

## Acknowledgments

First and foremost, I would like to express my deepest gratitude to **Daniela Klaproth-Andrade**. Thank you for your unwavering support at any time of the day, for patiently dealing with my moments of despair at the most frustrating hurdles, and for constantly reminding me to stay rational and focused when things got complicated. Your guidance was the backbone of this work.

I am sincerely grateful to **Prof. Dr. Julien Gagneur** and **Prof. Dr. Mathias Wilhelm** for their incredible expertise and sharp analytical insights. Your conceptual guidance and ability to prioritize the most impactful directions were essential in shaping this thesis and pushing it to its full potential.

A special thanks goes to my colleagues and friends **Yanik Bruns**, **Niklas Feuerstein**, **Franziska Koller**, **Samuel Khan**, and **Wassim Gabriel**. Thank you for the countless ideas, the technical help, and the inspiring discussions that enriched this project.

I would also like to thank **Prof. Dr. Bernhard Küster**, **Cecilia Jensen**, and **Johanna Tüshaus** from the Küster Lab. Your profound biological expertise and the swift provision of the glioma datasets were crucial for the success of this research. Thank you for always being available for quick answers and valuable feedback.

Finally, I want to thank my **family and friends**. Thank you for your endless mental support, for keeping me grounded, and for the hours spent proofreading this thesis. Your encouragement made this journey possible.



## Abstract

Mass spectrometry (MS)-based proteomics is an essential tool in clinical cancer research, providing functional insights into tumor mechanisms that extend beyond genomic data. While database-driven identification strategies remain the gold standard, they are inherently limited by a predefined search space, often overlooking rare post-translational modifications (PTMs) and single amino acids variations caused by genomic mutations. De novo peptide sequencing, particularly through modern deep learning architectures like Transformers, offers an unbiased alternative. However, these models currently struggle with Tandem Mass Tag (TMT) labeled data due to systematic mass shifts and altered fragmentation patterns.

In this thesis, an adaptation of the Transformer-based model *Modanovo* is presented to bridge the gap between de novo sequencing and TMT multiplexing. By expanding the vocabulary to include TMT-specific tokens and integrating a covariate embedding for TMT status, the model was trained to explicitly account for the chemical signatures of the label. Fine-tuning was performed on a comprehensive dataset consisting of 82% TMT-labeled multi-PTM spectra and an 18% “replay set” of unlabeled data to prevent catastrophic forgetting.

The results on the test dataset demonstrate robust performance with an Area Under the Precision-Recall Curve (AUPCC) of 0.89 for TMT data. Notably, the identification of modifications such as ubiquitination and monomethylation was partially enhanced in comparison to non-TMT data. Iterative hyperparameter optimization, specifically expanding the isotope error range and utilizing a beam search size of 5, proved crucial for managing increased spectral complexity.

Applied to an independent glioma TMT dataset comprising 6.48 million spectra, the model significantly expanded the identified proteome. Compared to MaxQuant, the de novo approach identified 40,000 additional unique peptides and provided high-confidence sequence suggestions for over 670,000 previously unidentified spectra. Biological highlights include the identification of 694 genetically validated SNPs and the detection of phosphorylation at Serine 15 of PYGL, a key metabolic switch for tumor cell survival under hypoxia. This work demonstrates that integrating TMT-specific knowledge into Transformer models unlocks the “dark proteome” of clinical samples, offering a scalable platform for personalized proteogenomics.



## Kurzzusammenfassung

Die Massenspektrometrie (MS)-basierte Proteomik ist ein essenzielles Werkzeug in der klinischen Krebsforschung, da sie funktionelle Einblicke in Tumormechanismen ermöglicht, die über genomische Daten hinausgehen. Während datenbankgestützte Identifizierungsstrategien weiterhin der Goldstandard sind, unterliegen sie einer inhärenten Beschränkung durch den vordefinierten Suchraum. Dabei werden seltene posttranskriptionale Modifikationen (PTMs) und durch genomische Mutationen verursachte Aminosäurevariationen (SNPs) häufig übersehen. De-novo-Peptidsequenzierung, insbesondere durch moderne Deep-Learning-Architekturen wie Transformer, bietet hier eine unvorensehbare Alternative. Aktuelle Modelle haben jedoch Schwierigkeiten mit Tandem Mass Tag (TMT)-markierten Daten, bedingt durch systematische Massenverschiebungen und veränderte Fragmentierungsmuster.

In dieser Arbeit wird eine Adaption des Transformer-basierten Modells Modanovo vorgestellt, um die Lücke zwischen De-novo-Sequenzierung und TMT-Multiplexing zu schließen. Durch die Erweiterung des Vokabulars um TMT-spezifische Tokens und die Integration eines Kovariaten-Embeddings für den TMT-Status wurde das Modell darauf trainiert, die chemischen Signaturen der Markierung explizit zu berücksichtigen. Das Fine-Tuning erfolgte auf einem umfassenden Datensatz, bestehend aus 82

Die Ergebnisse auf dem Testdatensatz zeigen eine robuste Performance mit einer Area Under the Precision-Recall Curve (AUPRC) von 0,89 für TMT-Daten. Bemerkenswerterweise wurde die Identifizierung von Modifikationen wie Ubiquitinierung und Monomethylierung im Vergleich zu Nicht-TMT-Daten teilweise verbessert. Eine iterative Hyperparameter-Optimierung, insbesondere die Erweiterung des Isotopenfehlerspektrums und die Nutzung einer Beam-Search-Größe von 5, erwies sich als entscheidend für den Umgang mit der erhöhten spektralen Komplexität.

Angewendet auf einen unabhängigen Gliom-TMT-Datensatz mit 6,48 Millionen Spektren, konnte das Modell das identifizierte Proteom signifikant erweitern. Im Vergleich zu MaxQuant identifizierte der De-novo-Ansatz 40.000 zusätzliche einzigartige Peptide und lieferte hochkonfidente Sequenzvorschläge für über 670.000 zuvor nicht identifizierte Spektren. Zu den biologischen Highlights zählen die Identifizierung von 694 genomisch validierten SNPs sowie der Nachweis der Phosphorylierung an Serin 15 von PYGL – einem zentralen Stoffwechselregulator für das Überleben von Tumorzellen unter Hypoxie. Diese Arbeit zeigt, dass die Integration von TMT-spezifischem Wissen in Transformer-Modelle das „dunkle Proteom“ klinischer Proben erschließt und eine skalierbare Plattform für die personalisierte Proteogenomik bietet.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	The Challenge of the “Dark Proteome” and PTMs . . . . .	1
1.2	De Novo Sequencing and the TMT Gap . . . . .	1
1.3	Objectives of this Thesis . . . . .	2
<b>2</b>	<b>Background</b>	<b>3</b>
2.1	Mass Spectrometry-Based Proteomics . . . . .	3
2.1.1	Bottom-Up Proteomics Workflow . . . . .	3
2.1.2	Tandem Mass Spectrometry (MS/MS) and Peptide Fragment Ion Theory . . . . .	3
2.1.3	Tandem Mass Tag (TMT) Labeling . . . . .	4
2.1.4	Impact of TMT on Fragmentation Patterns . . . . .	4
2.2	Peptide Identification Strategies . . . . .	5
2.2.1	Database Search Engines . . . . .	5
2.2.2	De Novo Peptide Sequencing . . . . .	5
2.2.3	Classical De Novo Peptide Sequencing Approaches . . . . .	6
2.3	Deep Learning and Transformer Models in de novo sequencing . . . . .	6
2.3.1	The new modeling approach . . . . .	6
2.3.2	Transformer Architecture and Self-Attention . . . . .	7
2.3.3	Spectrum Encoding and Embedding . . . . .	7
2.3.4	Autoregressive Sequence Generation and Beam Search . . . . .	7
<b>3</b>	<b>Methods</b>	<b>9</b>
3.1	Fine-tuning Strategy and Model Adaptations . . . . .	9
3.1.1	Tokenization and Vocabulary Expansion . . . . .	9
3.1.2	TMT Covariate Embedding . . . . .	10
3.1.3	Multi-task Learning Architecture . . . . .	11
3.2	Finetuning Datasets . . . . .	12
3.2.1	Data Selection and Composition . . . . .	12
3.2.2	TMT-labeled Data . . . . .	12
3.2.3	Non-TMT Data (Replay Set) . . . . .	13
3.2.4	Modification Distribution and Heatmap Analysis . . . . .	13
3.2.5	Quality Filtering and Pre-processing . . . . .	14

3.3	Training Protocol . . . . .	15
3.4	Evaluation Strategy and Performance Metrics . . . . .	16
3.4.1	Confidence Scoring and Peptide Ranking . . . . .	16
3.4.2	Precision-Coverage Analysis and Stratification . . . . .	16
3.5	Application to Clinical Glioma Data . . . . .	17
3.5.1	Experimental Dataset: The ClinSpect-M Glioma Cohort . . . . .	17
3.5.2	Data Acquisition and Large-Scale Preprocessing . . . . .	17
3.5.3	Peptide Identification Pipeline . . . . .	18
3.5.4	Peptide Alignment and Sequence Validation . . . . .	19
3.5.5	Genomic Validation and Integration of Panel Sequencing Data . . . . .	21
3.6	Spectral Evidence and Intensity Prediction . . . . .	22
3.6.1	Spectral Angle (SA) as a Quality Metric . . . . .	22
3.6.2	Intensity Prediction via Koina and Prosit . . . . .	23
<b>4</b>	<b>Results</b>	<b>25</b>
4.1	Model Training and Convergence . . . . .	25
4.2	Performance Evaluation on the test set . . . . .	25
4.2.1	Comparative Performance: Non-TMT and Modanovo Baseline . . . . .	26
4.2.2	Generalization on TMT-Labeled Data . . . . .	27
4.3	Hyperparameter Optimization and Isotopic Sensitivity . . . . .	27
4.3.1	Isotope Error Range and Beam Search Synergy . . . . .	28
4.3.2	Quantitative Impact of Configuration Changes . . . . .	28
4.3.3	Impact on Specific Modification Classes . . . . .	28
4.4	Application on independent Glioma TMT Dataset . . . . .	29
4.4.1	Proteome Alignment and Score Calibration . . . . .	29
4.4.2	Stratified Performance and Modification Stability . . . . .	30
4.4.3	Summary of Validation . . . . .	32
4.5	Uncovering the Dark Proteome: Comparison with MaxQuant . . . . .	32
4.5.1	Unique Peptide Identifications . . . . .	32
4.5.2	Identification of Previously Unidentified Spectra . . . . .	33
4.5.3	From Global Discovery to Specific Variants . . . . .	34
4.5.4	Discovery of SNP-Related Variants . . . . .	34
4.5.5	Discovery of PTM Sites . . . . .	36
<b>5</b>	<b>Discussion</b>	<b>41</b>
5.1	Key Findings . . . . .	41
5.2	Limitations and Methodological Constraints . . . . .	41
5.2.1	Future Improvements and Clinical Utility . . . . .	42

