Modanovo: A Unified Model for Post-translational Modification-Aware *De Novo* Sequencing Using Experimental Spectra From *In Vivo* and Synthetic Peptides

Authors

Daniela Klaproth-Andrade, Yanik Bruns, Wassim Gabriel, Christian Nix, Valter Bergant, Andreas Pichlmair, Mathias Wilhelm, and Julien Gagneur

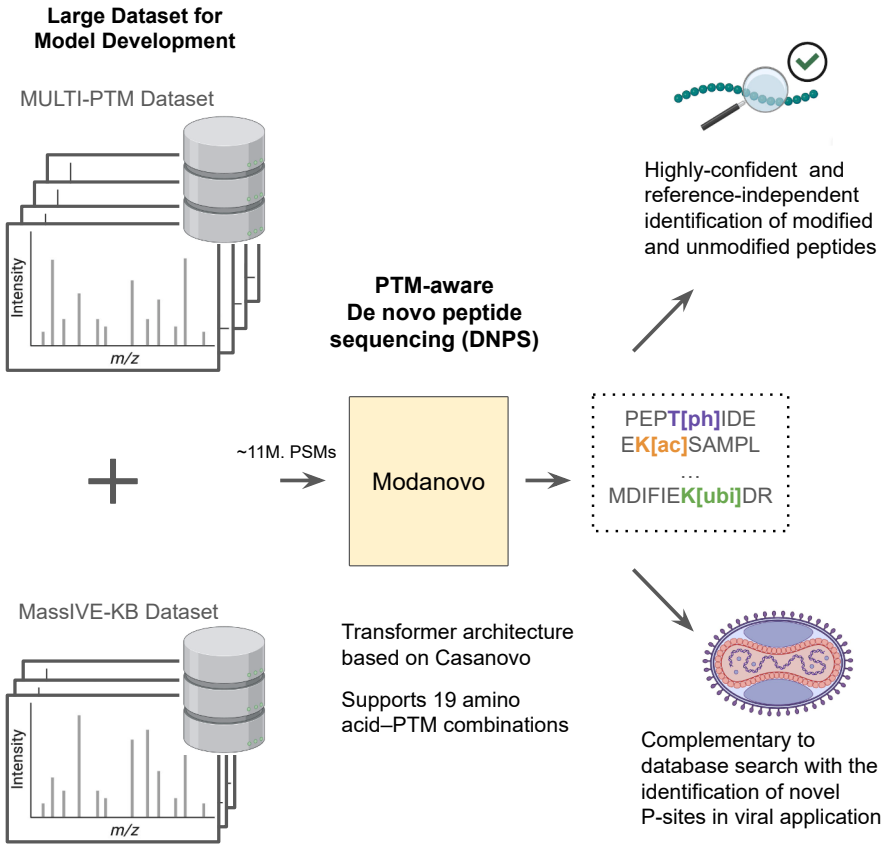
Correspondence

mathias.wilhelm@tum.de; gagneur@in.tum.de

In Brief

Modanovo is a transformer-based de novo peptide sequencing model that expands Casanovo to identify both modified and unmodified peptides. Trained on a large dataset spanning 19 biologically relevant amino acid–PTM combinations, it achieves robust performance across phosphorylation, acetylation, and ubiquitination while maintaining strong performance on unmodified peptides. Modanovo outperforms existing PTM-aware de novo peptide sequencing methods and complements database searches, enabling confident recovery of peptides and novel phosphosites, thereby providing a broadly applicable framework for comprehensive PTM-inclusive de novo sequencing.

Graphical Abstract



2026, Mol Cell Proteomics *25*(*2*), 101501

https://doi.org/10.1016/j.mcpro.2025.101501

Highlights

- Large dataset of PSM covering 19 biologically relevant amino acid-PTM combinations.
- De novo peptide sequencing model with 92% average precision across PTMs.
- Beats previous models on their restricted PTM set and matches Casanovo on unmodified.
- Beats database search in both 19 amino acid-PTM-restricted and open-search modes.
- Reveals new P-sites complementing database search on monkeypox virus-infected cells.

Modanovo: A Unified Model for Post-translational Modification-Aware *De Novo* Sequencing Using Experimental Spectra From *In Vivo* and Synthetic Peptides

Daniela Klaproth-Andrade[1], Yanik Bruns[1], Wassim Gabriel[2], Christian Nix[1], Valter Bergant[3,4], Andreas Pichlmair[3,5], Mathias Wilhelm[2,6,*], and Julien Gagneur[1,6,7,8,*]

Post-translational modifications (PTMs) play a central role in cellular regulation and are implicated in numerous diseases. Database searching remains the standard for identifying modified peptides from tandem mass spectra but is hindered by the combinatorial expansion of modification types and sites. De novo peptide sequencing offers an attractive alternative, yet existing methods remain limited to unmodified peptides or a narrow set of PTMs. Here, we curated a large dataset of spectra from endogenous and synthetic peptides from ProteomeTools spanning 19 biologically relevant amino acid-PTM combinations, covering phosphorylation, acetylation, and ubiquitination. We used this dataset to develop Modanovo, an extension of the Casanovo transformer architecture for de novo peptide sequencing. Modanovo achieved robust performance across these amino acid-PTM combinations (median area under the precision-coverage curve 0.92), while maintaining performance on unmodified peptides (0.93), nearly identical to Casanovo (0.94). The model outperformed π-PrimeNovo-PTM and InstaNovo-P and showed increased precision and complementarity to the database search tool MSFragger. Robustness was confirmed across independent datasets, particularly at peptide lengths frequently represented in the curated dataset. Applied to a phosphoproteomics dataset from monkeypox virus-infected cells, Modanovo recovered numerous confident peptides not reported by database search, including new viral phosphosites supported by spectral evidence, thereby demonstrating its complementarity to database-driven identification approaches. These results establish Modanovo as a broadly applicable model for comprehensive de novo sequencing of both modified and unmodified peptides.

Post-translational modifications (PTMs) enhance the functional diversity of proteins by altering amino acid side chains after translation (1). These chemical modifications are relevant to protein activity, localization, and interactions and are often implicated in disease (2). As a result, identifying PTMs is a central goal in proteomics and is essential for understanding many biological processes and diseases (3).

Bottom-up proteomics, typically using liquid chromatography-tandem mass spectrometry, is the primary method for large-scale protein and PTM identification (4, 5). Here, proteins are enzymatically digested into peptides, which are first analyzed in a first mass spectrometry scan to measure their mass-to-charge (m/z) ratios (6–8). Selected precursor ions are then isolated and fragmented, and the resulting fragment ions are measured in a second scan, producing a tandem mass spectrum (9, 10). Peptide sequences can, in principle, be inferred from mass differences between fragment ions (11, 12), but this remains challenging due to spectral noise, missing peaks, co-fragmented peptides, and ambiguity in ion series assignment. Consequently, peptide identification, which serves as the foundation for subsequent protein inference and quantification, is typically performed through database searching (13, 14). Here, experimental spectra are matched against theoretical spectra generated from the reference proteome of the organism of

From the [1]TUM School of Computation, Information and Technology, Technical University of Munich, Garching, Germany; [2]TUM School of Life Sciences, Technical University of Munich, Munich, Germany; [3]Institute of Virology, Technical University of Munich, School of Medicine, Munich, Germany; [4]Department of Molecular Biology and Nanobiotechnology, National Institute of Chemistry, Ljubljana, Slovenia; [5]German Centre for Infection Research (DZIF), Partner Site Munich, Munich, Germany; [6]Munich Data Science Institute (MDSI), Technical University of Munich, Garching, Germany; [7]Institute of Human Genetics, School of Medicine, Technical University of Munich, Munich, Germany; and [8]Computational Health Center, Helmholtz Center Munich, Neuherberg, Germany

*For correspondence: Mathias Wilhelm, mathias.wilhelm@tum.de; Julien Gagneur, gagneur@in.tum.de.

https://doi.org/10.1016/j.mcpro.2025.101501

interest (15, 16). However, including PTMs in the context of database searching exponentially increases the search space, making exhaustive searches increasingly difficult (17).

An alternative or complementary avenue to database search is de novo peptide sequencing (DNPS), which bypasses the reference database entirely, enabling the discovery of novel peptides, rare PTMs, and proteins from unsequenced organisms (18). Recent DNPS tools, including transformer-based models like Casanovo (19–21), have achieved strong results on benchmark datasets (22). However, identifying post-translationally modified peptides remains a major challenge. Casanovo and most of its peer-reviewed successor models (23–29) support only the seven amino acid-PTM combinations defined in the first release of the MassIVE-KB human spectral libraries (30), which include N-terminal modifications (acetylation, carbamylation, loss of ammonia), deamidation, and oxidation. In contrast, π-PrimeNovo (25) expands PTM coverage by fine-tuning on additional PTMs one at a time. It was trained separately for each of the PTM types contained in a dataset covering multiple PTMs (31), demonstrating flexibility through multiple single-PTM models, though the trained weights for all models have not been released publicly. A phosphorylation-specific π-PrimeNovo model trained on the 2020-Cell-LUAD dataset (32) has been released publicly. Similarly, the latest release of InstaNovo broadened its PTM coverage by adding support for phosphorylation (33). Although this single-PTM approach increases PTM diversity, it requires training individual models for each PTM, which is limited by the scarcity of large, diverse PTM datasets and the substantial computational costs of per-PTM training. Furthermore, in cases where multiple PTM types are of interest, each corresponding model must be run separately, and the user must determine which identification to trust, resulting in an approach that is user-biased and less practical for broad or ambiguous modification searches.

To the best of our knowledge, DNPS models that can simultaneously identify a broad spectrum of PTM types within a single unified framework while maintaining robust performance on unmodified peptides are lacking. This limitation partly arises from the scarcity of large-scale, high-quality experimental datasets covering diverse PTMs.

Here, we compiled a sequence-annotated dataset of tandem mass spectra including unmodified peptides and 19 amino acid-PTM combinations, drawing from the MassIVE-KB spectral libraries (30) and the MULTI-PTM dataset, which is part of PROSPECT-PTM (34, 35) and ProteomeTools (36). Using this resource, we developed Modanovo, a unified DNPS model built on the Casanovo architecture and expanding its PTM coverage by 12 amino acid-PTM combinations. Modanovo supports 39 tokens for canonical and modified residues, enabling broad PTM coverage without training separate models and while maintaining strong performance on unmodified peptides. Importantly, Modanovo identifies ubiquitinated, acetylated, and phosphorylated peptides, among others, reflecting PTMs with broad roles in protein regulation and cellular function. We demonstrate the utility of Modanovo by analyzing human foreskin fibroblast (HFF) cells infected with monkeypox virus (MPXV), revealing relevant new phosphorylated peptides missed by conventional database search.

EXPERIMENTAL PROCEDURES

Experimental Design and Statistical Rationale

The primary objective of this study was to develop and evaluate Modanovo, a de novo peptide sequencing (DNPS) model capable of handling datasets containing 19 amino acid-PTM combinations. A detailed description of the data processing workflow and model architecture is provided in subsequent sections. Modanovo was trained on experimental spectra from *in vivo* experiments from the MassIVE-KB (v1) human spectral libraries (30) and synthetic peptides from the MULTI-PTM dataset (34). Its generalization ability was evaluated on three independent publicly available datasets: 21 PTMs (31), MassIVE-KB (v2), and a dataset of MPXV-infected cells (37). To ensure strict separation of peptide sequences, peptide-spectrum matches (PSMs) were partitioned into training, validation, and test sets following the non-overlapping peptide splits established in the Casanovo study (19). Model performance was assessed using standard metrics for DNPS, including peptide-level precision-coverage curves and their corresponding area under the curve on held-out data. Benchmarking was performed against Casanovo, π-PrimeNovo (25), InstaNovo-P (33), and MSFragger (38, 39). No biological replicates were generated, as the study leveraged existing large-scale repositories.

