. M., Markuson, K. A., Russell, J. H., Sutherland, E., Veth, T. S., Zhang, R., Duboff, A. G., Tejus, V. R., McDermott, L. E., Dräger, L. S., and Riley, N. M. Extracting informative glycan-specific ions from glycopeptide MS/MS spectra with GlyCounter, March 2025. URL https://www.biorxiv.org/content/10.1101/2025.03.24.645139v1. Pages: 2025.03.24.645139 Section: New Results.

[17] Li, J., Guo, B., Zhang, W., Yue, S., Huang, S., Gao, S., Ma, J., Cipollo, J. F., and Yang, S. Recent advances in demystifying O-glycosylation in health and disease. *PROTEOMICS*, 22(23-24):2200156, 2022. ISSN 1615-9861. doi: 10.1002/pmic.202200156. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pmic.202200156. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.202200156.

[18] Liu, X., Fields, R., Schweppe, D. K., and Paulo, J. A. Strategies for Mass Spectrometry-based Phosphoproteomics using Isobaric Tagging. *Expert review of proteomics*, 18(9):795–807, September 2021. ISSN 1478-9450. doi: 10.1080/14789450.2021.1994390. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8595857/.

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

[20] Ma, Z.-Q., Chambers, M. C., Ham, A.-J. L., Cheek, K. L., Whitwell, C. W., Aerni, H.-R., Schilling, B., Miller, A. W., Caprioli, R. M., and Tabb, D. L. ScanRanker: Quality assessment of tandem mass spectra via sequence tagging. *Journal of Proteome Research*, 10(7):2896–2904, 2011.

[21] MacCoss, M. J., Alfaro, J. A., Faivre, D. A., Wu, C. C., Wanunu, M., and Slavov, N. Sampling the proteome by emerging single-molecule and mass spectrometry methods. *Nature methods*, 20(3):339–346, March 2023. ISSN 1548-7091. doi: 10.1038/s41592-023-01802-5. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10044470/.

[22] Ochoa, D., Jarnuczak, A. F., Viéitez, C., Gehre, M., Soucheray, M., Mateus, A., Kleefeldt, A. A., Hill, A., Garcia-Alonso, L., Stein, F., Krogan, N. J., Savitski, M. M., Swaney, D. L., Vizcaíno, J. A., Noh, K.-M., and Beltrao, P. The functional landscape of the human phosphoproteome. *Nature Biotechnology*, 38(3):365–373, March 2020. ISSN 1546-1696. doi: 10.1038/s41587-019-0344-3.

[23] Polasky, D. A., Geiszler, D. J., Yu, F., Li, K., Teo, G. C., and Nesvizhskii, A. I. MSFragger-Labile: A flexible method to improve labile PTM analysis in proteomics. *Molecular & Cellular Proteomics*, 22(5), 2023.

[24] Potel, C. M., Burtscher, M. L., Garrido-Rodriguez, M., Brauer-Nikonow, A., Becher, I., Le Sueur, C., Typas, A., Zimmermann, M., and Savitski, M. M. Uncovering protein glycosylation dynamics and heterogeneity using deep quantitative glycoprofiling (DQGlyco). *Nature Structural & Molecular Biology*, pp. 1–16, February 2025. ISSN 1545-9985. doi: 10.1038/s41594-025-01485-w. URL https://www.nature.com/articles/s41594-025-01485-w. Publisher: Nature Publishing Group.

[25] Purvine, S., Kolker, N., and Kolker, E. Spectral quality assessment for high-throughput tandem mass spectrometry proteomics. *Omics: a journal of integrative biology*, 8(3):255–265, 2004.

[26] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. 2018.

[27] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

[28] Riley, N. M., Malaker, S. A., Driessen, M. D., and Bertozzi, C. R. Optimal Dissociation Methods Differ for N- and O-Glycopeptides. *Journal of Proteome Research*, 19(8):3286–3301, August 2020. ISSN 1535-3893. doi: 10.1021/acs.jproteome.0c00218. URL https://doi.org/10.1021/acs.jproteome.0c00218. Publisher: American Chemical Society.

[29] Saba, J., Dutta, S., Hemenway, E., and Viner, R. Increasing the Productivity of Glycopeptides Analysis by Using Higher-Energy Collision Dissociation-Accurate Mass-Product-Dependent Electron Transfer Dissociation. *International Journal of Proteomics*, 2012(1):560391, 2012. ISSN 2090-2174. doi: 10.1155/2012/560391. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2012/560391. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2012/560391.