<b>Supplementary Material</b>	<b>43</b>
5.3    Zusätzliche Abbildungen . . . . .	43
5.4    Zusätzliche Tabellen . . . . .	43

x

# List of Figures

3.1	Schematic representation of the adapted transformer architecture based on Casanova. Allowing the identification of tmt-labeled post-translationally modified peptides directly from tandem mass (MS2) spectra. The model is trained starting with weight initialization from Modanovo's pre-trained weights. The model components for the amino acid (AA) embeddings and final linear layer are expanded to allow the TMT-tokens. The TMT status is fed as a covariate embedding into the decoder alongside the precursor and previous predicted residues. . . . .	11
3.2	Composition of the fine-tuning dataset. The training dataset is dominated by TMT-labeled multi-PTM peptide-spectrum matches (82%), supplemented by a 18% Replay Set to preserve model robustness and recall on unlabeled spectra. . . . .	12
3.3	Distribution of identified post-translational modifications (PTMs) across specific amino acid residues. The heatmaps display log-scaled spectral counts for various modifications as a function of the modified residue. (a) shows the TMT-labeled dataset (b) represents the Replay Set . . . . .	14
4.1	Precision-coverage curves at the peptide level across different PTMs. PTM types (Acetylation, Citrullination (Citrull_deamid), Monomethylation, OGlcNAc/OGlcNAc (O-Glyco), Oxidation, Phosphorylation, Pyro-glu, and Ubiquitination) are shown in the different panels. Colors represent the performance of the final model on TMT-labeled (orange) and non-TMT (blue) spectra. Unmodified peptides are shown in light grey for comparison. . .	26

4.2	Cumulative alignment of <i>de novo</i> predictions to the human reference proteome across confidence score thresholds. The plot illustrates the proportion of sequences matching the reference with zero (blue), up to one (orange), and up to two (green) amino acid substitutions. The segments between the curves represent peptides with specific mismatch counts, potentially accounting for Single Nucleotide Polymorphisms (SNPs) or technical sequencing errors. The increasing trend demonstrates the correlation between the model's self-reported confidence and sequence accuracy.	30
4.3	Proportion of perfectly aligned PSMs stratified by modification type and ranked by confidence score. . . . .	31
4.4	Stratified precision-rank analysis of the adapted model. The plot shows the proportion of Peptide-Spectrum Matches (PSMs) achieving a perfect <i>blastp</i> alignment to the human reference proteome, ranked by descending model confidence scores. Curves are stratified by modification type: unmodified or just oxidation/acetylation(green), phosphorylation (red), monomethylation (blue), and other PTMs (purple). Symbols represent specific confidence score thresholds ( $\geq 0.80, 0.90, 0.95$ ). While performance remains high (> 90%) for most classes at high confidence, monomethylated peptides show a significant precision drop, indicating potential challenges in mass-equivalent deconvolution or model overconfidence for this specific modification. . . . .	31
4.5	Overlap of unique peptide identifications between the adapted <i>de novo</i> model and MaxQuant. The model successfully recovers the majority of MQ identifications while contributing 40,000 additional unique sequences. . . . .	33
4.6	Left: Probit-predicted spectrum (top) and experimental spectrum (bottom) for the reference peptide sequence YAALLK. The cartoon illustrates the relevant nucleotide sequence and fragment ion series assuming the reference genome allele. Right: Same as for left, but for the peptide sequence YAASLK predicted by our model for the same experimental spectrum. The higher degree of overlap between experimental and theoretical spectrum for the mutated peptide (right) validate the models predictipn. . . . .	36
4.7	Distribution of unique phosphorylation sites identified by MaxQuant (MQ) and our model(DNPS) for selected cancer-related proteins. The results demonstrate a significant expansion of the detectable phospho-landscape. . . . .	37

4.8	Confidence distribution for the predictions validating p-Ser15 on PYGL. The identification is supported by 279 PSMs with a mean spectral angle (SA) of 0.7, indicating high reproducibility and spectral fidelity. . . . .	39
S1	JUST FOR TESTING PRUPOSES,Histogram of peptide lengths for all data . . . . .	43



# List of Tables

4.1 Comparison of model performance across different configurations. 28



# 1 Introduction

## 1.1 Proteins and PTMs in Cancer

Proteins are the primary functional units of the cell, executing the vast majority of biological processes. While the genome provides the blueprint, the proteome reflects the actual physiological state of an organism. In the context of complex diseases such as cancer, the importance of studying proteins becomes even more critical. Malignant transformation is often driven not just by the presence of certain proteins, but by their dynamic regulation through post-translational modifications (PTMs) [Mertins2016].

PTMs, such as phosphorylation, acetylation, and methylation, act as molecular switches that control protein activity, localization, and interaction networks. In neuro-oncology, particularly in the study of aggressive brain tumors like glioma, aberrant phosphorylation patterns are known to drive oncogenic signaling pathways, contributing to tumor growth and therapy resistance [Smith2019]. Identifying these modified proteins is essential for understanding the molecular landscape of a patient’s tumor and for the development of targeted therapies. However, a significant portion of these critical biological signals remains hidden from researchers—a phenomenon often described as the “dark proteome.”

## 1.2 Mass Spectrometry: state of the art to analyze the Proteome

To decipher this complexity, Mass Spectrometry (MS)-based proteomics has emerged as the gold-standard technology. In a typical “bottom-up” workflow, proteins extracted from clinical samples, such as glioma biopsies, are digested into smaller fragments called peptides. These peptides are then analyzed by a mass spectrometer, which measures their mass-to-charge ratio ( $m/z$ ).

In a process known as Tandem Mass Spectrometry (MS/MS), specific peptides are isolated and energetically fragmented into even smaller pieces. The resulting MS/MS spectrum is a “fingerprint” of the peptide, showing the masses of its fragments. By analyzing the gaps between these fragment peaks, one can

theoretically reconstruct the amino acid sequence. To increase efficiency in clinical studies, multiple patient samples are often labeled with Tandem Mass Tags (TMT). TMT is a chemical labeling technology that allows for the simultaneous analysis (multiplexing) of up to 18 samples in a single experiment, enabling precise comparison of protein levels across different patients [Thompson2003].

### 1.3 Mass Spectrometry: State of the Art to Analyze the Proteome

To decipher the complexity of the proteome, Mass Spectrometry (MS)-based proteomics has emerged as the gold-standard technology. At its core, a mass spectrometer acts as an extremely precise molecular scale. In a typical “bottom-up” workflow, proteins are extracted from biological samples and digested into smaller fragments called peptides [Aebersold2016]. These peptides are easier to measure and serve as proxies for the original proteins.

The identification of these peptides occurs via Tandem Mass Spectrometry (MS/MS). In this process, the instrument first weighs the intact peptide and then energetically breaks it into smaller fragment ions. The resulting MS/MS spectrum is a “fingerprint” showing the masses of these fragments. By analyzing the gaps between these mass peaks, which correspond to specific amino acids, the peptide’s sequence can be reconstructed [Steen2004].

To increase efficiency in clinical or large-scale research, multiple samples are often labeled with Tandem Mass Tags (TMT). This chemical labeling technology allows for the simultaneous analysis (multiplexing) of up to 18 samples in a single experiment, enabling a direct comparison of protein levels across different conditions or patients [Thompson2003].

### 1.4 Database-Driven Searches

The standard way of interpreting the spectrum (output of the mass spectrometer) is using a database search engine. This method compare experimental spectra against a predefined spectral library of known peptide sequences. While effective for “standard” proteins, database search engines struggle with PTMs. To find a modified peptide, the search engine must be told exactly which modification to look for. Including many possible PTMs leads to a “combinatorial explosion” that exponentially increases computation time and leads to higher false-discovery rates [Chick2015]. Consequently, many spectra originating from unexpected PTMs or single amino acid variants (SAVs) in cancer cells are simply discarded as “unidentified.”

*De novo* peptide sequencing offers a complementary solution by predicting sequences directly from the spectra without needing a database. While recent deep learning models like *Modanova* have revolutionized this field [Klaproth2024], they face a hurdle: they were not designed for the systematic mass shifts and altered fragmentation patterns introduced by TMT labeling. Since TMT is often used for high-quality clinical data, this creates a gap: we have the data to find new cancer insights, but our most advanced discovery tools (*de novo* sequencing) cannot read it correctly.

## 1.5 Objectives of this Thesis

The primary objective of this thesis is not merely to expand a model, but to unlock the biological potential of TMT-labeled clinical datasets through adapted *de novo* sequencing. By bridging the gap between deep learning and multiplexed proteomics, we aim to uncover biological insights in glioma data that remain invisible to standard workflows.

The specific goals of this thesis are:

1. **Model Adaptation and Data Curation:** To adapt the *Modanova* architecture to handle the specific chemical signatures of TMT labeling. A significant part of this goal is the curation of high-quality TMT-labeled datasets to serve as a training foundation.
2. **Uncovering the Dark Proteome in Glioma:** To apply the adapted model to actual patient data from glioma studies. We aim to identify novel PTMs and SAVs that have been missed by database searches, thereby providing a more comprehensive molecular characterization of the tumor.
3. **Clinical and Methodological Impact:** To evaluate the potential of this approach to improve clinical proteogenomics, providing a pathway for more sensitive biomarker discovery and a better understanding of the regulatory networks in cancer.

Ultimately, this thesis explores the potential of *de novo* sequencing to transcend the limitations of predefined databases, enabling a more comprehensive view of the proteome in multiplexed quantitative studies.

## *1 Introduction*

---

## 2 Background

In this chapter, the fundamental principles of mass spectrometry-based proteomics and the computational strategies for peptide identification are discussed. Particular focus is placed on the challenges introduced by chemical labeling and the emergence of deep learning models in *de novo* sequencing.

### 2.1 Mass Spectrometry-Based Proteomics

#### 2.1.1 Bottom-Up Proteomics Workflow

Mass spectrometry (MS)-based proteomics has become the gold standard for the large-scale analysis of proteins in complex biological samples. The most widely adopted strategy is the “bottom-up” approach. In this workflow, proteins are extracted from a biological source and enzymatically digested—typically using trypsin—into smaller peptides before being analyzed by the mass spectrometer [Gevaert2003]. This enzymatic cleavage is essential because peptides are easier to fractionate, ionize, and fragment than intact proteins. Following digestion, the resulting peptide mixture is usually separated by liquid chromatography (LC) and ionized (e.g., via Electrospray Ionization, ESI) to be transferred into the gas phase for analysis [Aebersold2016].

The analysis occurs in two primary stages within the mass spectrometer. In the first stage (MS1), the instrument measures the mass-to-charge ratio ( $m/z$ ) and intensity of the intact peptides (precursors). This provides a snapshot of the peptide population in a sample at a given time. From this MS1 information, specific precursor ions are selected for the second stage, tandem mass spectrometry (MS2), to extract structural information for identification.

#### 2.1.2 Tandem Mass Spectrometry (MS/MS) and Peptide Fragment Ion Theory

The identification of the amino acid sequence is achieved through Tandem Mass Spectrometry (MS/MS or MS2). In this process, a specific precursor ion is isolated based on its  $m/z$  and subsequently subjected to fragmentation [Steen2004]. In high-resolution instruments like the Orbitrap, Higher-energy

## 2 Background

---

Collisional Dissociation (HCD) is the preferred method. HCD is a beam-type collision-induced dissociation technique where ions collide with an inert gas (e.g., Nitrogen), leading to internal energy buildup and subsequent bond breakage, producing a predictable pattern of fragment ions.

According to the Roepstorff-Fohlman nomenclature for peptide fragment ion theory, the fragmentation of the peptide backbone occurs primarily at the amide bonds. In the experimental setting of HCD, this gas-phase dissociation predominantly targets the peptide bonds. This results in two main series of ions: b-ions, where the charge remains on the N-terminal fragment, and y-ions, where the charge remains on the C-terminal fragment [Roepstorff2010].

### 2.1.3 Tandem Mass Tag (TMT) Labeling

Tandem Mass Tag (TMT) labeling is a powerful chemical labeling strategy used for high-throughput multiplexed quantitative proteomics. The TMT molecule is an isobaric tag consisting of three functional groups: a reactive NHS-ester group for covalent attachment to peptide N-termini and lysine side chains, a mass reporter group, and a mass normalizer group [Thompson2003].