For reproducibility, the source code is publicly available at: https://github.com/gagneurlab/Modanovo. An official code release of Modanovo (v1.0.0) is available on Zenodo at 10.5281/zenodo.17668430 (40). Trained model weights and data splits, along with identification files, are available at 10.57967/hf/6452 (41) and 10.57967/hf/6451 (42) and 10.5281/zenodo.17640938 (43).

Datasets

MassIVE-KB (v1) Dataset—For model development, we downloaded a subset of the MassIVE knowledge base (MassIVE-KB) peptide spectral libraries (30) consisting of ~30 million PSMs, which were also used for training Casanovo (19, 20). These "high-quality" PSMs were identified by applying a very strict PSM-level false discovery rate (FDR) filter and selecting at most 100 PSMs for each precursor charge and modified peptide combination (30). The dataset includes both unmodified peptides and peptides bearing variable PTMs at specific residues. The variable PTM-residue combinations were: methionine oxidation (Unimod ID: 35), deamidation of asparagine and glutamine (Unimod ID: 7), N-terminal acetylation (Unimod ID: 1), N-terminal carbamylation (Unimod ID: 5), and N-terminal loss of ammonia (Unimod ID: 385). In addition, all cysteine residues were treated as fixed, modified with carbamidomethylation.

We considered the same peptide-disjoint training, validation, and test sets that were used to develop Casanovo. From these training, validation, and test sets, we randomly selected a subset (~9%) of

~2.1 M, ~0.2 M, and ~0.2 M PSMs, respectively, to be used for training and evaluation purposes of our model. Retaining a representative proportion of unmodified peptides and those PTMs unique to this dataset was intended to prevent forgetting during fine-tuning, while subsampling reduced computational cost and emphasized the additional PTMs not included in this dataset.

MULTI-PTM Dataset—In the absence of large *in vivo* datasets covering a wide range of PTMs for model development, we complemented our training data with a subset from PROSPECT-PTM (34), part of the ProteomeTools project (36). This "MULTI-PTM dataset" comprises high-quality tandem mass spectra of synthetic peptides with 14 distinct variable PTM-amino acid combinations and was used for model training and evaluation. All modifications in this dataset are precisely localized, and each peptide contains at most one variable PTM in addition to methionine oxidation and the fixed carbamidomethylation of cysteine. The included PTM-residue combinations are: methionine oxidation; N-terminal and lysine acetylation (Unimod ID: 1); arginine citrullination (Unimod ID: 7); monomethylation of lysine and arginine (Unimod ID: 34); phosphorylation of serine, threonine, and tyrosine (Unimod ID: 21); lysine ubiquitinylation (Unimod ID: 121); pyroglutamate formation (pyro-Glu) from glutamic acid and glutamine (Unimod IDs: 27, 28); and O-GalNAc and O-GlcNAc modifications of serine and threonine (Unimod ID: 43).

We restricted the dataset to PSMs from higher-energy collisional dissociation (HCD; beam-type collision-induced dissociation) spectra and to those analyzed with an Orbitrap mass analyzer. Moreover, we randomly subsetted the dataset to contain at most 100 instances of the same unmodified peptide and at most 200 instances of the same modified peptide. This resulted in a total of over eight million PSMs, which were randomly divided into training, validation, and test sets comprising approximately 90%, 5%, and 5% of the data, respectively. The defined sets are disjoint at the peptide level and follow the same split strategy as used in Casanovo. Moreover, the modified and unmodified counterparts of each peptide sequence were always placed in the same split.

21-PTM Dataset—We obtained the raw files and MaxQuant (44) identification results for the 21-PTM dataset (31), part of the ProteomeTools project (36), which contains synthetic peptides with 21 distinct amino acid-PTM combinations, from the PRIDE repository PXD009449. We restricted the dataset to PSMs derived from HCD fragmentation acquired on Orbitrap mass analyzers. We subset the dataset to contain only PTMs present in the MULTI-PTM and MassIVE-KB (v1) datasets, resulting in a total of six amino acid-PTM combinations. Nevertheless, we continue to refer to it as the "21-PTM dataset", consistent with its common usage in the field.

MassIVE-kb (v2) Dataset—We downloaded the second release of the MassIVE Knowledge Base (MassIVE-KB v.2024–05–24, (30)), which extends the variable PTMs cataloged in the first release by including phosphorylation and ubiquitination. Consistent with the download process for the initial release, we applied stringent quality controls by selecting "high-quality" PSMs and limited the dataset to a maximum of 100 PSMs per unique combination of precursor charge and modified peptide sequence. For this study, we further restricted the dataset to PSMs corresponding exclusively to post-translationally modified peptides, resulting in ~150 thousand PSMs.

Monkeypox Virus (MPXV) Dataset—We obtained the raw files and MaxQuant identification files for the full proteome and phosphorylation-enriched datasets of human foreskin fibroblasts (HFF) cells infected with monkeypox virus (MPXV) from the PRIDE repositories PXD040811 and PXD040889 (37). We discarded decoy and secondary peptides so that each experimental spectrum is attributed to at most one peptide sequence, namely the one with the highest score.

Identification of Post-translationally Modified Peptides With Modanovo

Fine-Tuning of a Transformer Model for Peptide Prediction From Tandem Mass Spectra—We based Modanovo on the Casanovo transformer architecture, which employs a sequence-to-sequence framework to predict peptide sequences directly from tandem mass spectra. The model formulates peptide sequencing as a next-token prediction task, where each token represents either a canonical amino acid or a PTM-amino acid combination.

The original Casanovo vocabulary consisted of 28 tokens, including the special stop token and seven amino acid-PTM combinations. To accommodate an expanded vocabulary of 40 tokens reflecting the inclusion of 12 new PTM-amino acid combinations as distinct tokens, we increased the dimensions of the input embedding layer and the output projection layer accordingly. The model with extended token vocabulary was initialized with pre-trained weights from Casanovo (v4.0.0), originally trained on the MassIVE-KB (v1) dataset comprising primarily unmodified peptides. Embeddings for the new tokens were initialized by averaging those of canonical amino acids, providing a meaningful initialization that facilitated efficient learning. We fine-tuned the entire model end-to-end on the combined training dataset containing spectra from both unmodified and modified peptides, enabling the model to learn fragmentation patterns associated with a broader spectrum of PTMs.

Training was conducted with a learning rate of $1 \cdot 10^{-5}$, a dropout probability of 0.1, a batch size of 32, and a maximum of 12 epochs. Model performance was monitored on the validation set to mitigate overfitting, and the model exhibiting the lowest validation loss was selected as the final version for all subsequent analyses. All other training strategies and preprocessing parameters were kept consistent with the default settings established in Casanovo.

Evaluation and Alternative Methods

Peptide Prediction Score—For ranking and evaluation, each PSM was associated with a confidence score. For Modanovo and Casanovo, a peptide-level score was calculated as the mean of the amino acid scores, taken directly from the model's softmax output at each decoding step. To stabilize residue-level confidences, each amino acid score was averaged with the overall peptide score. If the predicted sequence did not match the precursor mass within the specified tolerance, a penalty of −1 was applied to the peptide-level score.

For π-PrimeNovo, peptide confidence was defined as the mean predicted probability across all residues in a sequence, as reported by the non-autoregressive transformer. For MSFragger, peptide scores were taken from the search engine's reported spectral matching score (Hyperscore).

Evaluation Metrics—During model training and evaluation, isoleucine and leucine were treated as indistinguishable due to their identical masses. Additionally, pyroglutamate formation from glutamic acid and glutamine is treated as a single PTM-amino acid combination, reflecting the chemical equivalence of the resulting pyroglutamate residue.

We evaluated the model performance of each DNPS tool using peptide-level precision-coverage curves. For each spectrum, we compared the peptide predicted by a tool to the corresponding peptide identified by a database search engine at a specified FDR threshold. The score refers to the confidence score assigned by the tool. A PSM was considered correct if it exactly matched the database-identified peptide, allowing for substitutions mentioned above. PSMs with a ground truth database identification but no prediction were considered incorrect and assigned the lowest possible score. Precision and coverage were computed across

varying score thresholds t among all PSMs identified by the database search engine, and were defined as:

$$precision\ (t) = \frac{\#\ correct\ PSMs\ with\ score \geq t}{\#PSMs\ with\ score \geq t}$$

$$coverage\ (t) = \frac{\#\ PSMs\ with\ score \geq t}{\#PSMs}$$

The area under the precision-coverage curve (AUPCC) was computed using the trapezoidal method as implemented in the function "auc" of the scikit-learn package (45). In addition to the AUPCC, we often reported the final precision (at coverage 1) for each tool, which is equivalent to the peptide recall obtained by the tool in comparison to the ground truth identifications.

To compute peptide-level precision-coverage curves for a given PTM-amino acid combination, we include all ground truth peptides that contain that specific modification, regardless of whether additional modifications are also present. For example, the ground truth peptide "PEPT[+79.966]IDEK[+14.016]" contributes to both the peptide-level precision-coverage curves for phosphorylated threonine (T[+79.966]) and monomethylated lysine (K[+14.016]).

Peptide Alignment—Peptide alignments were obtained by running blastp (version 2.12.0+) against sequences of the reviewed human proteome, including isoforms (Uniprot, Taxon ID 9606) and Monkeypox virus proteins (GenBank: ON563414.3), downloaded from the PRIDE repository PXD040811. As a scoring matrix, we used the identity matrix, modified such that leucine and isoleucine were considered equivalent. All other blastp settings were set to their default values. We restricted the output of blastp to at most one hit per queried peptide sequence. If the search returned multiple hits, we selected the hit with the lowest e-value. We defined a query peptide to be a perfect alignment if the peptide is identical to the target peptide (except for differences between leucine and isoleucine).

Validation with Prosit—We obtained Prosit (46) spectrum predictions through Koina (47) for unmodified peptide identifications in the MPXV dataset using Prosit. Peptides longer than 30 residues, shorter than seven residues, or with a charge state exceeding six were excluded. We ran Prosit using a normalized collision energy of 30, which yielded the highest spectral angles for most sequences deemed correct when compared with MaxQuant predictions. Prosit-predicted spectra for a given peptide sequence were matched to the corresponding experimental spectra by aligning each Prosit-predicted peak to the nearest experimental peak within a 20 ppm tolerance window, if present. The normalized spectral angle (SA) between the predicted and experimental spectra was defined as:

$$SA = \frac{\sum_i I_{pred_i} \cdot I_{exp_i}}{\sqrt{\sum_i {I_{pred_i}}^2} \cdot \sqrt{\sum_i {I_{exp_i}}^2}}$$

where I_{pred_i} and I_{exp_i} denote the intensities of the i-th matched peak in the predicted and experimental spectra, respectively. Both experimental and theoretical intensities were normalized using base-peak normalization before SA computation. If no experimental peak was found within the tolerance window to match an expected peak, the intensity I_{exp_i} was set to zero. The SA ranged from 0 to 1, with values closer to one indicating greater similarity.