[30] Salmi, J., Moulder, R., Filen, J.-J., Nevalainen, O. S., Nyman, T. A., Lahesmaa, R., and Aittokallio, T. Quality classification of tandem mass spectrometry data. *Bioinformatics*, 22(4):400–406, 2006.

[31] Schwarz, F. and Aebi, M. Mechanisms and principles of N-linked protein glycosylation. *Current Opinion in Structural Biology*, 21(5): 576–582, October 2011. ISSN 0959-440X. doi: 10.1016/j.sbi.2011.08.005. URL https://www.sciencedirect.com/science/article/pii/S0959440X11001400.

[32] Singh, C., Zampronio, C. G., Creese, A. J., and Cooper, H. J. Higher Energy Collision Dissociation (HCD) Product Ion-Triggered Electron Transfer Dissociation (ETD) Mass Spectrometry for the Analysis of N-Linked Glycoproteins. *Journal of Proteome Research*, 11(9):4517–4525, September 2012. ISSN 1535-3893. doi: 10.1021/pr300257c. URL https://doi.org/10.1021/pr300257c. Publisher: American Chemical Society.

[33] Sutherland, E., Veth, T. S., Barshop, W. D., Russell, J. H., Kothlow, K., Canterbury, J. D., Mullen, C., Bergen, D., Huang, J., Zabrouskov, V., Huguet, R., McAlister, G. C., and Riley, N. M. Autonomous Dissociation-type Selection for Glycoproteomics Using a Real-Time Library Search. *Journal of Proteome Research*, 23(12):5606–5614, December 2024. ISSN 1535-3893. doi: 10.1021/acs.jproteome.4c00723. URL https://doi.org/10.1021/acs.jproteome.4c00723. Publisher: American Chemical Society.

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

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

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

[37] Wen, B., Hsu, C., Zeng, W.-F., Riffle, M., Chang, A., Mudge, M., Nunn, B. L., Berg, M. D., Villen, J., MacCoss, M. J., et al. Carafe enables high quality in silico spectral library generation for data-independent acquisition proteomics. *bioRxiv*, pp. 2024–10, 2024.

[38] Wu, F.-X., Gagné, P., Droit, A., and Poirier, G. G. Quality assessment of peptide tandem mass spectra. *BMC Bioinformatics*, 9(6):1–9, 2008.

[39] Wu, S.-W., Pu, T.-H., Viner, R., and Khoo, K.-H. Novel LC-MS2 Product Dependent Parallel Data Acquisition Function and Data Analysis Workflow for Sequencing and Identification of Intact Glycopeptides. *Analytical Chemistry*, 86(11):5478–5486, June 2014. ISSN 0003-2700. doi: 10.1021/ac500945m. URL https://doi.org/10.1021/ac500945m. Publisher: American Chemical Society.

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

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

[42] Yu, F., Deng, Y., and Nesvizhskii, A. I. MSFragger-DDA+ enhances peptide identification sensitivity with full isolation window search. *Nature Communications*, 16(1):3329, April 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-58728-z. URL https://www.nature.com/articles/s41467-025-58728-z. Publisher: Nature Publishing Group.

[43] Zhao, P., Viner, R., Teo, C. F., Boons, G.-J., Horn, D., and Wells, L. Combining High-Energy C-Trap Dissociation and Electron Transfer Dissociation for Protein O-GlcNAc Modification Site Assignment. *Journal of Proteome Research*, 10(9): 4088–4104, September 2011. ISSN 1535-3893. doi: 10.1021/pr2002726. URL https://doi.org/10.1021/pr2002726. Publisher: American Chemical Society.

S1. Supplementary Tables

Table S1. **Phosphorylation task performance.** Comparison of AHLF, the end-to-end transformer baseline, Casanovo Foundation, and multi-task trained Casanovo Foundation across phosphorylation detection datasets. The 25 datasets listed correspond to the holdout split *a* described in [2]. The first two columns indicate the number of non-phosphorylated versus phosphorylated spectra in each dataset. The reported performance metrics are F_1 score and AUROC. The best performance on each metric in each row is indicated in bold. AHLF results are directly taken from the paper and thus are only available to two significant figures as originally reported.