Because the tags are isobaric, peptides from different biological samples (up to 18-plex) are labeled, pooled, and appear as a single precursor peak in the MS1 scan [Werner2014]. By enabling the simultaneous analysis of up to 18 samples in a single LC-MS/MS run, TMT labeling minimizes technical batch effects and ensures consistent quantification across channels, significantly reducing the “missing value” challenge compared to label-free workflows [Rauniyar2014]. This challenge, common in label-free proteomics, occurs when a peptide is inconsistently detected or fragmented across different runs, leading to incomplete data matrices. In TMT, the co-elution and simultaneous fragmentation of all labeled versions of a peptide ensure that if a signal is detected, quantitative information is typically retrieved for all multiplexed samples. Furthermore, a “carrier channel”—an isobaric spike-in of a high-abundance proteome—can be used to boost the precursor signal of low-input samples, enhancing identification and sequencing depth in single-cell applications [Budnik2018].

### 2.1.4 Impact of TMT on Fragmentation Patterns

The use of TMT tags introduces systematic changes to the peptide fragment spectra. Upon fragmentation via HCD, the isobaric tag cleaves at a specific linker region, releasing the low-molecular-weight reporter ions in the  $m/z$  126–135 range for quantification [McAlister2014]. For identification, TMT labeling presents a challenge: the tag adds a constant mass shift to the N-terminus

and lysine residues. Furthermore, the presence of the bulky tag can alter the gas-phase basicity and fragmentation efficiency, often leading to different relative intensities of b- and y-ions compared to unlabeled peptides [Hogrebe2018].

As we move from the physical process of generating these spectra, the focus shifts to the computational interpretation of this data, which leads to the different identification approaches.

## 2.2 Peptide Identification Strategies

### 2.2.1 Database Search Engines

The most prevalent method for peptide identification is database searching. This strategy relies on a predefined protein sequence database, such as UniProt [TheUniProtConsortium2023]. Computational search engines, including Mascot, SEQUEST, or MaxQuant (Andromeda), perform an *in silico* digestion of these sequences to generate a library of theoretical spectra [Cox2008]. Each experimental MS/MS spectrum is then compared against these theoretical candidates using scoring functions to determine the best match, known as a Peptide-Spectrum Match (PSM) [Eng1994].

While robust, database searching is inherently limited by the predefined sequence search space. Traditional closed searches require an *a priori* definition of both the protein sequences and their expected modifications. Consequently, peptides from non-model organisms, genomic variants, or those carrying unexpected post-translational modifications (PTMs) are frequently missed. Although open search strategies allow for the identification of PTMs without prior definition by allowing larger precursor mass tolerances, the underlying peptide sequences must still be present in the database. Attempting to account for all possible PTMs within a traditional closed search would lead to a combinatorial explosion, drastically increasing false discovery rates (FDR) and computational costs [Nesvizhskii2010].

### 2.2.2 De Novo Peptide Sequencing

In contrast to database-driven methods, *de novo* peptide sequencing reconstructs the amino acid sequence directly from the fragment ion peaks in the MS/MS spectrum without any genomic or proteomic reference [Taylor1997]. This approach maps the mass differences between peaks directly to the masses of amino acids.

By measuring the mass difference between consecutive ions in a series, the corresponding amino acid can be inferred, as each residue (except for the iso-

mers Leucine and Isoleucine) possesses a unique residual mass. For example, a measured mass shift of 113.08 Da identifies Leucine or Isoleucine, whereas a shift of 71.04 Da identifies Alanine [Steen2004]. Historically, however, *de novo* sequencing was limited by spectral noise, low resolution, and incomplete fragmentation.

### 2.2.3 Classical De Novo Peptide Sequencing Approaches

Early *de novo* peptide sequencing algorithms, such as PEAKS, Novor, and PepNovo, are rooted in rule-based modeling of peptide fragmentation [Ma2003]. These methods typically represent an MS/MS spectrum as a spectrum graph, where nodes correspond to observed  $m/z$  values and edges represent mass differences matching specific amino acids. Sequence identification is framed as finding the optimal path through this graph, guided by hand-crafted scoring functions that account for peak intensities and ion types.

However, these classical approaches face significant limitations. They rely heavily on heuristic fragmentation rules which lack flexibility across different mass spectrometry instruments or chemistries. Incomplete fragmentation often leads to “broken” paths in the spectrum graph, while the presence of post-translational modifications (PTMs) or chemical labels like TMT exponentially increases the search space and resulting ambiguity.

## 2.3 Deep Learning and Transformer Models in *de novo* sequencing

The limitations of rule-based *de novo* sequencing approaches motivated the adoption of deep learning methods, which learn peptide fragmentation patterns directly from data. By leveraging large annotated MS/MS datasets, neural networks can model complex, instrument-specific fragmentation behavior and generalize across varying experimental conditions. This data-driven paradigm marked a fundamental shift in *de novo* peptide sequencing.

### 2.3.1 The new modeling approach

The transition to deep learning enabled the modeling of *de novo* sequencing as a sequence-to-sequence (seq2seq) task, translating spectral peak patterns into amino acid sequences. Convolutional Neural Networks (CNNs) were initially used to extract local spectral features, while Recurrent Neural Networks

## 2.3 Deep Learning and Transformer Models in *de novo* sequencing

---

(RNNs), such as Long Short-Term Memory (LSTM) networks, modeled sequential dependencies between amino acids [Tran2017]. Although these models, like DeepNovo, outperformed classical methods in many settings, they exhibited limitations in capturing long-range dependencies and global spectral context, particularly for longer peptides or spectra with sparse fragmentation.

### 2.3.2 Transformer Architecture and Self-Attention

The introduction of the Transformer architecture [Vaswani2017] addressed many of these limitations by replacing recurrence with self-attention mechanisms. Self-attention allows the model to assess relationships between all spectral features simultaneously, enabling the integration of complementary evidence such as b- and y-ion pairs distributed across the full  $m/z$  range.

In proteomics, this capability is especially beneficial, as peptide evidence is often fragmented and non-local. The Casanovo model was a landmark application of Transformers to *de novo* peptide sequencing, employing an encoder-decoder architecture to autoregressively predict peptide sequences [Yilmaz2022]. By combining global spectral context with precursor mass constraints and beam search decoding, Casanovo achieved state-of-the-art performance.

### 2.3.3 Spectrum Encoding and Embedding

A critical component of Transformer-based models is the encoding of MS/MS spectra into suitable input representations. Typically, spectra are transformed into fixed-length embeddings that incorporate  $m/z$  values, intensities, and positional information. These embeddings serve as the input tokens for the Transformer encoder, enabling the model to learn fragmentation-aware representations. Effective spectrum encoding is essential for robust performance, as it directly influences how well the model can distinguish informative peaks from noise and account for variations in fragmentation patterns.

### 2.3.4 Autoregressive Sequence Generation and Beam Search

DeepNovo pioneered the use of CNNs for feature extraction [Tran2017]. However, the autoregressive decoding strategy, where the model predicts one amino acid at a time conditioned on previously predicted residues and the encoded spectrum, was popularized by Casanovo [Yilmaz2022]. Beam search is commonly applied during inference to explore multiple high-probability candidate sequences simultaneously. This approach allows the model to balance local

## *2 Background*

---

confidence with global sequence plausibility while enforcing constraints such as precursor mass consistency. As a result, autoregressive decoding with beam search significantly improves identification accuracy, particularly in ambiguous or noisy spectra [Yilmaz2023].

Following Casanovo, several models have expanded the capabilities of *de novo* sequencing. InstaNovo and  $\pi$ -PrimeNovo introduce architectural innovations that, among other improvements, increase inference speed. To address the limitation of sparse PTM support in early models, Modanovo extended the token vocabulary to include a broad range of amino acid-PTM combinations, demonstrating that Transformers can scale to biologically diverse datasets [KlaprothAndrade2025].

# 3 Methods

## 3.1 Fine-tuning Strategy and Model Adaptations

The approach builds upon the Modanovo framework, a transformer-based architecture designed for the identification of post-translational modifications (PTMs) using experimental spectra [KlaprothAndrade2025].

The transition from unlabeled or label-free spectra to TMT-multiplexed data requires specific adaptations of the underlying deep learning model. In this work, the fine-tuning process involves adjusting the model to recognize TMT labels not as global experimental parameters, but as specific chemical modifications integrated into the sequencing vocabulary.

### 3.1.1 Tokenization and Vocabulary Expansion

To accommodate TMT labeling, the model’s tokenization strategy was expanded. Modanovo utilizes a residue-based vocabulary where each token represents either a standard amino acid or a specific amino acid-PTM combination [KlaprothAndrade2025]. For this study, the configuration was adjusted to include TMT-specific tokens. These tokens account for the fixed mass shifts on N-termini and Lysine (K) residues.

Specifically, the vocabulary was extended by the following residues and their corresponding mass shifts:

- **K[+229.163]**: Lysine with TMT10/16 label.
- **[+229.163]-**: TMT10/16 label at the peptide N-terminus.
- **K[+343.206]**: Lysine with both TMT and GlyGly (ubiquitination) modification.
- **K[+271.173]**: Lysine with both TMT and Acetyl modification.
- **K[+243.179]**: Lysine with both TMT and Methyl modification.

Following the Modanovo initialization protocol, the embeddings for these new tokens were initialized by averaging the embeddings of their

constituent components (e.g., the base amino acid embedding and the modification-specific shift) to leverage pre-learned chemical representations [KlaprothAndrade2025].

#### 3.1.2 TMT Covariate Embedding

To enable the decoder to account for systematic shifts in fragmentation patterns and physicochemical properties induced by TMT labeling, we introduce a categorical conditioning mechanism. This allows the model to explicitly distinguish between TMT-labeled and unlabeled spectra at a global level.

In the baseline transformer architecture (e.g., Casanova [Yilmaz2022]), the precursor embedding ( $\text{prec}_{\text{emb}}$ ) encodes the precursor ion’s mass-to-charge ratio ( $m/z$ ) and charge state ( $z$ ) using sinusoidal positional encodings passed through dedicated linear projection layers.

Analogous to the embedding of precursor features (precursor mass and charge), we define a learnable TMT-specific embedding. For each spectrum, a binary indicator  $f_{\text{TMT}} \in \{0, 1\}$  encodes the presence or absence of TMT labeling and is mapped through an embedding layer:

$$\mathbf{E}_{\text{TMT}} = \text{Embedding}(f_{\text{TMT}}) \in \mathbb{R}^{d_{\text{model}}}.$$

The resulting vector is integrated into the latent representation via additive fusion. Specifically, it is added to the precursor embedding prior to decoding:

$$\text{prec\_emb}_{\text{conditioned}} = \text{prec\_emb} + \mathbf{E}_{\text{TMT}}.$$

By injecting this information at the level of the precursor representation—effectively seeding the start of the decoding process—the transformer can adapt its internal representations to the chemical environment associated with TMT-labeled peptides. This conditioning strategy is computationally efficient, as it preserves the model dimensionality while providing a strong global signal that guides de novo sequencing depending on the labeling state of the sample.