Empirical Precision Estimates on the MPXV Dataset—We obtained empirical precision estimates on the phosphorylation-enriched and full proteome samples of the MPXV dataset by leveraging PSMs identified with MaxQuant as a reference. For each score threshold of Modanovo, we calculated the proportion of PSMs that could be matched to a MaxQuant peptide identification. This proportion was used as an empirical estimate of the precision at that threshold. By scanning across thresholds, we determined the score cutoffs that corresponded to target precision levels of 80%, 90%, and 95% for the full proteome and phospho-enriched samples separately. These empirically derived cutoffs were then applied to the set of PSMs without MaxQuant peptide identifications, enabling the extension of the empirical precision estimates to all identifications.

MPXV H5 Structure Modeling With AlphaFold and Electrostatic Surface Potential Analysis—In silico prediction of the structure of the MPXV H5 dimer was performed using the colab version of AlphaFold 2.3.1 (48) in the multichain mode using default parameters. The electrostatic surface potential of the modeled structure of the MPXV H5 dimer was calculated using the PyMOL plugin APBS electrostatics. Molecular graphics depictions were produced with the PyMOL software (49).

Alternative Methods—Casanovo. We downloaded the Casanovo (v4.0.0) model weights from https://github.com/Noble-Lab/casanovo/releases/tag/v4.0.0. Casanovo was trained on the MassIVE-KB (v1) dataset, and we obtained the corresponding training, validation, and test splits in April 2024 from https://noble.gs.washington.edu/~melih/mskb_casanovo_splits.zip.

π-PrimeNovo-PTM. We cloned the π-PrimeNovoPTM (25) code from https://github.com/PHOENIXcenter/pi-PrimeNovo/tree/main/pi-PrimeNovo-PTM and downloaded the model weights fine-tuned for phosphorylation. This model was trained on the 2020-Cell-LUAD dataset, which focuses on human lung adenocarcinoma and includes 103 LUAD tumor samples along with their matched non-cancerous adjacent tissues (32). This π-PrimeNovo-PTM model represents phosphorylation using a dedicated token "B" with a corresponding mass of 79.9663. For evaluation, we accepted predicted sequences as correct provided the phosphorylation was localized to the correct serine, threonine, or tyrosine residue, irrespective of whether the "B" token preceded or followed the residue.

InstaNovo-P. We installed and executed InstaNovo (v1.1.4) following the instructions provided on its PyPI page (https://pypi.org/project/instanovo). We ran the model using the publicly available phosphorylation fine-tuned weights, which were downloaded from the v1.1.2 release on GitHub (https://github.com/instadeepai/InstaNovo/releases/tag/1.1.2).

MSFragger. We ran MSFragger (v4.1 (38, 39)) on the MULTI-PTM dataset through the FragPipe interface (v22.0) using both closed and open search configurations. Carbamidomethylation of cysteine was specified as a fixed modification. In the closed-search setting, all PTM-amino acid combinations covered by Modanovo were explicitly specified as variable modifications in the search parameters. In the open search setting, at most three modifications per sequence were considered, with PTM mass tolerances ranging from −20 to +250 Da, and PTM-Shepherd (50) was subsequently applied for PTM characterization and localization. Spectra were searched against a forward and reverse version of the reviewed human proteome, without isoforms (UniProt, Taxon ID 9606, containing 20,596 proteins and downloaded November 17, 2023). The search was performed using tryptic digestion with precursor and fragment mass tolerances set to 20 ppm. PSM-level FDR filtering was performed using Percolator with an FDR threshold of 0.01. Peptide-level and protein-level FDR thresholds were set to one to avoid additional filtering beyond the spectrum level.

RESULTS

A Dataset for Developing de Novo PTM Identification Models

We compiled a dataset combining a subset of the MassIVE-KB human spectral libraries (v1), consisting mostly

of unmodified peptides and seven amino acid-PTM combinations, with a curated subset of spectra from MULTI-PTM (part of PROSPECT-PTM (34) and the ProteomeTools project (35, 36)). The latter includes synthetic peptides modified with a range of biologically relevant post-translational modifications (PTMs) such as phosphorylation, acetylation, ubiquitylation, and methylation, which were crucially missing in MassIVE-KB (v1). Due to the scarcity of large, well-annotated experimental datasets covering diverse PTMs, this combination allowed us to leverage both data from *in vivo* experiments as well as from synthetic peptides to improve model generalization. In total, this combined dataset, denoted as development dataset throughout this manuscript, consisted of approximately 11 million PSMs spanning 20 canonical amino acids and 19 distinct PTM-amino acid combinations, 12 of them not being covered by Casanovo and most of its successor models (Fig. 1, *A* and *B*, Supplementary Table S1), enabling comprehensive learning and evaluation across a broad spectrum of peptide modifications.

Modanovo Extends Casanovo to 12 new Amino Acid-PTM Combinations

Based on this development dataset, we developed Modanovo, a transformer-based model designed to identify modified peptides directly from tandem mass spectra. To enable this, we fine-tuned Casanovo's transformer architecture (Fig. 1*C*), initializing the model with weights from a Casanovo model previously trained on MassIVE-KB (v1). To accommodate the expanded vocabulary arising from PTM-amino acid combinations treated as distinct tokens, we adjusted the shapes of the input embedding layer and the final linear projection layer (Methods). The new parameters corresponding to the expanded token set were initialized by averaging the weights of existing tokens in their respective layers, providing a meaningful starting point rather than

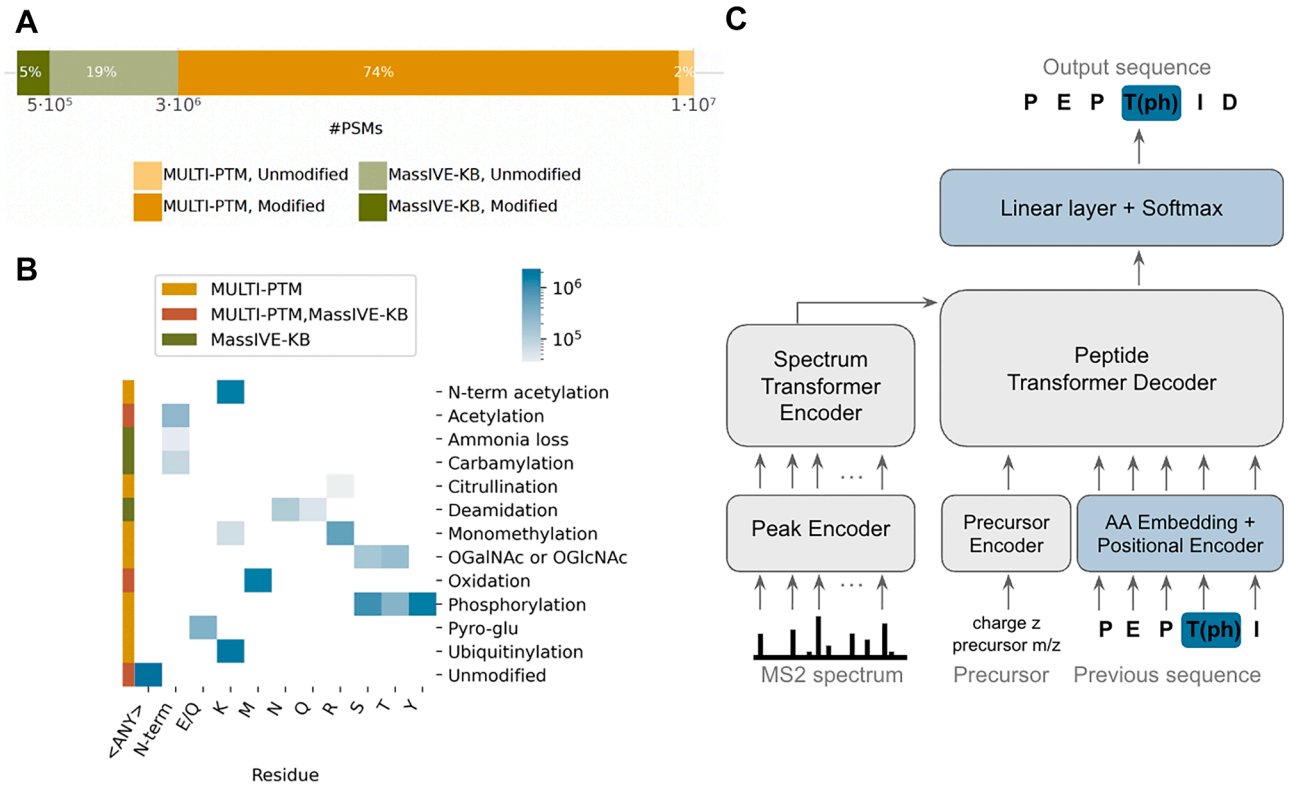


FIG 1. **Modanovo identifies post-translationally modified peptide sequences using a mixture of spectra from *in vivo* experiments and synthetic peptides for model development**. *A*, Number of peptide-spectrum matches (PSMs) for modified and unmodified peptides used for model development from two different data sources: the MassIVE-KB (v1) human spectral libraries (*dark* and *light green*) and from a subset of PROSPECT-PTM (MULTI-PTM, *dark* and *light mustard*). *B*, Heatmap showing the number of PSMs used for model development across PTM types and modified residues. Rows correspond to PTM types, columns to residues. The side color strip indicates the data source (MULTI-PTM, MassIVE-KB, or both). The heatmap is colored on a logarithmic scale, with *darker blue shades* representing higher PSM counts. *C*, An autoregressive transformer encoder and decoder architecture based on Casanovo allows the identification of post-translationally modified peptides directly from tandem mass (MS2) spectra. The model is trained starting with weight initialization from Casanovo's pre-trained weights. The model components for the amino acid (AA) embeddings and final linear layer (in *blue*) are expanded to allow the modification of new post-translationally modified residues.

random initialization. We then fine-tuned the entire model end-to-end, allowing it to retain knowledge of canonical peptide fragmentation patterns while adapting to the variability introduced by PTMs.

Modanovo Confidently Identifies Modified synthetic peptides

We first assessed whether Modanovo's performance remained comparable to that of Casanovo on unmodified peptides and the PTMs covered by Casanovo, which served as the starting point of our fine-tuning approach for PTM expansion. Specifically, we evaluated both models on the test set of the MassIVE-KB (v1) dataset, which was also used to originally develop and evaluate Casanovo and primarily consists of unmodified peptides. We found that Modanovo achieved nearly identical performance to Casanovo overall, with an area under the precision-coverage curve (AUPCC) of 0.93 versus 0.94 (on held-out data here and everywhere else, Supplementary Fig. S1*A*), indicating that it remained well suited for identifying unmodified peptides. In terms of AUPCC across specific PTM-amino acid combinations, Modanovo showed strong agreement with Casanovo (Supplementary Fig. S1*B*), with only minor variations observed across individual PTM categories. Slightly larger performance differences were seen for deamidation and N-terminal loss of ammonia. The reduced performance for peptides with N-terminal ammonia loss and deamidation may reflect added complexity introduced by the inclusion of pyroglutamate and citrullination modifications, which result in similar mass shifts and may interfere with model discrimination in this mass range.