Dataset	Number non-phospho	Number phospho	AHLF		End-to-end Transformer		Casanovo Foundation	
			F1	AUROC	F1	AUROC	F1	AUROC
OVAS	90936	37720	**0.92**	**0.99**	0.878	0.983	0.887	0.982
TOV-21-Primary	62350	26978	**0.92**	**0.99**	0.872	0.981	0.884	0.982
ES2-Primary	16297	6667	**0.91**	**0.99**	0.818	0.981	0.834	0.979
Daudi	150915	210916	0.90	0.96	**0.953**	**0.986**	0.924	0.980
U2OS	92329	205353	0.90	0.95	**0.913**	0.938	0.909	**0.964**
HaCaT	19216	113775	0.95	0.93	0.939	0.950	**0.964**	**0.978**
HT-29	1625	27531	0.97	0.92	0.990	**0.998**	0.973	0.988
HeLa	1469194	2949614	0.89	0.92	**0.923**	0.953	0.917	**0.955**
HEPG2	426	45416	0.98	0.92	**0.989**	**0.998**	0.971	0.983
A549	4068	172792	0.92	0.91	**0.977**	**0.988**	0.950	0.960
Colon	8359	28798	0.90	0.90	**0.938**	**0.965**	0.904	0.925
Primary-Gastro	22026	219767	**0.94**	0.88	0.931	**0.929**	0.910	**0.929**
LNCaP	53851	5200	0.45	0.87	**0.692**	**0.987**	0.346	0.877
RPMI-8226	1184	413	0.65	0.87	**0.992**	**0.999**	0.727	0.938
HEK293	322811	332690	0.73	0.86	**0.910**	**0.966**	0.760	0.837
Primary-Prostate	9223	100617	0.89	0.86	**0.954**	**0.982**	0.894	0.919
Primary-AML	494184	5893	0.29	0.85	0.755	**0.988**	0.583	0.945
Kasumi-1	2294	29470	0.88	0.85	**0.995**	**0.998**	0.902	0.844
HPAC	594	807	0.56	0.78	**0.962**	**0.985**	0.733	0.820
SU.86.86	909	984	0.53	0.77	**0.967**	**0.996**	0.664	0.782
CFPAC-1	999	780	0.52	0.76	**0.714**	**0.897**	0.600	0.803
PANC-05-04	1079	1426	0.55	0.74	**0.959**	**0.993**	0.685	0.78
PANC-02-03	273	815	0.56	0.72	0.671	**0.896**	**0.731**	0.819
OVSAYO	11515	28	0.02	0.69	0.021	0.825	**0.026**	**0.868**
HDMVEC	4320	2961	0.40	0.60	**0.873**	**0.970**	0.618	0.783

S2. Supplementary Figures

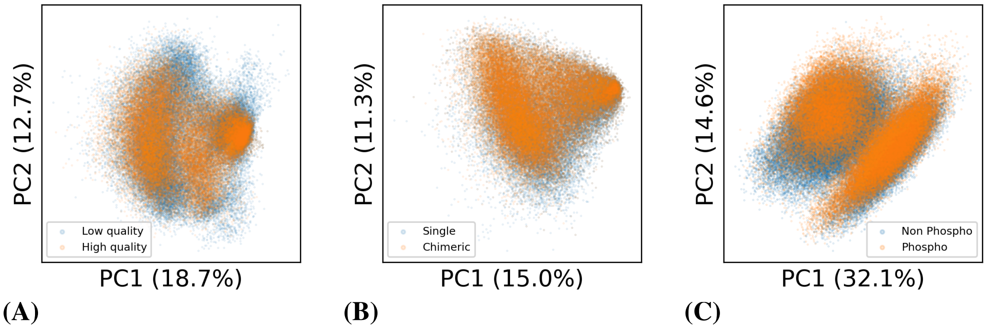

Figure S1. **Learned spectrum embeddings**. PCA plots visualizing the learned embeddings from the pre-trained encoder for test set spectra for (A) the quality prediction task, (B) the chimericity prediction task and (C) the phosphorylation prediction task.

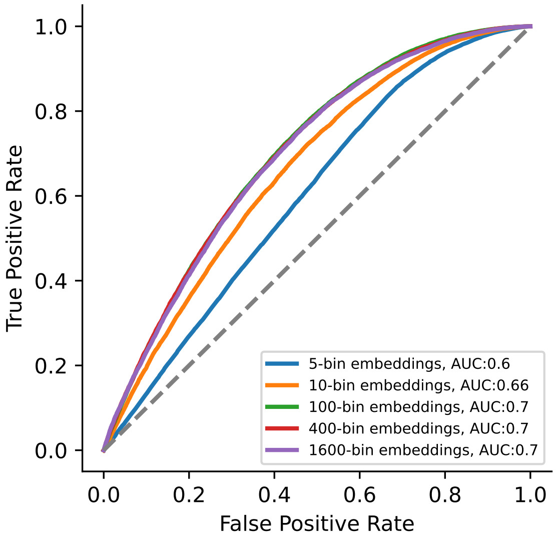

Figure S2. **Comparison of binned embedding performance at different binning resolutions on the chimericity prediction.** ROC curves and the area under the curve (AUC) are reported as a function of the number of bins used to represent peak intensities for the chimericity prediction validation set.