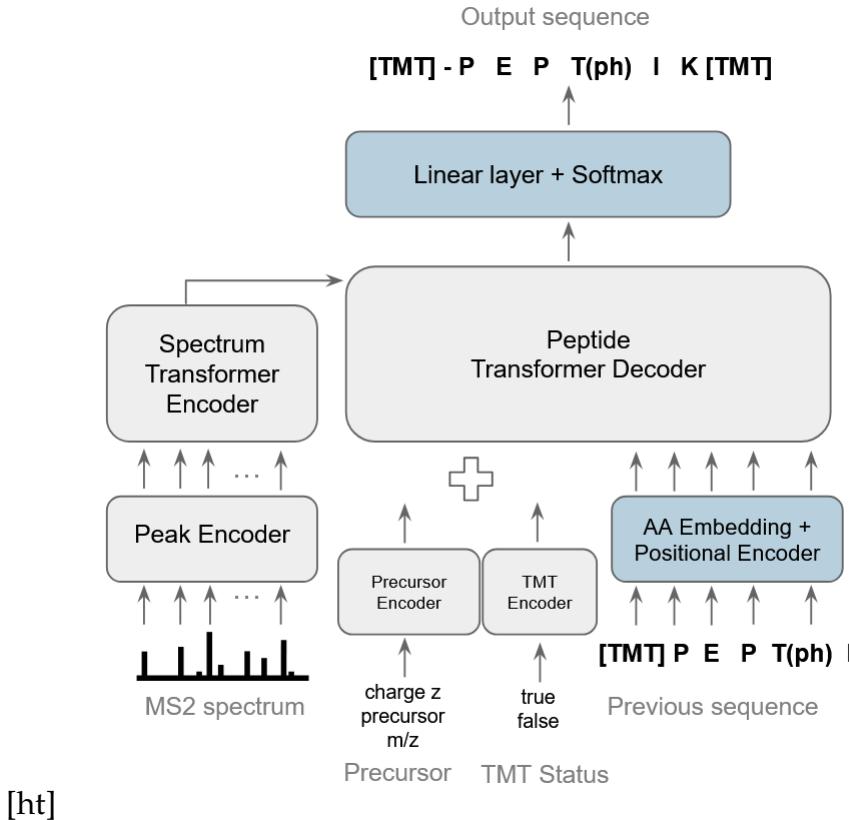


Figure 3.1: Schematic representation of the adapted transformer architecture based on Casanova. Allowing the identification of TMT-labeled post-translationally modified peptides directly from tandem mass (MS2) spectra. The model is trained starting with weight initialization from Modanovo's pre-trained weights. The model components for the amino acid (AA) embeddings and final linear layer are expanded to allow the TMT-tokens. The TMT status is fed as a covariate embedding into the decoder alongside the precursor and previous predicted residues.

## 3.2 Finetuning Datasets

### 3.2.1 Data Selection and Composition

For the training and evaluation of the adapted model, a robust dataset was curated to ensure high-quality spectral representations. The fundamental requirement for supervised learning in this context is the availability of database searching derived sequences associated with high-resolution fragment spectra. Data were integrated from various sources, initially stored in CSV and mzML

### 3 Methods

---

formats, and subsequently compiled into a unified Mascot Generic Format (MGF) file. This format allows for a streamlined input pipeline where the peptide sequence is explicitly linked to its corresponding spectrum [Deutsch2012].

The fine-tuning process utilizes two distinct datasets to adapt the model to TMT-labeled spectra while maintaining performance on unlabeled data (see Figure 3.2).

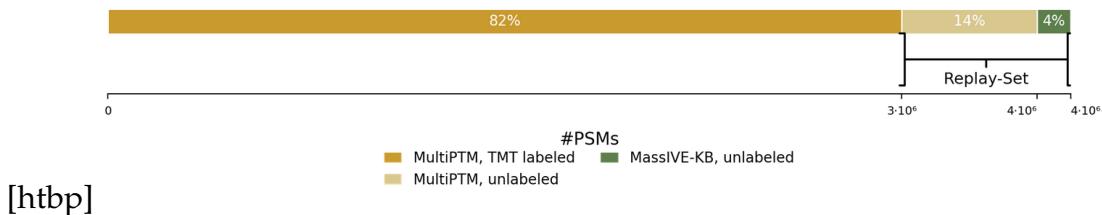


Figure 3.2: Composition of the fine-tuning dataset. The training dataset is dominated by TMT-labeled multi-PTM peptide-spectrum matches (82%), supplemented by a 18% Replay Set to preserve model robustness and recall on unlabeled spectra.

#### 3.2.2 TMT-labeled Data

The primary dataset for adapting the model to isobaric labeling is derived from the PROSPECT-MultiPTM collection [Zeng2024]. These data are based on the ProteomeTools project, a large-scale synthetic peptide library effort [Zolg2017].

**Instrumentation and Fragmentation** All spectra were acquired using Thermo Scientific Orbitrap instruments (Q Exactive and Orbitrap Fusion series). Fragmentation was performed exclusively using Higher-energy Collisional Dissociation (HCD), resulting in high-resolution MS2 spectra.

**Labeling and Modifications** The peptides in this dataset are labeled with Tandem Mass Tags (TMT). The chemical modification manifests as a specific mass shift at the peptide N-terminus and on the  $\epsilon$ -amino group of all Lysine (K) residues. The exact mass shifts follow the Unimod definitions and are encoded using the ProForma standard [Leis2022].

**Dataset Statistics** The TMT-labeled dataset is partitioned as follows:

- **Training Set:** 3,683,888 PSMs covering 75,268 unique peptides.
- **Validation Set:** 363,612 PSMs covering 7,751 unique peptides.

- **Test Set:** 364,867 PSMs covering 9,415 unique peptides.

### 3.2.3 Non-TMT Data (Replay Set)

To prevent catastrophic forgetting during the fine-tuning process, a diverse reference dataset of unlabeled (non-TMT) spectra is included. This “Replay Set” consists of a mixture of 80% MultiPTM data and 20% data from the MassIVE Knowledge Base (MassIVE-KB) [Wang2018].

**Dataset Statistics** The non-TMT reference data is partitioned as follows:

- **Training Set:** 784,128 PSMs covering 289,568 unique peptides.
- **Validation Set:** 98,396 PSMs covering 23,004 unique peptides.
- **Test Set:** 93,453 PSMs covering 19,141 unique peptides.

### 3.2.4 Quality Filtering and Pre-processing

To minimize noise and prevent the model from learning experimental artifacts, several quality filtering steps were applied:

- **Peak Cleaning:** Spectra containing no intensity information or empty peaks were removed.
- **Bias Mitigation:** To prevent overfitting to hyper-abundant peptides, a threshold of 229 PSMs per unique peptide sequence was enforced.
- **Data Leakage Prevention:** Following the *Modanovo* protocol, a strict data split was implemented. Peptides with specific modifications (e.g., *PEP[ph]*) were assigned to the same split as their unmodified counterparts (*PEP*) to ensure the model learns chemical principles rather than memorizing sequences [KlaprothAndrade2025].

The final dataset was structured into an 80/10/10 split (training, validation, and testing). To specifically address the TMT expansion, an 80/20 balance between TMT-labeled and unlabeled spectra was maintained, and all non-TMT spectra originating from TMT-specific experiments were removed to ensure label consistency. Furthermore, Unimod syntax was translated into mass-shift syntax (e.g., [+229.163]) to align with the model’s vocabulary.

### 3.3 Training Protocol

Model fine-tuning was initialized from the publicly available Modanovo checkpoint, allowing the model to build upon previously learned representations of peptide fragmentation and spectral structure [KlaprothAndrade2025]. In contrast to partial adaptation strategies, all model parameters were updated during fine-tuning, i.e. no layers were frozen, enabling global adaptation to TMT-specific fragmentation effects and modification patterns.

The underlying transformer architecture was kept identical to the base Modanovo configuration, comprising a model dimension of  $d_{\text{model}} = 512$ , 8 self-attention heads, a feed-forward dimension of 1024, and 9 layers each in the encoder and decoder stacks. This architectural consistency ensures that any observed performance differences can be attributed to the fine-tuning procedure rather than structural changes.

Optimization hyperparameters were deliberately chosen to favor stable adaptation of the pre-trained weights. A low learning rate of  $1 \times 10^{-6}$  was used to prevent catastrophic forgetting while still permitting gradual adjustment to TMT-induced shifts in fragmentation behavior. To further stabilize early training dynamics, a warm-up phase of two epochs was applied. Regularization was introduced via a weight decay of  $1 \times 10^{-5}$  and label smoothing with a factor of 0.01, improving generalization in the presence of heterogeneous modification patterns.

Training was performed using mixed-precision arithmetic with *bf16* precision, reducing memory consumption and improving computational efficiency on modern GPU architectures without compromising numerical stability [Micikevicius2017]. Model selection was based on validation loss, and the checkpoint with the lowest validation loss was retained for all downstream analyses.

### 3.4 Evaluation Strategy and Performance Metrics

To rigorously assess the performance of the TMT-adapted de novo sequencing model, a multi-faceted evaluation framework was established. The primary objective is to determine how well the model generalizes to TMT-labeled spectra and various post-translational modifications (PTMs) compared to traditional database-driven assignments.

### 3.4.1 Confidence Scoring and Peptide Ranking

Each peptide-spectrum match (PSM) generated by the model is assigned a confidence score to facilitate ranking and quality control. Following the architecture of Transformer-based models like Casanova and Modanovo, we derive a peptide-level score by calculating the arithmetic mean of the individual amino acid confidence scores, which are obtained from the softmax output at each decoding step [Yilmaz2022].

To ensure the plausibility of the predictions, a mass-matching constraint is applied. If the calculated mass of the predicted sequence (including PTMs and TMT labels) deviates from the observed precursor mass beyond a defined tolerance (e.g., 10 ppm), the peptide score is penalized. This integration of spectral evidence and thermodynamic constraints is crucial for distinguishing between high-confidence sequences and plausible but incorrect mass-shift combinations.

### 3.4.2 Precision-Coverage Analysis and Stratification

The core metric for evaluating the model's predictive power is the precision-coverage curve. This allows for a threshold-independent assessment of how many peptides can be identified at a given reliability level. For this study, we specifically focus on the Area Under the Precision-Coverage Curve (AUPCC), calculated using the trapezoidal rule [Pedregosa2011].

A critical aspect of our evaluation is the \*\*stratified analysis\*\*. To understand the specific impact of the TMT expansion, we evaluate the performance separately for:

- **TMT-labeled spectra:** To measure the success of the model adaptation to systematic mass shifts.
- **Unlabeled (non-TMT) spectra:** To ensure that the model retains its general sequencing capabilities without losing performance on standard data (preventing “catastrophic forgetting”).

Precision ( $P$ ) and Coverage ( $C$ ) at a score threshold  $t$  are defined as:

$$P(t) = \frac{|\text{Correct PSMs with score } \geq t|}{|\text{Total predictions with score } \geq t|} \quad (3.1)$$

$$C(t) = \frac{|\text{Predictions with score } \geq t|}{|\text{Total ground truth identifications}|} \quad (3.2)$$

A PSM is considered correct if the sequence exactly matches the ground truth identified by a database search (e.g., MaxQuant or MSFragger), treating isobaric

### 3 Methods

---

amino acids such as Leucine, Isoleucine and PyroGlu-Q, PyroGlu-E as equivalent [KlaprothAndrade2025]. We evaluate the precision for specific PTM-amino acid combinations. For a given modification (e.g., Phosphorylation at T), we subset the ground truth data to include all peptides containing this specific shift. This granular view ensures that the model’s ability to handle complex, multiplexed PTM patterns is validated across both TMT and non-TMT backgrounds.

The evaluation follows a peptide-centric approach: a single ground truth peptide contributes to the curves of all modifications it contains. For instance, a peptide sequence such as “[+229.997]-PEPT[+79.966]IDEK[+14.016]” is included in both the precision-coverage curve for phosphorylated threonine (T[+79.966]) and monomethylated lysine (K[+14.016]). In all subsequent analyses, the performance on unmodified peptides (light grey in Figure 4.1) serves as a reference baseline.

## 3.5 Application to Clinical Glioma Data

To evaluate the practical utility and robustness of the TMT-adapted model in a real-world clinical context, we applied it to a large-scale, independent glioma dataset. Unlike the synthetic and curated libraries used for fine-tuning, this dataset represents the inherent complexity of clinical proteomics, characterized by high dynamic range and heterogeneous post-translational modifications.