Having confirmed that Modanovo maintained strong performance on unmodified peptides from *in vivo* experiments and the set of PTMs covered by Casanovo, we evaluated Modanovo on the MULTI-PTM proportion of the development dataset. The MULTI-PTM dataset contained 12 distinct PTM-amino acid combinations, which did not overlap with those covered in Casanovo. Across these distinct PTM-amino acid combinations, Modanovo achieved a median AUPCC of 0.92 ([0.70, 0.96] 95% confidence interval) and a median final peptide precision of 0.68 ([0.48, 0.74] 95% confidence interval), measured against ground truth peptide sequences reported by PTM-specific MaxQuant (44) searches (Fig. 2*A*). Comparably, on unmodified peptides from the development dataset, Modanovo attained a median AUPCC of 0.93 and a final peptide precision of 0.71 (Fig. 2*A*). Particularly, phosphorylated peptides were accurately identified by Modanovo, with AUPCC values of 0.96, 0.93, and 0.94, and final precision values of 0.78, 0.71, and 0.72 for sequences containing phosphorylated serine (S[+79.966]), threonine (T[+79.966]), and tyrosine (Y[+79.966]) residues, respectively. Comparable performance was observed for peptides bearing acetylation (K[+42.011] and [+42.011]-), ubiquitination (K[+114.043]), and citrullination (R[+0.984]). Notably, Modanovo achieved an AUPCC of 0.96 for sequences containing pyroglutamate residues (E[-18.011] and Q[-17.027]). This performance, which surpassed that observed for unmodified peptides, may be due to the model successfully learning that pyroglutamate formation is restricted to the first residue position in a peptide sequence.

In contrast to the consistently high performance observed across most PTM types, Modanovo showed reduced performance on peptides containing O-GalNAc and O-GlcNAc modifications of serine and threonine (AUPCC of 0.25 and 0.5). Perhaps, this was partly because these glycosylation events produce complex and often heterogeneous fragmentation patterns, with reduced fragment ion coverage and intensity that hinder reliable sequence reconstruction (51). Moreover, both modifications are encoded using the same token in the model due to their mass equivalence and equal Unimod identifier (52), despite their structural differences. The challenge is compounded by the limited number of examples for these glycopeptides in the dataset (Fig. 2*B*, ~19,000 glycopeptides vs. a mean number of test PSMs per PTM type in the MULTI-PTM dataset of 66,109). Interestingly, when allowing modified peptide sequences to be considered correct despite shifts in the O-GalNAc/O-GlcNAc modification site between serine and threonine (e.g., treating PEPTI**S[+203.079]**ER and PEP**T[+203.079]**ISER equivalently), or between two serine or threonine residues, the AUPCC improved (from 0.25 to 0.58 for serine residues and from 0.50 to 0.70 for threonine residues, Supplementary Fig. S2), although it still fell short of the levels achieved for other PTM types. This improvement showed that while the model sometimes struggled to localize the modification to the exact residue, it more often correctly called the modification at the peptide level and recovered the underlying unmodified peptide sequence. These shifts likely reflect the inherent difficulty of pinpointing the modification site when spectra lack clear site-determining fragment ions. Similarly, the somewhat decreased performance of monomethylated peptides could be partly attributed to disagreements in the localization sites between the ground truth and predicted peptide sequences. Allowing for different monomethylation sites increased the AUPCC from 0.83 to 0.85 for arginine residues and from 0.85 to 0.91 for lysine residues (Supplementary Fig. S2).

Overall, a moderate positive correlation was observed between the AUPCC and the number of PSMs in the test set, but the relationship was not statistically significant (Fig. 2*B*, Spearman's $\rho = 0.45$, $p = 0.13$). This suggests that factors beyond dataset size, such as fragmentation behavior or the structural properties of specific PTMs, played an important role in model performance. Notably, some PTMs with relatively few examples, such as citrullination, still achieved competitive AUPCC values, while others with larger sample sizes, such as arginine monomethylation, performed more modestly. These findings indicate that PTM-specific learnability may outweigh the absolute amount of training data in determining predictive accuracy.

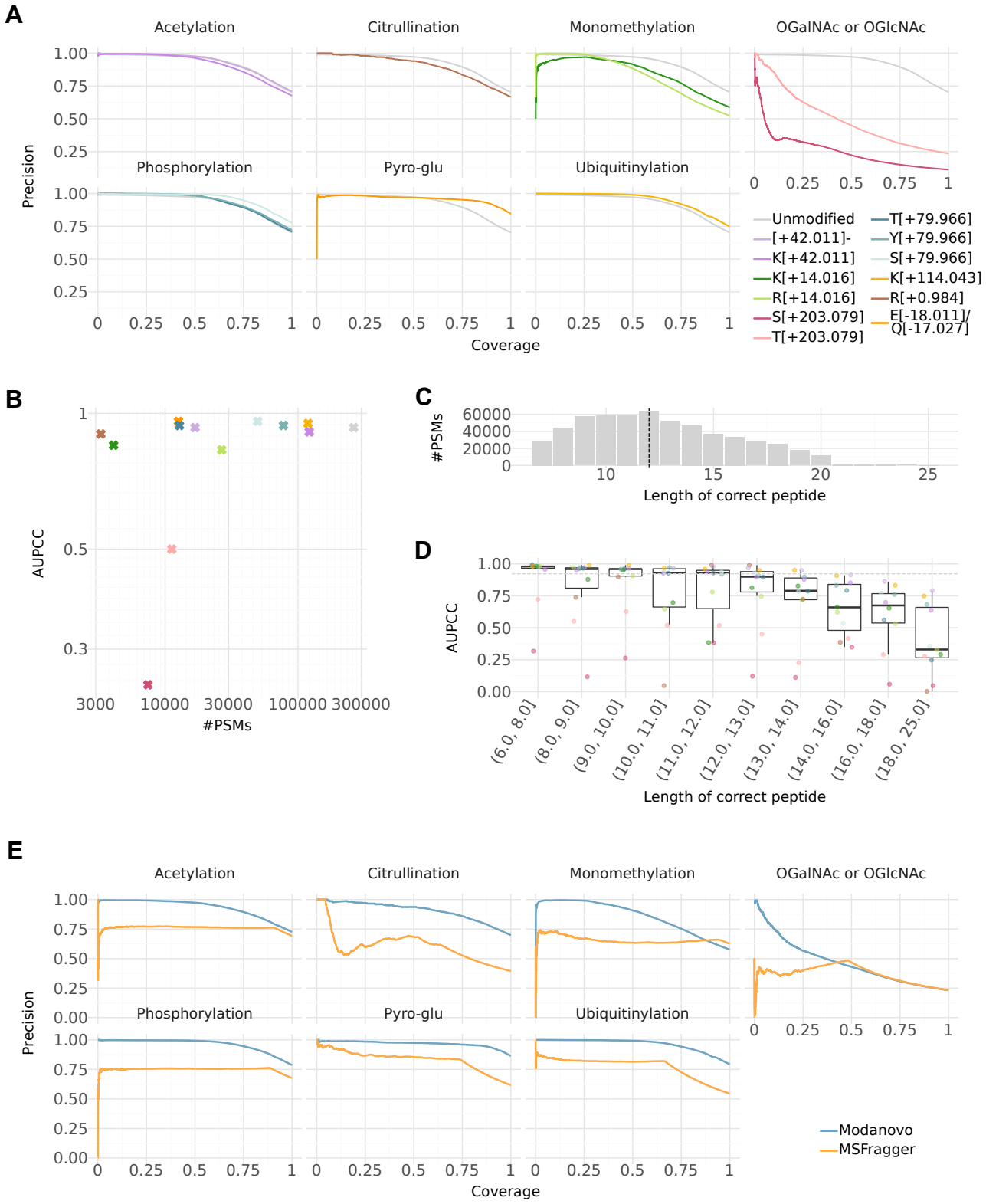


FIG 2. **Model performance on the test set of the development dataset.** *A*, Precision-coverage curves at the peptide level for the 12 new PTM-amino acid combinations covered by Modanovo, as well as methionine oxidation. PTM types (Acetylation, Citrullination, Monomethylation, OGalNAc/OGlcNAc, Oxidation, Phosphorylation, Pyro-glu, and Ubiquitination) are shown in the different panels, and colors represent different PTM-residue combinations. Performance for unmodified peptides (*light grey*) is shown in each panel for comparison.

We reasoned that achieving high performance on longer peptides is more challenging, since every amino acid and its modifications must be correctly identified, the spectra of long peptides are often of lower quality, and errors accumulate due to the autoregressive nature of the transformer model. Therefore, we evaluated the impact of peptide length on Modanovo performance. Peptides on the MULTI-PTM subset were 12 residues long in median (Fig. 2*C*). Modanovo achieved improved or comparable median AUPCC across PTM-residue combinations, relative to the overall median AUPCC (0.92), for peptides up to 13 residues in length (Fig. 2*D*). However, model performance declined for longer peptides. For example, the AUPCC for sequences containing phosphorylated serine dropped from 0.93 to 0.78 when comparing peptides of length 12 to 13 to those of length 13 to 14. Similar observations were made on the Massive-KB (v1) dataset for both Modanovo and Casanovo, whose performance decreased for peptides longer than the median length of 14 (Supplementary Fig. S3). These findings suggest that Modanovo performs most reliably within the peptide length range it was most frequently exposed during training.

Among state-of-the-art DNPS tools claiming to expand PTM coverage beyond Casanovo's vocabulary, publicly available model weights are limited to versions supporting phosphorylation only. Therefore, we compared Modanovo to π-PrimeNovo-PTM (25) and InstaNovo-P (33) on phosphorylated peptides from the development dataset. Modanovo substantially outperformed π-PrimeNovo-PTM and modestly improved InstaNovo-P on all phosphorylated sequences of the dataset (Supplementary Fig. S4).

Having established Modanovo as a DNPS model covering 12 new amino acid-PTM combinations, we next asked whether it could serve as a useful complementary tool to state-of-the-art database search approaches. For this, we first compared the performance of Modanovo on the 12 new amino acid-PTM combinations to that of MSFragger (38), applying a 1% PSM-level FDR. As a realistic database search reflecting the application setting, in which multiple PTMs are of interest and the ground truth is unknown, we ran MSFragger against the full human proteome, allowing for the same set of possible modifications as Modanovo (Methods). We note that this database search setup differed substantially from the one initially used to establish high-quality ground truth annotations (34), where targeted searches were performed with MaxQuant, each restricted to the relevant modification and using a database limited to the synthesized peptides. Overall, Modanovo outperformed MSFragger in identifying modified peptides across all PTM types (Fig. 2*E*), highlighting its strong ability to identify correct PSMs even at high precision levels. Moreover, we ran MSFragger in an open search setting and applied PTM-Shepherd for PTM characterization and localization (Methods), reflecting a realistic scenario in which users have no prior knowledge of the modifications present in the samples. While the search summary correctly recovered all PTM types, except citrullination, at $\geq$1% PSM frequency in the MULTI-PTM dataset, the superiority of Modanovo remained evident (Supplementary Fig. S5), underscoring the added value of DNPS approaches in such applications.

Modanovo Generalizes to Independent Datasets

To evaluate Modanovo's generalizability, we applied it to two distinct datasets: the "21-PTM dataset" (31), consisting of modified synthetic peptides, and a modified-only subset of the latest release of MassIVE-KB (v2), which, in contrast to the first release, included phosphorylated, acetylated, and ubiquitinated residues. We focused our assessment on the 12 new amino acid-PTM combinations covered by Modanovo, resulting in six and five amino acid-PTM combinations from the 21-PTM and MassIVE-KB (v2) datasets. For consistency with common usage in the field, we still refer to the first dataset as the "21-PTM dataset", even though only a subset of PTMs was considered here.