S3. Additional Explanations and Experimental Details

S3.1. Quality task

To create a labeled dataset for this task, we randomly sample 20 human Orbitrap HCD mass spectrometry runs from the MassIVE-KB data. The experiments MSV000080254, MSV000080255, MSV000081563, MSV000081607, MSV000081649, MSV000083508, MSV000083961, MSV000083966, MSV000083967 were used in the train set, MSV000083978, MSV000083983, MSV000086369, MSV000086385, MSV000086389 in the validation set, and MSV000086439, MSV000086448, MSV000086491, MSV000088236, MSV000088405 in the test set. We perform a database search for each run against the reference human proteome (UniProt ID UP000005640) using Sage (version 0.14.7) with the default workflow. Spectra that are matched to a peptide under 1% false discovery rate (FDR) are labeled as high quality, whereas spectra that failed to be matched are annotated as low quality.

S3.2. Chimericity task

The samples were prepared using the method described in [37] and analyzed using an Orbitrap Fusion Lumos mass spectrometer. Raw MS/MS data were converted to mzML files using MSConvert with peak picking enabled in ProteoWizard (version 3.0.24031) [8]. The human, mouse, and yeast MS/MS data were then searched against a human (20,597 proteins, 02/2024), mouse (21,701 proteins, 02/2024), and yeast (6060 proteins, 02/2024) proteome database, respectively, using FragPipe (version 22.0) with the default workflow and “DDA+” mode (i.e., wide window database search). Database search results were filtered at a 1% PSM-level FDR. Spectra were assigned as chimeric if involved in more than one high-confidence PSM.

Unfortunately, the model weights for SPEQ [12], the only other deep learning method for this task, are not published so we are unable to benchmark against this method. However, our end-to-end transformer serves as a methodologically similar baseline.

S3.3. Phosphorylation task

The raw data and database search results from the human phosphoproteome dataset were downloaded from ProteomeXchange PXD012174 [22], which contains data from 101 human cell and tissue types analyzed using phospho-enrichment assays. Data was prepared following the pre-processing scripts used by AHLF [2], which were shared by the authors. These filtered spectra at a 1% FDR at the PSM, protein, and phosphosite localization level. Additionally, PSMs were filtered based on a minimum score for modified peptides of 40, and a minimum delta score for modified peptides of 6. Spectra assigned only to phosporylated peptides were assigned a positive label and spectra assigned only to unphosphorylated peptides were assigned a negative label. Remaining spectra were discarded. Finally, the data was split into train/validation/test sets at the cell/tissue type level following the same splits used by Altenberg *et al* [2].

S3.4. Glycosylation task

The raw data from the 48 mouse brain HCD runs generated by Potel *et al.* was downloaded from ProteomeXchange PXD052447 [24], along with the FragPipe [23] N- and O-glycopeptide search results. Raw MS/MS data were converted to mzML files using MSConvert with peak picking enabled in ProteoWizard (version 3.0.24031) [8]. The data were randomly split at the run level into train/validation/test sets containing 36/6/6 runs each. N-glyco PSMs in the MSFragger search results were filtered for assigned modifications at asparagine residues; O-glyco PSMs were filtered for assigned modifications at either serine or threonine residues. From there, results were filtered based on a hyperscore greater than 16 and a glycan q-value less than 0.01 to obtain a 1% FDR for glycan assignment. For confident classification labels, cases of co-occupancy of O- and N-glycosites on the same PSM were filtered out. Spectra identified as containing an O-glycopeptide were labeled as positive examples, while spectra containing an N-glycopeptide were assigned negative labels. Spectra not identified with a glycopeptide were discarded. GlyCounter [16] was run on each of these spectra, yielding a list of 54 oxonium ions, which were in turn used to calculate the m/z 138/144 ratio.