### 3.5.1 Experimental Dataset: The ClinSpect-M Glioma Cohort

The model was used on an unpublished clinical dataset provided by the Kusterlab [reference tbd]. This cohort comprises approximately 300 Glioma patient samples, representing a diverse spectrum of tumor grades and molecular subtypes like Astrocytoma, Oligodendroglioma, and Glioblastoma.

### 3.5.2 Data Acquisition and Large-Scale Preprocessing

The phospho-enriched dataset consists of approximately 6.5 million tandem mass spectra, originally acquired as Orbitrap-based raw files (.raw). To ensure compatibility with the model input, raw files were converted into Mascot Generic Format (MGF) using the ThermoRawFileParser (v2.0.0) [Hulstaert2020].

During preprocessing, all MS1 precursor scans and MS3 reporter ion scans were excluded from the sequencing workflow. The focus was restricted to MS2 fragment spectra, which contain the peptide backbone information necessary for sequence reconstruction.

### 3.5.3 Peptide Identification Pipeline

To achieve a comprehensive analysis of the Glioma proteome, we employed a dual-strategy approach: *de novo* sequencing for discovery and a database-driven search as a validation baseline.

**De Novo Sequencing Configuration** The sequencing of the TMT-labeled spectra was performed using the adapted Modanovo framework. To ensure high-quality sequence predictions and to accommodate the systematic shifts introduced by TMT, the following parameters were applied:

- **Mass Tolerances:** The precursor mass tolerance was set to 50 ppm to account for potential drift in large-scale clinical datasets. The isotope error range was restricted to [0, 3].
- **Sequence Constraints:** A minimum peptide length of 6 amino acids and a maximum of 100 were enforced. For the decoding process, a beam search with a width of  $n\_beams = 1$  was utilized, focusing on the top-ranked match to maximize throughput.

**Database Search Configuration (MaxQuant Baseline)** To provide a complementary ground-truth baseline and validate the *de novo* results, all Glioma datasets were processed using MaxQuant (version 2.1.3.0) [Tyanova2016]. This step was done by our collaborators and we obtained the results from them. This step is crucial to determine the “searchable” fraction of the proteome and to identify which spectra remain “unexplained” by traditional methods, thereby highlighting the discovery potential of our model beyond database-driven searches.

The search was conducted against the human reference proteome (UniProt UP000005640) with parameters harmonized to the experimental design:

- **Protease:** Trypsin/P was specified, allowing for cleavage C-terminal to Lysine and Arginine, even when followed by Proline.
- **Fixed Modifications:** Carbamidomethylation of cysteine (+57.021 Da) and TMT labeling of Lysine and the peptide N-terminus (+229.163 Da) were set as static modifications.
- **Variable Modifications:** Oxidation (M), acetylation (n-Term) and phosphorylation (S/T/Y) were included in the search space.

#### 3.5.4 Peptide Alignment and Sequence Validation

To validate the biological origin of the predicted *de novo* sequences and to distinguish between known peptides and potential novel discoveries, a sequence alignment against the reference proteome was performed. This step is crucial because *de novo* models predict sequences solely based on spectral features, which may include errors or biologically plausible variations not present in the reference database.

**Mass-Spectrometric Ambiguities and Encoding Strategy** A fundamental challenge in *de novo* peptide sequencing is the existence of isobaric or near-isobaric amino acid residues.

Since deep learning models primarily operate on mass-to-charge ratios ( $m/z$ ) and the resulting mass differences, they essentially identify mass shifts rather than unique chemical identities [Muth2012]. While transformer-based architectures can theoretically also learn to consider aminoacid context when mass shifts overlap, this still remains unevaluated and therefore we resolve the ambiguities.

We identified several critical ambiguities:

- **I/L Equivalence:** Leucine (L) and Isoleucine (I) are structural isomers with an identical monoisotopic mass of 113.08406 Da, making them indistinguishable in HCD spectra.
- **Deamidation-induced Ambiguities:** The deamidation of Glutamine (Q) and Asparagine (N) results in a mass increase of +0.984016 Da. This shift leads to these pairs:
  - Glutamic acid (E, 129.042593 Da) and deamidated Glutamine (Q[+0.98], 129.042594 Da).
  - Aspartic acid (D, 115.026943 Da) and deamidated Asparagine (N[+0.98], 115.026943 Da).
- **Pyro-glutamate Formations:** Ambiguities also occur between Pyro-glu E and Pyro-glu Q. However, as these appeared in less than 1% of the detected spectra in our dataset, they were not explicitly encoded to avoid over-complicating the search space.

**BLASTp Configuration and Ambiguity Handling** Alignment was executed using Protein-Protein BLAST (version 2.17.0+) against the human reference proteome (UniProt UP000005640). Because the reference proteome consists of unmodified sequences, a direct match of a deamidated peptide (predicted as

D or E by the model) against a genomic N or Q would fail under standard parameters.

To resolve this, we implemented a specialized encoding strategy using IUPAC ambiguity codes:

- All predicted **D** residues (which could be  $N[+0.98]$ ) were replaced with **B** (asparagine or aspartic acid).
- All predicted **E** residues (which could be  $Q[+0.98]$ ) were replaced with **Z** (glutamine or glutamic acid).

An alternative approach would have been the implementation of a custom substitution matrix. However, the BLASTp does not support user-defined matrices for short peptide searches without extensive modification of the source code. Therefore, we utilized the PAM30 matrix, which is optimized for short sequences and natively supports the B and Z codes.

#### Example of IUPAC Encoding for Deamidation Alignment:

<b>Reference Proteome:</b>	A V G <b>D</b> L T S <b>Q</b> R	
<b>De-novo Prediction:</b>	A V G <b>N</b> <b>L</b> T S <b>E</b> R	(Potential Mismatches)
<b>Standard BLASTp:</b>	A V G <b>-</b> <b>L</b> T S <b>-</b> R	<b>Mismatch</b>
<b>Our Strategy (PAM30):</b>	A V G <b>B</b> <b>L</b> T S <b>Z</b> R	<b>Perfect Match*</b>

\*Note: Both sequences are converted to IUPAC codes B (N/D) and Z (Q/E) prior to alignment, allowing the PAM30 matrix to score them as identical residues.

**Alignment Parameters** The alignment parameters were adjusted to account for the short nature of tryptic peptides. We set an E-value threshold of 2000, as the small search space of a single peptide often results in high E-values despite perfect sequence identity. Furthermore, we applied a query coverage threshold (`qcov_hsp_perc`) of 80%. This constraint ensures that at least 80% of the *de novo* predicted sequence aligns with the database entry, thereby filtering out spurious, short-match alignments.

Another adjustment was the deactivation of composition-based statistics (`comp_based_stats`). In default settings, many alignment tools (such as BLAST) adjust the scoring matrices based on the amino acid composition of the sequences being compared to account for biased distributions [Altschul1997]. However, in the context of *de novo* sequencing, where sequences are often short, these statistics can lead to score inflation or, conversely, penalize short but biologically valid matches [Kersey2004].

### 3 Methods

---

**Post-Alignment Filtering and SNP Analysis** The resulting alignments were further enriched to identify Single Nucleotide Polymorphisms (SNPs) and truncation events. To account for sequences extending beyond protein termini or alignment gaps, the full query and target sequences were retrieved. The number of mismatches was calculated by considering the B and Z equivalences. For each remaining mismatch, an automated codon-lookup was performed. A mismatch was classified as “SNP-explainable” if the transition between the predicted amino acid and the reference residue could be achieved by a single nucleotide substitution in the underlying codon.

Cases involving gaps or truncations where the query extended beyond the reference boundaries were excluded from the SNP analysis to maintain high confidence in the mutation mapping.

#### 3.5.5 Genomic Validation and Integration of Panel Sequencing Data

To substantiate the biological relevance of the *de novo* predicted sequences, particularly those harboring potential amino acid substitutions, we integrated patient-specific genomic evidence. For the ClinSpect-M cohort, genomic panel sequencing data (covering approximately 500 cancer-relevant genes) was available.

**Processing of Genomic Variants** The genomic data, provided in annotated JSON format, was processed using the Nirvana clinical variant annotator (v3.2.3) based on the GRCh37 (hg19) genome assembly. To identify potential tumor-specific peptides and single amino acid variants (SAAVs), somatic mutations were filtered using the Variant Effect Predictor (VEP) [McLaren2016].

**Mapping and Evidence Filtering** For each *de novo* peptide that aligned to a protein originating from the 500-gene panel with only a few mismatches, we performed a precise positional mapping. The Uniprot identifiers from the proteomic alignment were mapped to Ensembl transcript IDs used in the VEP output. A peptide candidate was considered “genomically validated” if:

1. The predicted mismatch relative to the UniProt reference sequence occurred at the exact position of a genomic variant identified in the panel sequencing.
2. The amino acid substitution predicted by the *de novo* model matched the transcript-level prediction (HGVSp) from the genomic data.

This integrated workflow allows for the identification of non-canonical peptides that are absent from standard reference databases but supported by the patient's individual mutational landscape.

## 3.6 Identification and Validation of Phosphorylation Sites

Beyond sequence variations, the precise localization of post-translational modifications (PTMs), particularly phosphorylations, is critical for understanding cellular signaling. We implemented a standardized workflow to map and validate predicted modification sites.

The localization of phosphorylation sites (p-sites) was performed by integrating *de novo* predictions from Modanovo-TMT and traditional database-driven results from MaxQuant.

For each predicted phosphopeptide, the unmodified sequence was first aligned to the reference proteome as described in the previous section. By utilizing the precise start and end coordinates of the alignment, we translated the local modification index (the position within the peptide) into a global protein position. This allows for a direct comparison of modification sites across different peptides and experiments, even if the underlying peptide sequences differ due to alternative cleavage or truncations.

To ensure the reliability of the identified p-sites, we had a look at the evidence supporting each site taking into account the following factors:

- **Site Redundancy:** Number of supporting Peptide-Spectrum Matches (PSMs) and overlapping sequences from alternative cleavage events.
- **Quality Metrics:** *De novo* confidence scores for the PSMs and the alignment query coverage.
- **Cross-Platform Consistency:** Identification of “consensus sites” predicted by both MaxQuant and *ModanovoTMT*.
- **Literature Comparison:** Cross-referencing mapped positions with previous studies and databases like PhosphoSitePlus [Hornbeck2015].

## 3.7 Spectral Evidence and Intensity Prediction

While sequence alignment and genomic mapping provide biological context, the physical validity of a *de novo* prediction must be confirmed by comparing

### 3 Methods

---

the experimental MS/MS spectrum with the expected fragmentation pattern of the predicted sequence.

#### 3.7.1 Spectral Angle (SA) as a Quality Metric

To quantify the similarity between the experimental spectrum and a theoretical prediction, we utilize the Spectral Angle (*SA*). Unlike the traditional Pearson correlation, the *SA* is less sensitive to high-intensity peaks and provides a more robust measure of relative fragment ion intensities [Toprak2014]. The *SA* is defined as:

$$SA = 1 - \frac{2 \cdot \arccos\left(\frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \cdot \|\mathbf{v}\|}\right)}{\pi} \quad (3.3)$$

where  $\mathbf{u}$  represents the vector of intensities from the experimental spectrum and  $\mathbf{v}$  the predicted intensities. A value of 1 indicates a perfect match, while 0 indicates no similarity.