While model performance generally decreased compared to that observed in the development dataset, it remained strong across a diverse range of amino acid-PTM combinations (Fig. 3*A*, Supplementary Fig. S6), though the quality of identification varied depending on the dataset and modification. For instance, the model identified peptides with lysine ubiquitylation (K[+114.043]) with an AUPCC of 0.93 and 0.82 on the 21-PTM and MassIVE-KB (v2) datasets, respectively, compared to 0.95 on the development dataset. Similarly, it reached an AUPCC of 0.89 for lysine acetylation (K[+42.011]) on the 21-PTM dataset, compared to 0.91 for the development dataset. The AUPCC was generally lower on MassIVE-KB (v2) for the remaining PTM types (acetylation and phosphorylation, Fig. 3*A*), with the most pronounced drop observed for phosphorylation (mean AUPCC: 0.94 vs. 0.65).

Precision-coverage curves for the remaining PTM-amino acid combinations, which were covered by Casanovo before, are found in Supplementary Fig. S1. *B*, Area under the precision-coverage curve (AUPCC) against the number of peptide-spectrum matches (PSMs) per PTM-residue combination on the test set, using the same color scheme as in *panel A*. Statistical significance was assessed using Spearman correlation ($\rho = 0.45$, $p = 0.13$). *C*, Distribution of peptide sequence lengths in the test set of the MULTI-PTM dataset. The dashed vertical line marks the median peptide length. *D*, AUPCC at the peptide level across peptide length bins, evaluated on the test set. Each point represents a PTM-amino acid combination, using the same color scheme as in *panel A*. Bins were constructed to contain approximately equal numbers of PSMs. The dashed grey line indicates the AUPCC across all peptide lengths. *E*, Precision-coverage curves at the peptide level comparing Modanovo (*blue*) to MSFragger (claimed 1% FDR, *orange*) on the test set of the MULTI-PTM dataset, faceted by the different PTM types. MSFragger often does not propose a peptide for a given spectrum. These are ranked last and cause the hyperbole sections in the higher coverage range.

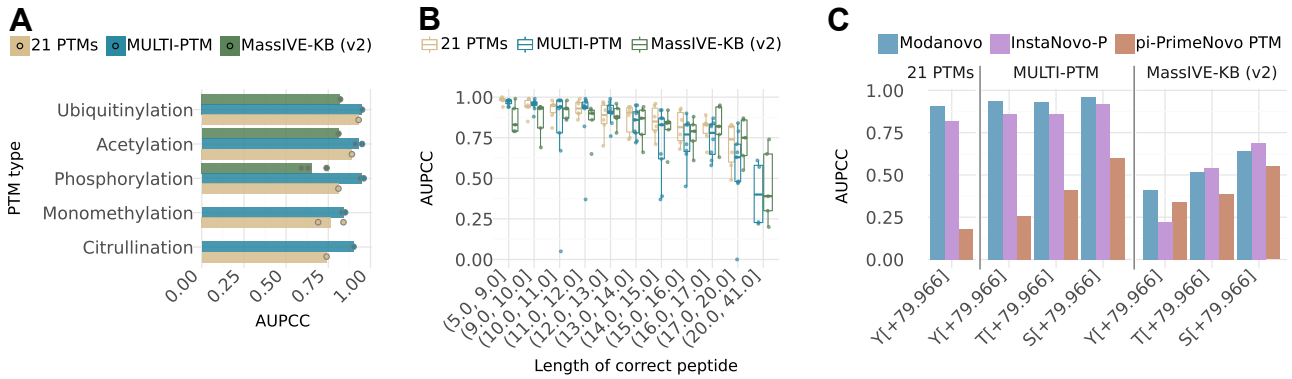


FIG 3. **Model performance on alternative datasets and comparison to other methods**. *A*, Area under the precision-coverage curve (AUPCC) for the different PTM types contained in the three different datasets, MassIVE-KB (v2, *green*), MULTI-PTM (*blue*), and 21 PTMs (*light brown*). Data points represent individual PTM-residue combinations, restricted to the overlap between amino acid-PTM combinations present in the MassIVE-KB or 21-PTM dataset and new 12 amino acid-PTM combinations introduced in Modanovo. *B*, AUPCC across peptide length bins for the three different datasets, MassIVE-KB (v2, *green*), MULTI-PTM (*blue*), and 21 PTMs (*light brown*). Each point represents an amino acid-PTM combination. Bins were constructed to contain approximately equal numbers of PSMs. *C*, AUPCC for the different phosphorylated residues by Modanovo (*blue*) compared to π-PrimeNovo-PTM (*brown*) and InstaNovo-P (*violet*) on the MassIVE-KB (v2), MULTI-PTM, and 21-PTM datasets.

However, this decrease in performance was largely explained by differences in peptide length distributions: MassIVE-KB (v2) contains a greater proportion of longer peptides than the 21-PTM and development datasets (median lengths: 19, 14, and 12, Supplementary Fig. S7). When controlling for peptide length, the AUPCC remained comparable between the three datasets (Fig. 3*B*), indicating that Modanovo's learned representations generalize well to *in vivo*-derived spectra. This consistent performance at matched lengths supports the conclusion that Modanovo is capable of accurately identifying modified peptides from complex biological samples, despite being trained primarily on data from synthetic peptides.

Unlike the peptides in the development dataset, the 21-PTM dataset contains sequences with more than one PTM type (beyond methionine oxidation). To evaluate Modanovo's ability to generalize to these more complex cases, which were not seen during training, we stratified performance by PTM-type combinations. The model was able to sequence multiply modified peptides for most combinations (Supplementary Fig. S8). For example, Modanovo achieved an AUPCC of 0.74 for peptides containing both acetylated and monomethylated residues (K[+42.011] and R[+14.016]), which, while lower than the values observed for singly modified peptides (0.76 for monomethylation and 0.95 for acetylation), still indicated reasonable generalization to peptides carrying multiple co-occurring modifications despite being trained only on singly modified peptides (apart from optional methionine oxidation).

Phosphorylation site (P-site) localization remains a long-standing challenge in mass spectrometry-based proteomics. To assess Modanovo's robustness to this ambiguity, we evaluated its ability to identify phosphorylated peptides in the MassIVE-KB (v2) and 21-PTM datasets while allowing for alternative P-site placements relative to those assigned by database searching. While performance remained unchanged on the 21-PTM dataset when accepting predictions with different P-sites, either on the same residue type or on a different phosphorylatable residue, it increased for the MassIVE-KB (v2) dataset (Supplementary Fig. S9). For example, among peptides with a phosphorylated tyrosine as the ground truth, the AUPCC improved from 0.74 to 0.79 when predictions placing the phosphate on serine or threonine residues were also considered correct. These results underscore Modanovo's flexibility in recovering plausible phosphopeptides despite the inherent uncertainty of P-site assignment.

We further compared Modanovo to the publicly available π-PrimeNovo-PTM (25) and InstaNovo-P (33) models, which allow the prediction of phosphorylated residues, ensuring that all tools were evaluated on the same set of PSMs after removing any PSMs used for training of any tool. Consistent with the observations in the development dataset (Supplementary Fig. S4), Modanovo consistently achieved a higher AUPCC than π-PrimeNovo-PTM across all three phosphorylated residues (Fig. 3C, Supplementary Fig. S10), with gains of 0.41 vs. 0.34 for tyrosine, 0.52 vs. 0.39 for threonine, and 0.64 vs. 0.55 for serine, on the MassIVE-KB (v2) dataset; and substantially outperformed π-PrimeNovo-PTM on the 21-PTM dataset. These results suggest that explicitly modeling PTM-amino acid combinations (e.g., S[+79.966]) as single tokens, rather than treating the PTM mass shift (e.g., [+79.966]) as an independent token that can appear anywhere in the sequence, may contribute to improved predictive performance. Combined with training on a larger and more diverse dataset containing several PTM

types, this design choice could help better capture biologically plausible fragmentation patterns across different PTM types. Compared to InstaNovo-P, Modanovo showed improved performance on the MULTI-PTM and 21-PTM datasets (AUPCC values of 0.91 vs. 0.82 on the 21-PTM dataset, and similar values on the MULTI-PTM dataset). On the MassIVE-KB (v2) dataset, it outperformed InstaNovo-P on tyrosine-phosphorylated sequences (AUPCC values of 0.41 vs. 0.22), while exhibiting minor reductions on threonine/serine-phosphorylated sequences (AUPCC values of 0.52 vs. 0.54 and 0.64 vs. 0.69), an acceptable trade-off given Modanovo's substantially broader PTM coverage.

Modanovo Identifies Phosphopeptides from Monkeypox Virus-Infected Samples

Having established Modanovo as a DNPS tool capable of confidently identifying post-translationally modified peptide sequences from tandem mass spectra, we applied it to a time-resolved dataset of human foreskin fibroblast (HFF) cells infected with monkeypox virus (MPXV), comprising both full proteome and phosphoproteome measurements (37). This recently published dataset represents a compelling use case, as it captures the complex interplay of PTMs in both the virus and the host during the course of infection. Furthermore, DNPS approaches are particularly attractive in viral proteomics, since viral genomes frequently undergo mutations that can complicate database-driven peptide identification. The dataset was originally analyzed with the database search engine MaxQuant. Here, we evaluated the extent to which Modanovo provides complementary insights to this initial analysis.

MaxQuant identified a peptide sequence at 1% FDR for only 15% and 30% of the spectra in the phospho-enriched samples and full proteome samples, respectively. Considering these identifications as ground truth, Modanovo demonstrated strong performance on unmodified peptides and peptides containing only common modifications (N-terminal acetylation or methionine oxidation), achieving an AUPCC of 0.95 and 0.88 on phosphorylation-enriched and full proteome samples, respectively (Fig. 4*A*, Supplementary Fig. S11). This was consistent with the performance on the development dataset (AUPCC of 0.93) and further confirmed the robustness of the model. For peptides containing a single phosphorylated residue, the AUPCC decreased to 0.78. While this represents a drop in performance, it remained comparable to Modanovo's performance on unmodified peptides and phosphopeptides of similar lengths in the development dataset (Fig. 2*D*, Supplementary Fig. S12). Hence, the lower performance on phosphopeptides in the MPXV dataset is attributable to a shift in peptide length distribution: the median sequence lengths were 16 and 17 amino acids for singly and multiply phosphorylated peptides, respectively (Supplementary Fig. S13), longer than those typically observed during training. As expected, model performance declined with increasing proportions of missing y-ion fragments (Supplementary Fig. S13). Notably, the model performance for singly phosphorylated peptides closely matched that of unmodified peptides when controlling for the proportion of missing y-ions (Supplementary Fig. S14). Overall, Modanovo maintained competitive performance when using MaxQuant identifications as ground truth, highlighting its ability to generalize to more complex datasets with longer peptides and variable fragment coverage.

We next analyzed Modanovo peptide predictions beyond the spectra identified by MaxQuant. For this, we set cutoffs on Modanovo confidence scores by leveraging the MaxQuant-identified PSMs as ground truth to map score cutoffs to precision estimates (Methods). As an initial positive control, we observed that, despite having no prior information on the sample preparation or enrichment strategy, Modanovo predominantly predicted unmodified peptides and peptides containing only common modifications (N-terminal acetylation or methionine oxidation) in the full proteome samples and phosphopeptides in the phospho-enriched samples for high precision estimates (Supplementary Fig. S15).