S3.5. N- vs O-glycosylation

In some cases, distinguishing N-glycosylation from O-glycosylation is straightforward using ratios of ions that indicate the presence of N-acetylglycosamine (GlcNAc) or N-acetylgalactoseamine (GalNAc). This classification can be simplified to a comparison of m/z 138 to m/z 144, where a 1:1 ratio indicates the presence of GalNAc, but not GlcNAc, in a glycan composition. This ratio is useful for classifying N-glycopeptides, which have GlcNAc but not GalNAc residues, relative to simple core 1 O-glycopeptides, which only contain GalNAc. This task becomes more challenging when considering elongated core-1 O-glycans and core 2-8 O-glycans that contain both GalNAc and GlcNac moieties. For example, core 2 glycans are relatively common in mammalian glycoproteomic datasets, and the GlcNAc residues in these O-glycopeptides mean they produce oxonium ion patterns that look more similar to N-glycopeptides than core 1 O-glycopeptides that lack GlcNAc.

Tryptic N-glycopeptides typically only have a single potential glycosite, meaning that higher-energy collisional dissociation (HCD) is sufficient for both identifying and localizing N-glycosylation [28]. On the other hand, O-glycopeptide sequences can often contain multiple potential O-glycosites per peptide, and thus require the collection of alternative dissociation methods, e.g., electron-transfer dissociation (ETD), to generate peptide fragment ions that retain glycan modifications that facilitate localization. Acquiring ETD spectra incurs a significant overhead in instrument time. Thus, by predicting whether a given HCD spectrum contains an N- versus an O-glycopeptide, we can intelligently guide the data acquisition to spend instrument time acquiring ETD spectra only for the precursor ions for which it is necessary [33].

S4. Training Settings and Hyper-parameters

S4.1. Supervised pre-training

The weights for the pre-trained Casanovo model checkpoint 4.0.0 (Apache 2.0 license) were downloaded from GitHub. This model was trained on the MassIVE-KB dataset using the supervised *de novo* sequencing task as described in Yilmaz *et al.* [41]. Only the weights for the spectrum encoder from the encoder-decoder Casanovo model were used. This gives an encoder-only model with nine transformer encoder block layers, an embedding size of 512, and eight attention heads. Overall spectrum representations were obtained from this encoder by taking the mean of the individual peak embeddings.

S4.2. Task-specific training

Binned baseline. To pre-process the input for our binned baseline models, we discretize the *m/z* axis into equal-width bins between 150 and 2000 *m/z*. 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 S2). Peaks outside the range 140–2000 *m/z* are filtered out, and the remaining peak intensities are binned at 18.6 *m/z* resolution. We then train a gradient-boosted decision tree classifier on these representations [9] using the validation set for early stopping based on validation AUROC. The hyperparameter early_stopping_rounds was set to 32, and n_iters was chosen to be sufficiently large that training is always terminated by early stopping. Otherwise, default parameters were used.

Glycounter baseline. Similar to the binned baseline, the GlyCounter baseline for the glycosylation status prediction task represents each spectrum as a 54-dimensional vector of intensities for a pre-defined set of oxonium ions known to be produced by glycan fragmentation. An XGBoost classifier is likewise trained on these representations.

End-to-end transformer. For the end-to-end transformer pipeline, we train the transformer spectrum encoder and MLP classifier head end-to-end on each task. The transformer encoder is implemented using depthcharge components to have the same architecture as the Casanovo encoder, except for the number of transformer layers, which was optimized based on validation set performance from the interval [1-9]. Gradient updates during were performed using the Adam optimizer [15] with a learning rate of 1e-4 and a weight decay of 1e-6. Training is terminated with early stopping based on AUROC on the validation set with a patience of to 5 epochs.

Casanovo Foundation. To apply Casanovo Foundation to a downstream task, we first use the pretrained encoder from Casanovo version 4.0.0 (Apache License 2.0) to obtain 512-dimensional spectrum embeddings for each spectrum. We then train a small two layer dense network on these embeddings. The dense model has one 512-dimensional hidden layer with ReLU activation. The model is trained with a learning rate of 1e-3, and is also terminated via early stopping with a patience of 5 epochs.

S4.3. Timing and compute resources

Pre-training Casanovo on MassIVE-KB took 8 days on 4 RTX 2080 Ti GPUs. Training models for each of the downstream tasks was done using 2 L40S 48GB GPUs and took ∼8 days in total. The majority of this time was spent training the end-to-end transformer model on the phosphorylation task. All remaining experiments were done on a CPU workstation with 16x Intel Xeon CPU E5-2680 @ 2.70GHz and 64GB of RAM in relatively negligible time.

This dependence on large compute resources, GPUs in particular, is a notable current limitation of Casanovo Foundation. In practice, many mass spectrometry proteomics labs do not have access to or familiarity with using GPUs. However, as deep learning is becoming more widespread in the field, labs are beginning to invest more in local and cloud-based compute resources. Additionally, future engineering efforts to accelerate the inference time of Casanovo Foundation can further bridge this gap.