#### 3.7.2 Intensity Prediction via Koina and Prosit

Since TMT labeling and various PTMs significantly alter fragmentation energy and peak intensities, we utilized the `Prosit_2024_intensity_PTMs_g1` model, accessed via the Koina federated prediction service [Gessulat2019, Wilhelm2024]. This deep learning model is specifically trained to handle TMT-labeled peptides and a wide array of PTMs.

**Implementation Workflow** For each high-scoring *de novo* peptide, the predicted sequence (including modifications) was converted into the standardized Unimod format. These strings were sent to the Koina API along with the experimental parameters, specifically setting the Collision Energy (CE) to 32 to match the Orbitrap acquisition settings. The resulting theoretical intensities for b- and y-ions were then used to calculate the spectral angle *SA*. This step ensures that our discoveries are not only biologically plausible but also physically consistent with the raw mass spectrometric data.

# 4 Results

Starting from the *Modanovo* checkpoint and expanding the architecture, the primary objective was to bridge the gap between specialized PTM identification and large-scale quantitative TMT workflows. We first characterize the underlying dataset and modification coverage before analyzing the model’s sequencing accuracy through precision-coverage metrics.

## 4.1 A dataset for robust expansion: TMT-labeled spectra with diverse PTMs

The performance of deep learning models in *de novo* sequencing is fundamentally tied to the diversity and quality of the training data. Our fine-tuning dataset was strategically compiled to represent both the “classical” PTM landscape and the specific challenges of TMT-multiplexed proteomics. The final dataset consists of a curated mixture of TMT-labeled spectra and a “Replay Set” of non-TMT spectra, maintaining an approximate 80:20 ratio to ensure the model retains its ability to generalize across different experimental conditions.

To demonstrate the PTM-coverage of the training data, we visualized the modification density across different residues.

## 4 Results

---

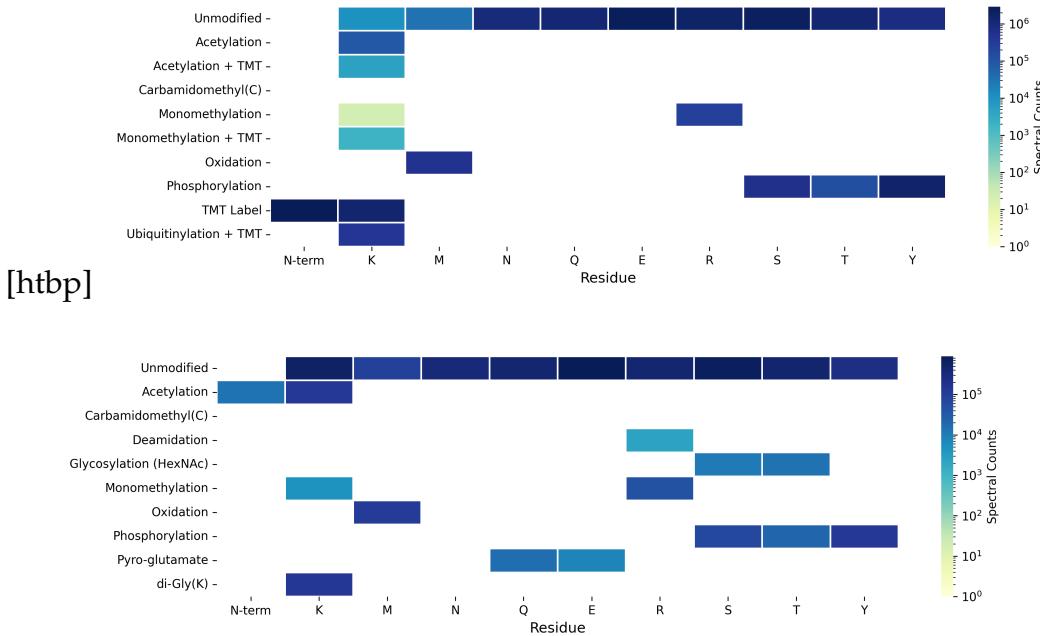


Figure 4.1: Distribution of identified post-translational modifications (PTMs) across specific amino acid residues. The heatmaps display log-scaled spectral counts for various modifications as a function of the modified residue. (a) shows the TMT-labeled dataset (b) represents the Replay Set.

As shown in Figure 3.3, the compiled data set provides high coverage for common PTMs such as Oxidation (M) and Phosphorylation (S, T, Y) we tried to keep the TMT and non TMT PTM landscape similar to not have biases in it. However like in Modanovo, there remains some optimization potential. Specifically, certain rare PTM-residue combinations or complex multiplexed modifications (e.g., Ubiquitylation + TMT) exhibit lower spectral counts. This sparsity is a known challenge in training robust *de novo* models, caused by the limited number of spectra with data base search identification.

## 4.2 Modanovo-TMT confidently identifies peptides in multiplexed proteomics data

The model was initialized with weights from the Modanovo base, allowing it to leverage pre-learned features of peptide fragmentation. Training was conducted until the validation loss showed early signs of increase, indicating the onset of overfitting. Convergence was reached after 13 epochs, corresponding to approximately 1,550,000 training steps. This stability suggests that the model

## 4.2 Modanovo-TMT confidently identifies peptides in multiplexed proteomics data

successfully integrated the TMT-specific fragmentation patterns and expanded vocabulary.

To evaluate the predictive confidence, precision-coverage curves were generated for different PTM categories, stratified by TMT vs. non-TMT data (Figure 4.1). The performance is quantified using the Area Under the Precision-Coverage Curve (AUPCC), providing a robust measure of sequencing quality across varying confidence thresholds.

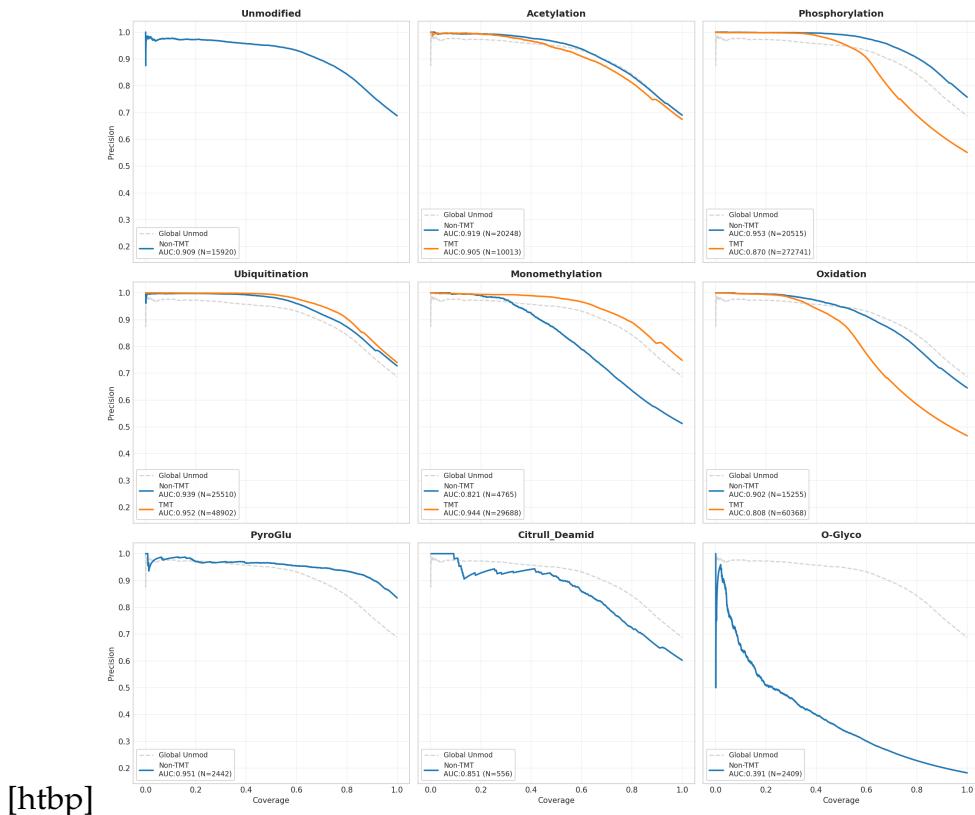


Figure 4.2: Precision-coverage curves at the peptide level across different PTMs. PTM types are shown in the different panels. Colors represent the performance of the final model on TMT-labeled (orange) and non-TMT (blue) spectra. Unmodified non-TMT peptides are shown in light grey for comparison.

A key requirement for the expanded model was to maintain high performance on standard (non-TMT) data. On non-TMT unmodified peptides, the model achieved an AUPCC of 0.92, which is highly consistent with the performance reported for the original Modanovo architecture (AUPCC of 0.93) [Yilmaz2022, KlaprothAndrade2025]. This demonstrates that the inclusion of TMT-related features does not degrade general sequencing accuracy.

## 4 Results

---

Specific PTM performance highlights the model’s strengths and limitations:

- **High-Performance PTMs:** The model demonstrated excellent accuracy for Phosphorylation (AUPCC 0.95) and Pyro-Glu (AUPCC 0.95). The high precision for Pyro-Glu is likely due to its restricted occurrence at the peptide N-terminus, a positional pattern that transformer architectures can effectively internalize [Yilmaz2022].
- **Challenging Modifications:** In contrast, O-Glycosylation (S/T[+203.079]) remains a significant challenge with an AUPCC of 0.39. This mirrors difficulties reported in literature, attributed to both the scarcity of training examples and the complex, often labile fragmentation patterns of glycopeptides which deviate from standard b- and y-ion series.

The global performance on TMT data (AUPCC 0.89) is slightly lower than on non-TMT data (AUPCC 0.92). Besides the lack of data it could also be explained by increased spectral complexity due to the mass shifts and altered fragmentation patterns introduced by TMT labeling. However, the model’s ability to maintain high precision across a wide range of PTMs in TMT-labeled spectra confirms that it has successfully learned to interpret the chemical signatures of multiplexed proteomics data. Specific findings for TMT-labeled PTMs include.

- **Ubiquitination and Monomethylation:** Interestingly, the model performed slightly better on TMT-labeled Ubiquitination (AUPCC 0.95) compared to its non-TMT counterpart (AUPCC 0.94). For Monomethylation, the gap was even more pronounced (TMT: 0.94 vs. non-TMT: 0.82), suggesting that the systematic mass shifts provided by TMT might aid in distinguishing these specific modifications in certain contexts.
- **Acetylation:** Performance remained robust across both categories (TMT: 0.91 vs. non-TMT: 0.92), indicating that Acetylation is identified with high confidence regardless of the labeling strategy.

The final precision at full coverage (e.g., 0.59 on all tmt-labeled spectra) demonstrates that the model provides a reliable foundation for uncovering biological insights in multiplexed quantitative proteomics experiments.

### 4.3 Isotopic Sensitivity and Beam Search in TMT Data

A critical challenge identified during the evaluation of TMT-labeled data was the increased frequency of non-monoisotopic precursor selection. In TMT multiplexed samples, the high density of ions and the chemical labeling itself often

lead to a shift where the instrument triggers fragmentation on an isotopic peak rather than the monoisotopic mass. To address this, the model configuration was iteratively optimized.

The isotope error range serves as a corrective buffer during inference for misassigned precursor masses — a frequent artifact where the mass spectrometer selects a  $^{13}C$  isotope peak instead of the monoisotopic mass. The initial baseline configuration (Config0) utilized a restrictive isotope error range of 0, 1, which often forced the model to strictly adhere to the theoretical monoisotopic mass, leading to sequence errors when the precursor mass was incorrectly assigned by the instrument. By expanding this range to 0, 3 (Config ISO), the model gained the necessary flexibility to account for these systematic mass shifts without being penalized for small mass deviation errors.