As an orthogonal proxy for peptide plausibility, we assessed the agreement between the experimental spectrum and the Prosit-predicted spectrum (46), using the spectral angle as a similarity metric (0 for no similarity, one for highest similarity, Methods). At the time of this analysis, the available Prosit model did not handle PTMs beyond methionine oxidation. Therefore, this analysis had to be restricted to unmodified peptides (and those containing only methionine oxidation). We stratified spectra by agreement between Modanovo and MaxQuant annotations. When considering all Modanovo predictions without applying any precision estimate, spectral angles were highest for sequences matching MaxQuant and lowest for spectra not identified by MaxQuant, presumably representing lower-quality spectra (Fig. 4*B*). A similar trend was observed when restricting to predictions with precursor mass agreement, although differences in median spectral angles were less pronounced. At the 90% precision threshold, Modanovo predictions matching MaxQuant showed high spectral similarity (median SA: 0.81), closely comparable to Modanovo-only predictions for spectra without MaxQuant identifications (median SA: 0.80; Fig. 4*B*). At 95% precision threshold, spectral angles increased slightly across all groups, reaching 0.84 for both MaxQuant-matching and Modanovo-only predictions. For spectra where Modanovo and MaxQuant disagreed, spectral similarity also improved with confidence, with median SA values of 0.80 and 0.90 at the 90% and 95% thresholds, respectively (Fig. 4*B*). These results support the validity of many high-confidence Modanovo identifications and demonstrate that the model's confidence score is an effective indicator of prediction quality. This is even evident for spectra not annotated by MaxQuant or where the two methods disagree, cases likely to be more challenging for peptide identification.

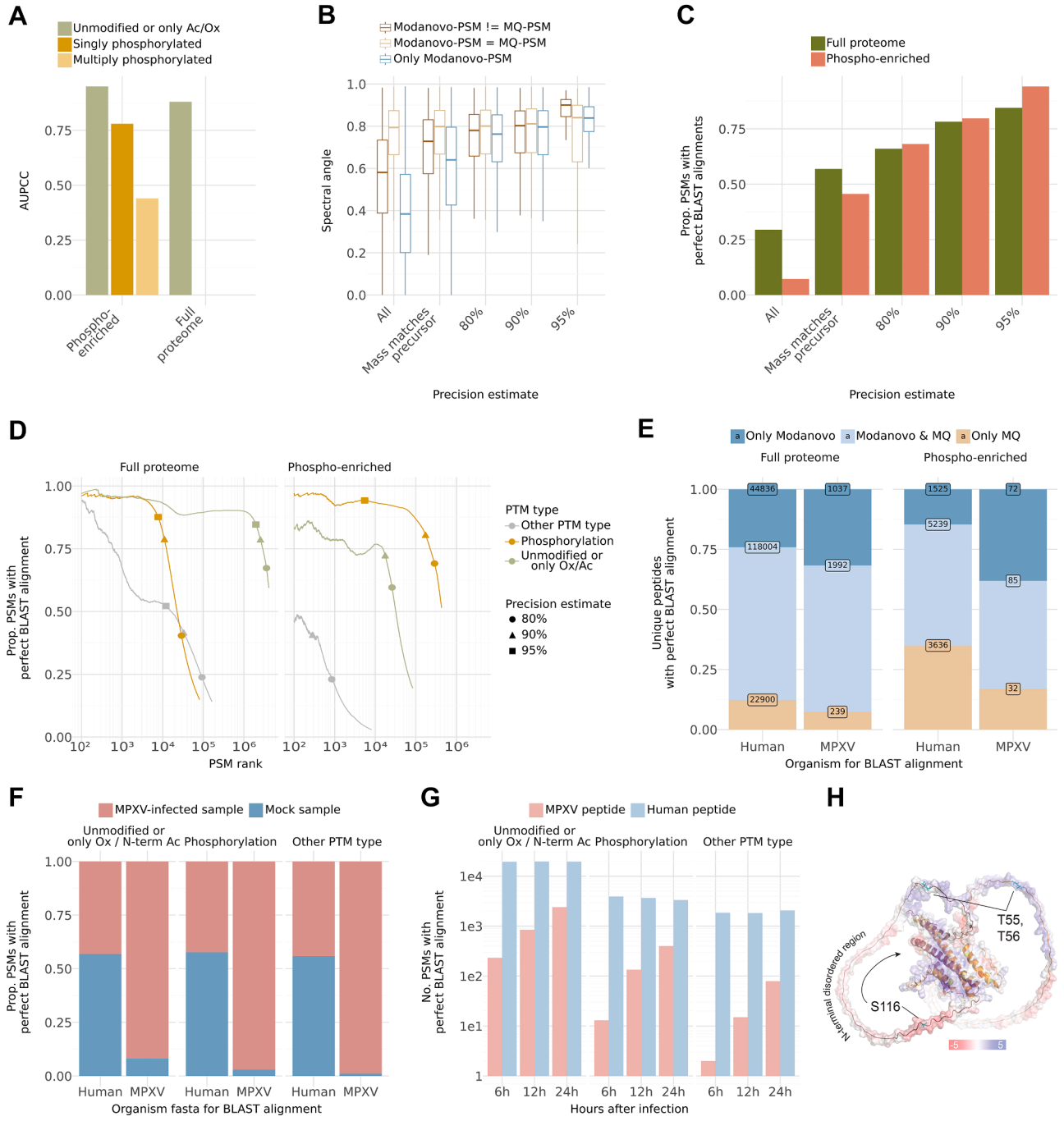


FIG 4. **Modanovo applied to a dataset of monkeypox virus (MPXV)-infected human foreskin fibroblast (HFF) cells**. *A*, Area under the precision-coverage curve (AUPCC) obtained with Modanovo's predictions with MaxQuant peptides as ground truth sequences for comparison for the full proteome samples and phosphorylation-enriched samples with the MaxQuant sequence having multiple phosphorylated residues (*light mustard*), one phosphorylation residue (*mustard*), and no phosphorylation residues (*green*). *B*, Spectral angles obtained from the comparison of experimental spectra and Prosit-predicted spectra for Modanovo's prediction for the whole set of predictions ("All"), for predictions with mass matching the precursor mass-to-charge (m/z) and at different precision estimates (80%, 90% and 95%), and for Modanovo's predictions matching the identified MaxQuant sequence (*light brown*), not matching the MaxQuant sequence (*dark brown*), and for spectra without a MaxQuant identification (*blue*). *C*, Proportion of peptide-spectrum matches (PSMs) with perfect BLAST alignments for the phosphorylation-enriched (*terracotta*) and full proteome (*green*) samples for all PSMs predicted by Modanovo ("All"), the PSMs predicted by Modanovo with calculated mass matching the precursor m/z ("Mass matches precursor"), and at different precision estimates (80%, 90%, and 95%). *D*, Proportion of PSMs with perfect BLAST alignments and mass consistent with the precursor m/z for the phosphorylation-enriched (*right*) and full proteome (*left*) samples, sorted by Modanovo's score for phosphorylated peptides (*mustard*), unmodified peptides, or peptides only containing methionine oxidation or N-terminal acetylation (*green*) and peptides containing other PTM types covered by Modanovo

As a further source of evidence of the plausibility of Modanovo identifications, we next evaluated the proportion of Modanovo-predicted peptides that perfectly aligned to the human or the MPXV proteome, allowing isoleucine-leucine mismatches (Methods). For PSMs from the phospho-enriched samples, the proportion of perfect alignments remained high, reaching 0.94 at an estimated precision of 95%, and 0.80 at 90% precision (Fig. 4*C*, Methods). In comparison, perfect alignment rates for the full proteome samples were slightly lower, reaching 0.84 and 0.72 at the 95% and 90% precision thresholds, respectively. Moreover, in both full proteome and phospho-enriched samples, PSMs corresponding to unmodified peptides or those containing only common modifications (methionine oxidation or N-terminal acetylation) exhibited consistently high alignment rates across the confidence range (Fig. 4*D*). For example, in the top 10,000 PSMs by confidence, the perfect alignment rate exceeded 93% and 77% for unmodified peptides on the full proteome samples and phospho-enriched samples, respectively (Fig. 4*D*). Notably, PSMs predicted to contain a phosphorylation event also maintained high perfect alignment rates, particularly in the phospho-enriched samples, where they reached a proportion of 0.93 perfect alignments in the top 10,000 PSMs and outperformed other PTM types across nearly the entire confidence spectrum. These results particularly support the reliability of phosphorylation predictions in the phospho-enriched samples.

We next assessed the novelty and complementarity of Modanovo's predictions relative to database search by comparing the set of perfectly aligning peptides identified by Modanovo and MaxQuant to the reference proteomes with computed mass matching the experimental precursor, stratified by sample type and species of origin (human vs. MPXV; Fig. 4*E*). Notably, Modanovo identified a substantial number of MPXV peptides in their unmodified version that were missed by MaxQuant, particularly in the phospho-enriched samples. Specifically, Modanovo recovered 72 unique MPXV peptides in the phospho-enriched dataset that were not identified by MaxQuant, more than twice the number of peptides uniquely identified by MaxQuant (n = 32), and 1037 MPXV peptides in the full proteome dataset, compared to 239 unique to MaxQuant. While the proportion of shared peptides was higher for human than for viral sequences, this is expected given the higher abundance of host proteins and potential database coverage biases.

Allowing for alignment mismatches is particularly relevant in the context of viral proteomes, which often exhibit high sequence variability due to rapid mutation rates, strain differences, or incomplete annotation of viral coding sequences. When allowing for amino acid mismatches during alignment in the phospho-enriched samples, Modanovo's advantage became more pronounced: the number of unique MPXV peptides increased by 129 and 247 when allowing for one and two mismatches, respectively (Supplementary Fig. S16). Moreover, aligning predicted peptides against a six-frame translation of the viral genome rather than the reference proteome alone revealed an additional set of unique MPXV peptides: 153 and 314 were identified by Modanovo under one-mismatch and two-mismatch conditions, respectively (Supplementary Fig. S16). These results highlight Modanovo's strength in discovering peptides from viral proteins, particularly under phospho-enriched conditions where modified viral peptides may evade database search detection.

Moreover, we examined the organismal origin of Modanovo's predictions by assessing the proportion of PSMs with perfect alignments to the human or MPXV proteome, stratified by PTM type and infection condition (Fig. 4*F*). Across all modification types, the majority of predicted peptides aligned to the expected species: human sequences appeared in both sample types (mock and MPXV-infected), while MPXV-specific peptides were observed almost exclusively in infected samples. Furthermore, to investigate temporal dynamics of host and viral peptide detection, we quantified the number of Modanovo-predicted PSMs from MPXV-infected samples with perfect alignment to the human or MPXV proteome across different time points post-infection (Fig. 4*G*). MPXV-derived peptides increased markedly over time. In contrast, human peptide counts remained relatively stable across all time points, regardless of modification type. Together, these results highlight Modanovo's ability to recover biologically meaningful peptide sequences and capture dynamic changes in viral expression and post-translational modification patterns across infection time points.