Furthermore, while literature for models like Casanova suggests that increasing the beam size has diminishing returns, our results indicate a strong synergy between isotope flexibility and search breadth. Increasing the beam size from 1 to 5 (Config FINAL) allowed the decoder to explore a wider sequence space, which proved essential for resolving the complex fragmentation patterns of TMT-labeled and modified peptides.

The optimization process led to a stepwise improvement in all global metrics (see Table 4.1). The global AUPCC increased from 0.8746 (Config0) to 0.8817 (ISO) and reached its peak at 0.8988 in the FINAL configuration. Notably, the most significant breakthrough occurred in the jump to the FINAL configuration, where the global precision at full coverage rose from 0.5726 to 0.6146.

Table 4.1: Comparison of model performance across different configurations.

Metric	Config0	Config ISO	Config FINAL
Global AUPCC	0.8746	0.8817	0.8988
Global Precision (Cov 1.0)	0.5726	0.5726	0.6146

The transition to Config FINAL particularly benefited complex PTMs:

- **Ubiquitination:** This category showed a precision jump from 0.6675 to 0.7359. The combination of expanded isotope ranges and increased beam size likely allowed the model to better identify the characteristic GlyGly-remnant fragmentation patterns, which are often obscured in TMT-labeled spectra.
- **Phosphorylation:** As the largest dataset ( $N=293,256$ ), the shift from 0.8556 (Config0) to 0.8784 (AUPCC) and a precision increase to 0.5656 is statistically the most significant indicator of the model's enhanced robustness.

- **Monomethylation:** The AUPCC rose from 0.9032 to 0.9326. Remarkably, TMT-labeled methylated peptides achieved a significantly higher AUPCC (0.9443) than their non-TMT counterparts (0.8211). This suggests that the model learned to leverage the TMT-induced mass shifts as a “fingerprint” to distinguish methylation from isobaric interferences more effectively than in unlabeled data.

The improvement in Global TMT AUPCC (from 0.8640 to 0.8717 in the ISO stage) confirms that accounting for isotopic uncertainty is a prerequisite for high-confidence *de novo* sequencing in multiplexed workflows.

## 4.4 Application on independent Glioma TMT Dataset

To evaluate the model’s discovery potential, inference was performed on the complete Glioma TMT dataset, comprising 6.48 million spectra. Given that *de novo* sequencing operates independently of protein databases, it is crucial to validate the predicted sequences against a reference proteome to distinguish between high-confidence identifications and potential false positives.

### 4.4.1 Proteome Alignment and Score Calibration

The predicted sequences for unmodified peptides were aligned against the human reference proteome using `blastp`. This alignment process serves as a first-tier validation of the model’s output quality. The sequences were categorized based on their alignment criteria: perfect matches (identity = 100%), sequences with one mismatch, and sequences with two mismatches.

Out of the total predictions, 1.34 million peptides aligned perfectly with the reference proteome, while 0.62 million and 0.8 million sequences exhibited one and two mismatches, respectively.

#### 4.4 Application on independent Glioma TMT Dataset

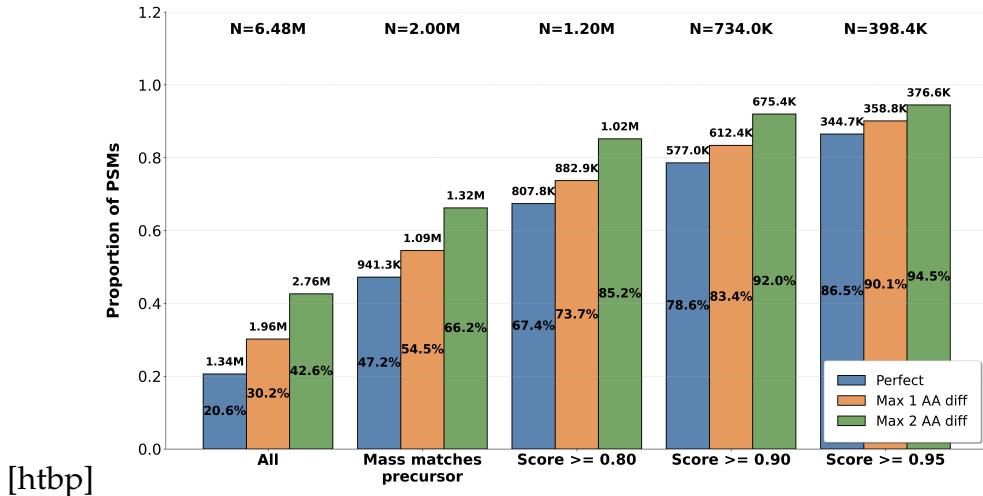


Figure 4.3: Cumulative alignment of *de novo* predictions to the human reference proteome across confidence score thresholds. The plot illustrates the proportion of sequences matching the reference with zero (blue), up to one (orange), and up to two (green) amino acid substitutions. The segments between the curves represent peptides with specific mismatch counts, potentially accounting for Single Nucleotide Polymorphisms (SNPs) or technical sequencing errors. The increasing trend demonstrates the correlation between the model’s self-reported confidence and sequence accuracy.

The upward trend of the bars in Figure 4.2 confirms that the internal scoring mechanism of the Transformer architecture effectively reflects the probability of a sequence being biologically “correct.” At the highest confidence bin ( $> 0.95$ ), the vast majority of sequences are plausible in sense that they appear in the proteome, providing a solid baseline for the subsequent analysis of modified and novel peptides. In addition to proteome alignment, the plausibility of the predictions was strengthened by comparing the theoretical mass of the predicted sequences with the observed precursor  $m/z$ . All the peptides starting from the second bin fulfill this requirement.

#### 4.4.2 Stratified Performance and Modification Stability

To further investigate the reliability across different chemical states, the precision of the predictions was stratified by modification type and ranked by confidence (PSM Rank). In this analysis, the “Rank 1” prediction represent the highest-confidence spectrum.

## 4 Results

---

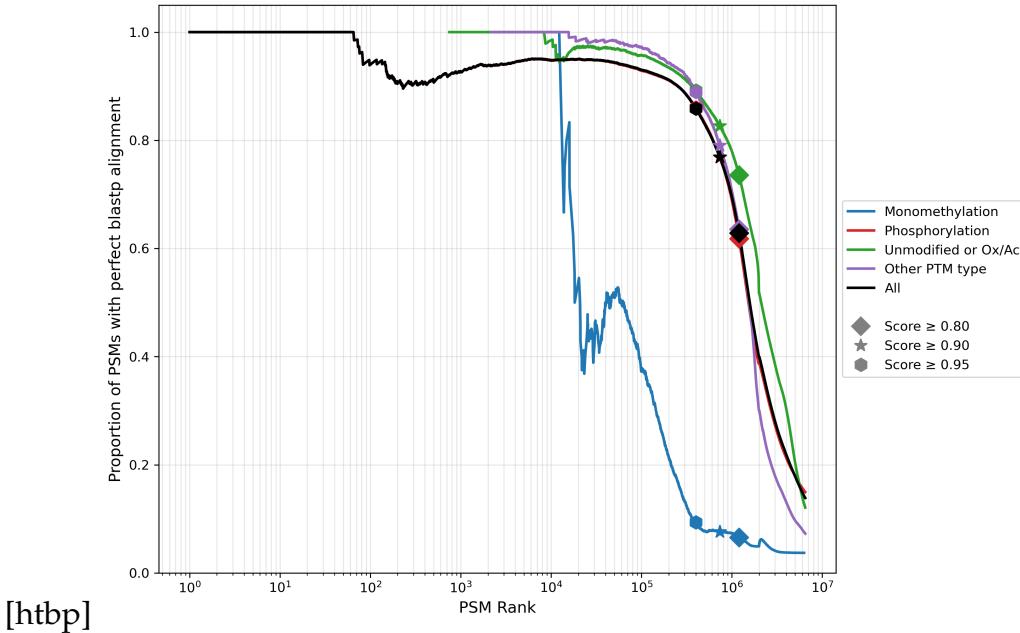


Figure 4.4: Stratified precision-rank analysis of the adapted model. The plot shows the proportion of Peptide-Spectrum Matches (PSMs) achieving a perfect *blastp* alignment to the human reference proteome, ranked by descending model confidence scores. Curves are stratified by modification type: unmodified or just oxidation/acetylation(green), phosphorylation (red), monomethylation (blue), and other PTMs (purple). Symbols represent specific confidence score thresholds ( $\geq 0.80, 0.90, 0.95$ ). While performance remains high (> 90%) for most classes at high confidence, monomethylated peptides show a significant precision drop, indicating potential challenges in mass-equivalent deconvolution or model overconfidence for this specific modification.

As shown in Figure 4.4, most modifications follow the expected trend where decreasing confidence ranks correlate with a lower proportion of perfect alignments. Phosphorylated peptides follow the global trend closely.

A notable outlier is Monomethylation. While this modification showed excellent performance in the AUPCC metrics on the test set, the proportion of perfect alignments drops more sharply with increasing rank compared to other modifications. This discrepancy suggests the absolute mass shifts or the localization of the methyl group might still lead to mismatches during genomic alignment in real-world samples.

The alignment results demonstrate that the expanded model produces biologically plausible peptide sequences. The high correlation between the model's confidence score and the genomic match rate validates the use of these scores

as a filter for discovery. This foundation allows for the exploration of spectra that remain unidentified by traditional database-driven methods, such as MaxQuant, which will be discussed in the following section.

## 4.5 Complementing the state of the art:MaxQuant

A primary objective of this study was to evaluate the extent to which *de novo* peptide sequencing can expand the identification landscape beyond the limitations of database-driven methods. By comparing our results with those obtained via MaxQuant (MQ), we can quantify the “dark matter” of the proteome—spectra that contain high-quality information but remain unidentified by MaxQuant.

### 4.5.1 Unique Peptide Identifications

The comparison of unique peptide sequences identified by both methods reveals a substantial expansion of the detectable proteome. As shown in Figure 4.5, the *de novo* approach identified approximately 40,000 unique peptides that were also found by MaxQuant, demonstrating high consistency in the “known” space.

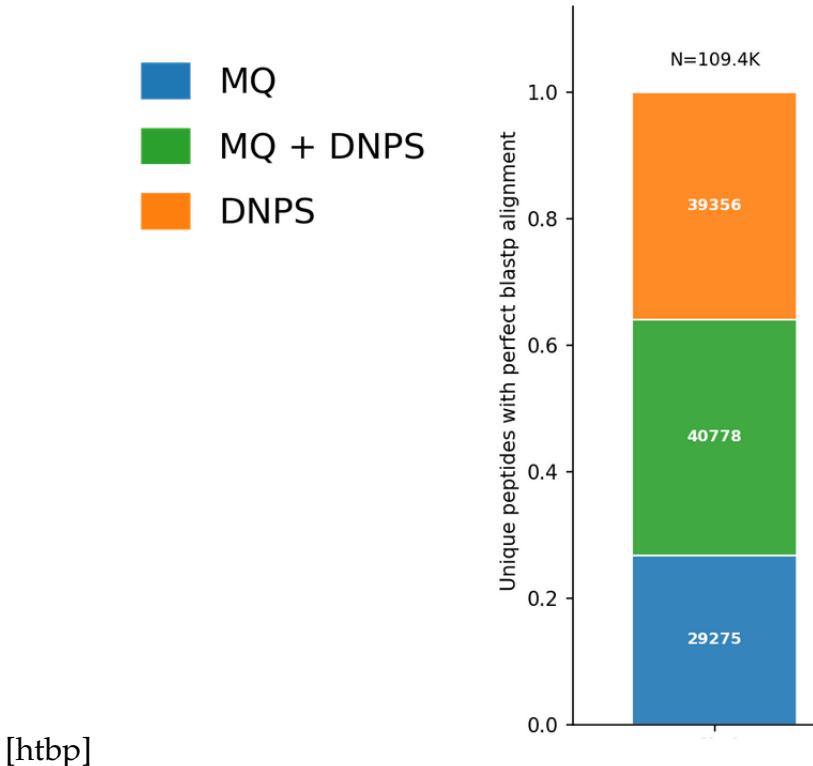


Figure 4.5: Overlap of unique peptide identifications between the adapted *de novo* model and MaxQuant. The model successfully recovers the majority of MQ identifications while contributing 40,000 additional unique sequences.