Multiple seminal poxvirus studies have highlighted the importance of phosphorylation dynamics during infection (37, 53–56). We inspected phosphorylation sites identified in the

(*grey*). For clarity, the first 100 PSMs are omitted. *E*, Proportion of unique peptides with perfect BLAST alignment and mass consistent with the precursor m/z for the phosphorylation-enriched (*right*) and full proteome (*left*) samples obtained by querying against the human and MPXV proteome for sequences only identified by Modanovo (*blue*), Modanovo and MaxQuant (MQ, *light blue*), and only MaxQuant (*brown*). *F*, Proportion of PSMs with perfect BLAST alignments and mass consistent with the precursor m/z for the phosphorylation-enriched (*right*) and full proteome (*left*) samples obtained by querying against the human and MPXV proteome for MPXV-infected samples (*terracotta*) and mock samples (*blue*). *G*, Number of PSMs with perfect BLAST alignments and mass consistent with the precursor m/z for the phosphorylation-enriched (*right*) and full proteome (*left*) samples for samples measured at different hours post-infection for MPXV peptides (*light terracotta*) and human peptides (*light blue*). *H*, In silico predicted structure of the MPXV H5 dimer by AlphaFold (48) overlaid with electrostatic surface potential analysis of non-phosphorylated form. Phosphosites S12/13/T15, S27, S134/137/140 and S176/181 are highlighted in *gold*, and sites T55/56 and S116 in teal and labelled.

MPXV multifunctional protein Cop-H5 (57). H5 was highlighted in the study that originally introduced this dataset as the most heavily phosphorylated viral protein with multiple known regulatory phosphosites (37). As an illustrative example, Modanovo consistently detected phosphosites at residues S12, S13, and T15 (cluster 1); S27; S134, S137, and S140 (cluster 2); S176, and S181; in agreement with those previously reported using MaxQuant (37). Phosphorylation of sites in cluster two and S176 was previously shown to regulate double-strand DNA binding activity, suggesting a dynamic role of this protein during the life cycle of this double-strand DNA virus (37). MaxQuant uniquely reported additional sites at S46 and T47, whereas Modanovo uniquely identified sites at T55, T56, and S116. One peptide was detected for sites T55 and T56 in two spectra, which were so far not described. For the target phosphosites at S116, Modanovo identified six distinct peptide sequences across 105 spectra, without a corresponding peptide reported by MaxQuant, predominantly at the later hours post-infection (12h and 24h, Supplementary Fig. S17). These PSMs showed high confidence scores and MS2 spectral evidence (Supplementary Fig. S18). The analogous phosphosite to S116 in the vaccinia virus (S109) in Cop-H5 has been reported previously as part of F10/H1 viral kinase/phosphatase phospho-network, a pivotal viral mechanism driving the phosphorylation dynamics during poxvirus lifecycle (56, 58). These three phosphosites fall within a predicted N-terminal unstructured region of the protein, and to highly charged amino acid stretches thereof as inferred from AlphaFold structural (48) and electrostatic models (Fig. 4H). These three sites, uniquely identified by Modanovo, could thereby hint at functional sites of a Cop-H5 disordered region, potentially regulating its activities or di/multimerization potential. Taken together, these observations provide a proof-of-concept illustration of how DNPS can detect biologically relevant phosphosites and demonstrate how the approach can complement standard database search pipelines.

DISCUSSION

In this work, we curated a comprehensive dataset comprising a large number of peptide-spectrum matches (PSMs) from both *in vivo* experiments and synthetic peptides, including both unmodified peptides and post-translationally modified peptides containing 19 distinct amino acid-PTM combinations. This dataset enables model development and benchmarking of de novo peptide sequencing (DNPS) algorithms supporting modified peptides. Leveraging this dataset, we developed Modanovo, built by extending Casanovo to support 19 PTM-amino acid combinations within a single, unified model, without sacrificing performance on unmodified peptides. Modanovo achieved strong performance across a broad spectrum of modifications on the development dataset and on independent datasets, validating it as a robust and practical extension suitable for downstream applications. The application to a monkeypox virus (MPXV) dataset demonstrated the complementarity of Modanovo to a state-of-the-art database search approach, revealing hundreds of well-supported peptides missed by database search and new MPXV phosphosites.

The comparison between open-search MSFragger and Modanovo on the MULTI-PTM dataset, where Modanovo performed favorably, highlights how a PTM-aware DNPS model can serve as an effective first-pass PTM discovery step: de novo predictions can reveal which modifications are present even without prior specification, driven mainly by mass-over-charge evidence and potentially augmented by fragment intensity information. These confidently detected PTMs can then be used to constrain a subsequent closed-search database run, reducing the search space and improving identification performance.

Modanovo was trained on synthetic peptides for most PTM types. Synthetic peptide data are less noisy and miss fewer fragment peaks than data from “real” experiments. Nevertheless, these differences did not pose significant challenges during model transfer, underscoring the robustness of the learned representations. Nonetheless, the accurate prediction of longer, heavily modified peptides remains difficult. This challenge arises from three main sources. First, spectra of longer peptides are more likely to miss some fragment peaks. Second, autoregressive transformer decoders inherently accumulate errors as sequences grow longer, making each subsequent prediction increasingly susceptible to earlier inaccuracies. Third, the dataset used for training contains few long synthetic modified peptides. Several avenues could be pursued to mitigate this limitation in future DNPS work on modified peptides. One option is to synthesize longer, modified peptides to enrich training datasets; however, this remains experimentally challenging. An alternative strategy is to synthesize only the long unmodified peptides, where current synthesis methods are more reliable, and then introduce specific modifications chemically or enzymatically. Although this approach is limited to certain PTMs, advances in chemoselective ligation and enzymatic modification systems are expanding the range of modifications that can be installed *ex vivo* (59). Another direction is to incorporate additional high-quality *in vivo* spectra of modified peptides from repositories such as MassIVE-KB and PRIDE. While this could increase coverage and diversity, it introduces notable complications: FDR is typically controlled only within individual projects, and although the MassIVE-KB (v2) spectral library re-identifies datasets and re-assesses confidence scores to ensure cross-project comparability, it still contains only a small number of confidently identified modified peptides. More broadly, database-derived “ground truth” from *in vivo* experiments can be particularly unreliable for PTMs, where errors in site localization may directly translate into labels that misguide model training. Beyond the acquisition of new data, data-augmentation strategies, such as those explored in recent

Prosit work with expanded PTM coverage (35), could offer a complementary approach to broaden the training distribution without relying solely on experimentally derived PSMs. Overall, combining richer PTM-datasets consisting of longer peptides with non-autoregressive decoding strategies may offer a promising direction to further reduce length-related constraints.

In addition to token-expansion approaches, including our work, π-PrimeNovo-PTM, and the recent model InstaNovo-P (33), one avenue for enhancement lies in the integration of open modification search tools, which could further expand PTM coverage without requiring explicit enumeration of all PTM-amino acid pairs during training. In this context, leveraging embeddings from chemical foundation models could enable representations that go beyond the mere residue masses, potentially resolving current ambiguities between residues of identical mass, such as isoleucine versus leucine, or O-GalNAc versus O-GlcNAc. In addition, coupling Modanovo with tools such as Prosit or data-driven rescoring pipelines such as Oktoberfest (60) may improve site localization by adding additional information such as retention times, particularly in spectra lacking strong site-determining fragment ions. The development and evaluation of such future tools could readily leverage the development dataset we provide, along with its splits into training, validation, and test sets.

In this study, we have applied Modanovo to datasets for which the reference proteome is known *a priori*. Remarkably, this still showed added value over database search. The advantages of DNPS are expected to be even more pronounced in scenarios where the reference proteome is poorly annotated, undergoes adaptive mutations, i.e., through selective pressure, or when proteome sequence information is unavailable. These scenarios arise in studies of how RNA modifications affect protein sequence, in rapidly mutating tumours and RNA viruses, and in phosphometaproteomics, where diverse microbial proteomes and widespread phosphorylation pose major challenges for database-driven approaches.

DATA AVAILABILITY

No original data were created for this study. The raw files and MaxQuant identification files of the 21-PTM and the Monkeypox virus datasets were obtained through PRIDE accessions PXD009449, PXD040811, and PXD040889. The spectra and ground truth peptides of the MassIVE-KB datasets (v1 and v2) were downloaded from https://massive.ucsd.edu/ProteoSAFe/static/massive-kb-libraries.jsp. Spectral data and sequence annotations for the MULTI-PTM dataset were downloaded from Zenodo (61).

Supplemental Data—This article contains supplemental data.

Acknowledgments—We would like to thank the members of the Gagneur lab and Wilhelm lab for their valuable feedback, especially Johannes Hingerl, Joel Lapin, and Ludwig Lautenbacher. We are also grateful to members of the Kuster and Wilhelm lab for their support with data transfer, and to the Casanovo team for sharing their train/validation/test data splits.

Author Contributions—J. G., M. W., and D. K. A. conceptualization; D. K. A. data curation; D. K. A., Y. B., C. N., W. G., and V. B. formal analysis; J. G. and M. W. funding acquisition; D. K. A., Y. B., and C. N. investigation; J. G., M. W., and D. K. A. methodology; J. G. and M. W. project administration; J. G. resources; D. K. A., Y. B., C. N., and W. G. software; J. G. and M. W. supervision; D. K. A., V. B., and A. P. validation; D. K. A. and V. B. visualization; J. G., M. W., C. N., V. B., and D. K. A. writing–original draft; J. G., M. W., D. K. A., Y. B., C. N., W. G., and V. B. writing–review & editing.

Funding and Additional Information—This study was supported by the German Bundesministerium für Bildung und Forschung (BMBF) supported through the project CLINSPECT-M [16LW0243 K to Y. B., D. K., J. G.]; by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via the IT Infrastructure for Computational Molecular Medicine [#461264291] and the collaborative research centers [TRR179-TP24, TRR237-A07, TRR353-B04 to A. P.]; by the European Research Council via an ERC Starting Grant [#101077037 to WG and MW] and by the State of Bavaria (BayVFP 2024–2027 "P3M"). VB acknowledges support by the Slovenian Research and Innovation Agency (P1-0242, J4-60082). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or any other granting authority, which cannot be held responsible for them.

Conflict of Interests—The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:Mathias Wilhelm is the founder and shareholder of MSAID GmbH and a scientific advisor of OmicScouts GmbH, a Momentum Biotechnologies Company, with no operational role in either company. The remaining authors declare no competing interests.

Abbreviations—The abbreviations used are: AUPCC, area under the precision-coverage curve; DNPS, de novo peptide sequencing; FDR, false discovery rate; HCD, higher-energy collision dissociation; HFF, human foreskin fibroblast; m/z, mass-to-charge; MPXV, monkeypox virus; PSM, peptide-spectrum match; PTM, post-translational modification; SA, spectral angle.

Received September 16, 2025, and in revised form, December 17, 2025 Published, MCPRO Papers in Press, December 24, 2025, https://doi.org/10.1016/j.mcpro.2025.101501

REFERENCES

1. Pagel, O., Loroch, S., Sickmann, A., and Zahedi, R. P. (2015) Current strategies and findings in clinically relevant post-translational modification-specific proteomics. *Expert Rev. Proteomics* **12**, 235–253
2. Klostermeier, D., and Rudolph, M. G. (2017) *Biophysical Chemistry*, 1st ed., CRC Press, Boca Raton, FL
3. Ramazi, S., and Zahiri, J. (2021) Post-translational modifications in proteins: resources, tools and prediction methods. *Database* **2021**, baab012
4. Aebersold, R., and Mann, M. (2016) Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355
5. Miller, R. M., and Smith, L. M. (2023) Overview and considerations in bottom-up proteomics. *Analyst* **148**, 475–486
6. Steen, H., and Mann, M. (2004) The abc's (and xyz's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol* **5**, 699–711
7. Duong, V. A., and Lee, H. (2023) Bottom-Up proteomics: advancements in sample preparation. *Int. J. Mol. Sci.* **24**, 5350
8. Dupree, E. J., Jayathirtha, M., Yorkey, H., Mihasan, M., Petre, B. A., and Darie, C. C. (2020) A critical review of Bottom-Up proteomics: the good, the bad, and the future of this field. *Proteomes* **8**, 14
9. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M. C., and Yates, I. I. I., JR. (2013) Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* **113**, 2343–2394
10. Shuken, S. R. (2023) An introduction to mass spectrometry-based proteomics. *J. Proteome Res.* **22**, 2151–2171
11. Dančík, V., Addona, T. A., Clauser, K. R., Vath, J. E., and Pevzner, P. A. (1999) De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **6**, 327–342
12. Taylor, J. A., and Johnson, R. S. (1997) Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **11**, 1067–1075
13. Muth, T., and Renard, B. Y. (2018) Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Brief Bioinform* **19**, 954–970
14. Huang, T., Wang, J., Yu, W., and He, Z. (2012) Protein inference: a review. *Brief Bioinform* **13**, 586–614
15. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989
16. Eng, J. K., Searle, B. C., Clauser, K. R., and Tabb, D. L. (2011) A face in the crowd: recognizing peptides through database search. *Mol. Cell Proteomics* **10**, R111.009522
17. Huang, X., Huang, L., Peng, H., Guru, A., Xue, W., Hong, S. Y., *et al*. (2013) ISPTM: an iterative search algorithm for systematic identification of post-translational modifications from complex proteome mixtures. *J. Proteome Res.* **12**, 3831–3842
18. Medzihradszky, K. F., and Chalkley, R. J. (2015) Lessons in *de novo* peptide sequencing by tandem mass spectrometry. *Mass Spectrom. Rev.* **34**, 43–63
19. Yilmaz, M., Fondrie, W. E., Bittremieux, W., Melendez, C. F., Nelson, R., Ananth, V., *et al*. (2024) Sequence-to-sequence translation from mass spectra to peptides with a transformer model. *Nat. Commun.* **15**, 6427
20. Straub, G., Ananth, V., Fondrie, W. E., Hsu, C., Klaproth-Andrade, D., Riffle, M., *et al*. (2025) Improvements to Casanovo, a deep learning *de novo* peptide sequencer. *Mol. Biol.*
21. Yilmaz, M., Fondrie, W., Bittremieux, W., Oh, S., and Noble, W. S. (2022) De novo mass spectrometry peptide sequencing with a transformer model. In: *International Conference on Machine Learning*, PMLR, Baltimore, MD: 25514–25522
22. Bittremieux, W., Ananth, V., Fondrie, W. E., Melendez, C., Pominova, M., Sanders, J., *et al*. (2024) Deep Learning Methods for De Novo Peptide Sequencing. *Mass Spectrom. Rev.*
23. Eloff, K., Kalogeropoulos, K., Mabona, A., Morell, O., Catzel, R., Rivera-de-Torre, E., *et al*. (2025) InstaNovo enables diffusion-powered de novo peptide sequencing in large-scale proteomics experiments. *Nat. Mach Intell.* **7**, 565–579
24. Xia, J., Chen, S., Zhou, J., Shan, X., Du, W., Gao, Z., *et al*. (2024) AdaNovo: towards Robust De Novo Peptide Sequencing in Proteomics against Data Biases. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., *et al*., eds. *Advances in Neural Information Processing Systems*, Curran Associates, Inc.: 1811–1828
25. Zhang, X., Ling, T., Jin, Z., Xu, S., Gao, Z., Sun, B., *et al*. (2025) π-PrimeNovo: an accurate and efficient non-autoregressive deep learning model for de novo peptide sequencing. *Nat. Commun.* **16**, 267
26. Petrovskiy, D. V., Nikolsky, K. S., Kulikova, L. I., Rudnev, V. R., Butkova, T. V., Malsagova, K. A., *et al*. (2024) PowerNovo: de novo peptide sequencing via tandem mass spectrometry using an ensemble of transformer and BERT models. *Sci. Rep.* **14**, 15000
27. Correction to: introducing π-HelixNovo for practical large-scale de novo peptide sequencing. *Brief Bioinform* **25**, (2024), bbae134
28. Lee, S., and Kim, H. (2024) Bidirectional de novo peptide sequencing using a transformer modelBittremieux, W., ed.. *PLoS Comput. Biol.* **20**, e1011892
29. Jin, Z., Xu, S., Zhang, X., Ling, T., Dong, N., Ouyang, W., *et al*. (2024) ContraNovo: a contrastive learning approach to enhance De Novo peptide sequencing. *Proc. AAAI Conf. Artif. Intell.* **38**, 144–152
30. Wang, M., Wang, J., Carver, J., Pullman, B. S., Cha, S. W., and Bandeira, N. (2018) Assembling the community-scale discoverable human proteome. *Cell Syst.* **7**, 412–421
31. Zolg, D. P., Wilhelm, M., Schmidt, T., Médard, G., Zerweck, J., Knaute, T., *et al*. (2018) ProteomeTools: systematic characterization of 21 post-translational protein modifications by liquid chromatography tandem mass spectrometry (LC-MS/MS) using synthetic peptides. *Mol. Cell Proteomics* **17**, 1850–1863
32. Xu, J. Y., Zhang, C., Wang, X., Zhai, L., Ma, Y., Mao, Y., *et al*. (2020) Integrative proteomic characterization of human lung adenocarcinoma. *Cell* **182**, 245–261.e17
33. Lauridsen, J., Ramasamy, P., Catzel, R., Canbay, V., Mabona, A., Eloff, K., *et al*. (2025) *InstaNovo-P: A De Novo Peptide Sequencing Model for Phosphoproteomics*. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
34. Gabriel, W., Shouman, O., Schroeder, A., Boessl, F., and Wilhelm, M. (2024) PROSPECT PTMs: Rich labeled tandem mass spectrometry dataset of modified peptides for machine learning in proteomics. *Adv. Neural Inf. Process. Syst.* **37**
35. Gabriel, W., Zolg, D. P., Giurcoiu, V., Shouman, O., Prokofeva, P., Seefried, F., *et al*. (2025) Learning the unseen: data-augmented deep learning for PTM discovery with prosit-PTM. *bioRxiv*. https://doi.org/10.1101/2025.11.07.687302
36. Zolg, D. P., Wilhelm, M., Schnatbaum, K., Zerweck, J., Knaute, T., Delanghe, B., *et al*. (2017) Building ProteomeTools based on a complete synthetic human proteome. *Nat. Methods* **14**, 259–262
37. Huang, Y., Bergant, V., Grass, V., Emslander, Q., Hamad, M. S., Hubel, P., *et al*. (2024) Multi-omics characterization of the monkeypox virus infection. *Nat. Commun.* **15**, 6778
38. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., and Nesvizhskii, A. I. (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat. Methods* **14**, 513–520
39. Yu, F., Teo, G. C., Kong, A. T., Haynes, S. E., Avtonomov, D. M., Geiszler, D. J., *et al*. (2020) Identification of modified peptides using localization-aware open search. *Nat. Commun.* **11**, 4065
40. *Andradesalazar. gagneurlab/Modanovo: Modanovo v1.0.0*. (2025). Zenodo
41. Klaproth-Andrade, D., Bruns, Y., Nix, C., Gabriel, W., Wilhelm, M., and Gagneur, J. (2025) *Modanovo-Model (Revision e93ba29)*. Hugging Face
42. Klaproth-Andrade, D., Bruns, Y., Nix, C., Gabriel, W., Wilhelm, M., and Gagneur, J. (2025) *Modanovo-Development-Dataset (Revision 40dc674)*. Hugging Face
43. Klaproth-Andrade, D., Bruns, Y., Nix, C., Gabriel, W., Wilhelm, M., and Gagneur, J. (2025) *Modanovo: Model Weights and Identification Files*. Zenodo
44. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372
45. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., *et al*. (2011) Scikit-learn: machine learning in python. *J. Mach Learn Res.* **12**, 2825–2830
46. Gessulat, S., Schmidt, T., Zolg, D. P., Samaras, P., Schnatbaum, K., Zerweck, J., *et al*. (2019) Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518

47. Lautenbacher, L., Yang, K. L., Kockmann, T., Panse, C., Chambers, M., Kahl, E., *et al*. (2025) Koina: democratizing machine learning for proteomics research. *Nat Commun* **16**, 9933
48. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., *et al*. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589
49. Schrödinger, L. L. C. (2015) *The PyMOL Molecular Graphics System*. New York, NY, Version 1.8
50. Geiszler, D. J., Kong, A. T., Avtonomov, D. M., Yu, F., Leprevost, F. D. V., and Nesvizhskii, A. I. (2021) PTM-Shepherd: analysis and summarization of post-translational and chemical modifications from open search results. *Mol. Cell Proteomics* **20**, 100018
51. Sun, W., Zhang, Q., Zhang, X., Tran, N. H., Ziaur Rahman, M., Chen, Z., *et al*. (2023) Glycopeptide database search and de novo sequencing with PEAKS GlycanFinder enable highly sensitive glycoproteomics. *Nat. Commun.* **14**, 4046
52. Creasy, D. M., and Cottrell, J. S. (2004) Unimod: protein modifications for mass spectrometry. *Proteomics* **4**, 1534–1536
53. Soday, L., Lu, Y., Albarnaz, J. D., Davies, C. T. R., Antrobus, R., Smith, G. L., *et al*. (2019) Quantitative temporal proteomic analysis of Vaccinia virus infection reveals regulation of histone deacetylases by an interferon antagonist. *Cell Rep.* **27**, 1920–1933.e7
54. Matson, J., Chou, W., Ngo, T., and Gershon, P. D. (2014) Static and dynamic protein phosphorylation in the Vaccinia virion. *Virology* **452–453**, 310–323
55. Brown, N. G., Nick Morrice, D., Beaud, G., Hardie, G., and Leader, D. P. (2000) Identification of sites phosphorylated by the vaccinia virus B1R kinase in viral protein H5R. *BMC Biochem.* **1**, 2
56. Novy, K., Kilcher, S., Omasits, U., Bleck, C. K. E., Beerli, C., Vowinckel, J., *et al*. (2018) Proteotype profiling unmasks a viral signalling network essential for poxvirus assembly and transcriptional competence. *Nat. Microbiol.* **3**, 588–599
57. D'Costa, S. M., Bainbridge, T. W., Kato, S. E., Prins, C., Kelley, K., and Condit, R. C. (2010) Vaccinia H5 is a multifunctional protein involved in viral DNA replication, postreplicative gene transcription, and virion morphogenesis. *Virology* **401**, 49–60
58. Ngo, T., Mirzakhanyan, Y., Moussatche, N., and Gershon, P. D. (2016) Protein primary structure of the vaccinia virion at increased resolution. *J. Virol.* **90**, 9905–9919
59. Schlatzer, T., Kriegesmann, J., Schröder, H., Trobe, M., Lembacher-Fadum, C., Santner, S., *et al*. (2019) Labeling and natural post-translational modification of peptides and proteins via chemoselective Pd-Catalyzed prenylation of cysteine. *J. Am. Chem. Soc.* **141**, 14931–14937
60. Picciani, M., Gabriel, W., Giurcoiu, V., Shouman, O., Hamood, F., Lautenbacher, L., *et al*. (2024) Oktoberfest: open-source spectral library generation and rescoring pipeline based on Prosit. *Proteomics* **24**, e2300112
61. Gabriel, W., Shouman, O., and Wilhelm, M. (2024) *PROSPECT Ptms: Multi-Ptm*. Zenodo