Modanovo-TMT demonstrated a high degree of overlap with the established workflow, successfully identifying 40,000 unique peptides that were also present in the MQ results. This substantial intersection further confirms the model's ability to reliably recover standard peptide sequences.

Beyond this shared proteomic space, MQ identified approximately 30,000 peptides that remained exclusive to the database search. In contrast, the *de novo* approach identified an additional 40,000 unique sequences that were not captured by the MQ search. These findings indicate that while both methods are complementary, the *de novo* model effectively doubles the number of unique peptide candidates by uncovering sequences that lie beyond the constraints of traditional database matching, while still maintaining high sensitivity for the core proteome.

### 4.5.2 Identification of Previously Unidentified Spectra

Beyond the unique peptide identifications we looked at the spectrum level (PSMs). In the analyzed dataset, MaxQuant provided identifications for 1,510,866 spectra. Our model was able to provide high-confidence sequences for a significant portion of the remaining “dark” spectra:

- **High-Confidence Matches (Score  $> 0.8$ ):** We identified 446,290 additional spectra with a confidence score above 0.8. Based on our previous alignment validation, this score range corresponds to highly reliable peptide sequences.
- **Precursor-Shifted Matches (Score  $-0.2 < s < 0$ ):** Interestingly, we found 231,208 spectra in a score range that indicates high sequencing confidence but carries a penalty ( $-1$ ) due to a precursor mass mismatch.

Summing these categories, the *de novo* approach provides interesting sequence candidates for over 670,000 spectra that were previously discarded in the MQ workflow. This substantial increase in spectral utilization must of course be further analyzed and validated with different approaches like aligning to a 6-frame translation.

### 4.5.3 Discovery of SNP-Related Variants

The identification of these additional sequences suggests that the “dark proteome” in these glioma samples is rich in biological variants. The high number of high-confidence predictions that do not perfectly match the database entries—especially those with slight precursor shifts—points towards the presence of non-canonical protein isoforms.

The discovery of over 670,000 previously unidentified spectra suggests that a significant portion of the “dark proteome” in these samples arises from sequences not represented in the canonical reference database or not passing the FDR threshold. To investigate whether these unique identifications stem from biological mutations, we analyzed peptides that showed minor sequence deviations (one or two amino acids) compared to the closest database entry. To ensure a conservative and robust analysis, we employed a “clean-hit” filtering strategy: any peptide that produced a perfect match in any experimental fraction or database search was excluded, focusing our analysis solely on truly novel sequence candidates.

By aligning these unique sequences back to the human proteome, we evaluated whether the observed amino acid substitutions could be explained by

## 4 Results

---

Single Nucleotide Polymorphisms (SNPs). A substitution was flagged as a potential SNP if the transition between the canonical and the *de novo* predicted amino acid could be achieved by a single nucleotide change within the corresponding codon [Wang2011]. Our analysis revealed that approximately 30% of all single amino acid deviations and 25% of dual amino acid deviations are directly explainable by such genomic point mutations. For instance, we identified cases where the model predicted a sequence such as *PAPTIT* instead of the canonical *PEDTIT*. Both mismatches in this example—Glutamate (E) to Alanine (A) and Aspartate (D) to Threonine (T)—are reachable via a single nucleotide exchange in their respective codons (e.g., GAG → GCG).

These findings demonstrate that the *de novo* model effectively captures protein-level manifestations of genetic variability that remain invisible to standard database-search workflows [Alfaro2014]. The high percentage of explainable substitutions confirms that these are not random sequencing errors, but likely reflect the actual biological diversity of the glioma samples. This capability to identify non-canonical isoforms and mutations without requiring matched genomic data highlights the transformative potential of deep learning-based sequencing for personalized proteogenomics [Choudhary2001].

The identification of these amino acid substitutions at the protein level, however, does not inherently guarantee the presence of a corresponding genomic variant. To validate our findings, we integrated the *de novo* results with genomic data from panel sequencing, covering 500 cancer driver genes. These genomic variants were annotated using the Ensembl Variant Effect Predictor (VEP) to identify potential Single Amino Acid Variants (SAAVs). By cross-referencing our data, we found that 694 identified SNPs were supported by direct genomic evidence, with the predicted SAAVs aligning with genomic mutations at high precision. In several instances, the evidence was further strengthened by multiple unique, overlapping peptides covering the same mutation site, providing independent proteomic confirmation for a single genomic event [Wang2014].

To ensure the spectral reliability of these novel identifications, we performed a validation based on spectral similarity. We compared the experimental spectra against theoretical fragment patterns using the spectral angle as a metric for similarity. This confirmed that the observed fragmentation matches the predicted pattern of the mutated sequence with high fidelity, whereas a comparison with the non-mutated, canonical sequence would result in significant mass shifts and poor spectral alignment, unless the substitution was isobaric (e.g., Leucine to Isoleucine) [Gessulat2019].

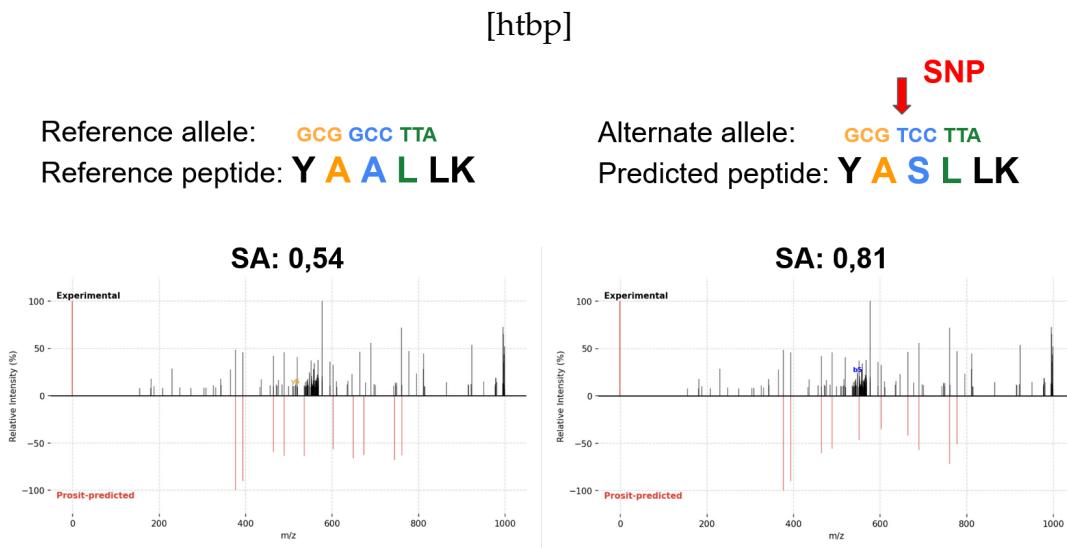


Figure 4.6: Left: Prosit-predicted spectrum (top) and experimental spectrum (bottom) for the reference peptide sequence YAALLK. The cartoon illustrates the relevant nucleotide sequence and fragment ion series assuming the reference genome allele. Right: Same as for left, but for the peptide sequence YAASLKD predicted by our model for the same experimental spectrum. The higher degree of overlap between experimental and theoretical spectrum for the mutated peptide (right) validate the models predictipn.

The true strength of this proteogenomic approach lies in the context of TMT multiplexing. While database-driven searches often struggle with the increased complexity and altered fragmentation of labeled peptides, TMT-based workflows allow for the simultaneous quantification of these variants across multiple samples. By linking the identified SAAVs with the specific MS3 reporter ion channels, it becomes possible to map a mutation directly to a specific patient within a multiplexed run [Pertosi2016]. This highlights the synergy of expanding *de novo* sequencing to TMT data: it not only uncovers mutations beyond the reach of standard databases but also preserves the quantitative resolution necessary for clinical and biological interpretation in large-scale cohorts.

#### 4.5.4 Discovery of PTM Sites

A key advantage of *de novo* peptide sequencing is the ability to identify post-translational modifications (PTMs) without the inherent bias of a restricted search space. To evaluate the model's capacity for PTM discovery, we systematically analyzed predicted phosphorylation sites (p-sites) that were not identified by the MaxQuant (MQ) search.

## 4 Results

The methodology for PTM mapping involved several steps: First, all *de novo* predicted peptides were aligned against the human reference proteome. We considered two distinct alignment scenarios: (1) *offby\_0*, representing perfect matches where the predicted sequence (including PTMs) matches the database entry exactly, and (2) *offby\_1\_snp*, allowing for a single mismatch to account for potential single nucleotide polymorphisms or sequence variants. Following alignment, the predicted modification residues were mapped to their specific positions within the protein sequence. To ensure a conservative estimate, we filtered the results to unique p-sites, aggregating redundant spectral identifications to a single site per protein.

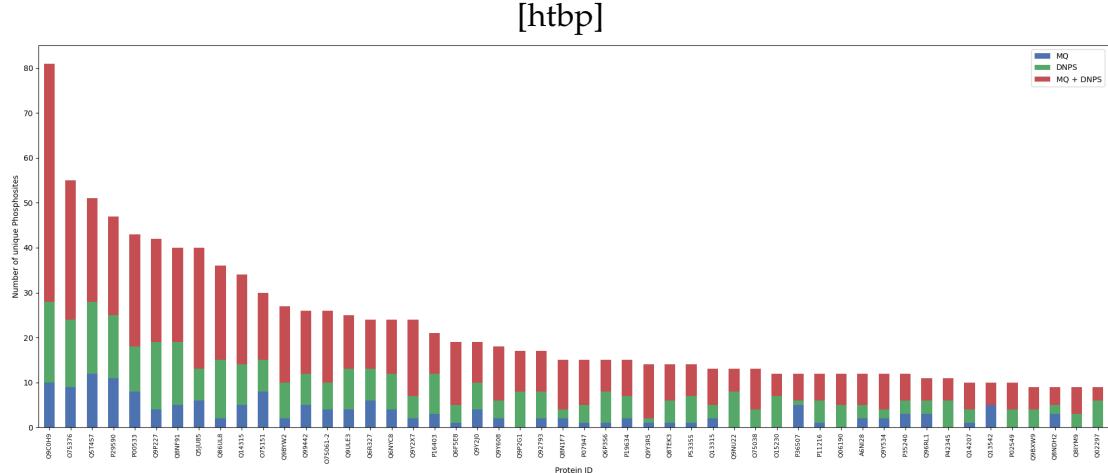


Figure 4.7: Distribution of unique phosphorylation sites identified by MaxQuant (MQ) and our model(DNPS) for selected cancer-related proteins. The results demonstrate a significant expansion of the detectable phospho-landscape.

The statistical evaluation highlights a massive expansion of the p-site landscape. For the *offby\_0* category, a total of 224,637 unique p-sites were identified across the dataset. While MaxQuant identified 40,598 sites, the *de novo* model (DNPS) contributed 103,438 sites, with an additional 180,526 sites overlapping or being newly discovered in total (see Figure 4.7). In the *offby\_1\_snp* category, the expansion is even more pronounced, with DNPS identifying 77,078 unique sites compared to only 4,977 by MQ. It should be noted, however, that while these numbers are promising, they likely contain a higher noise floor, as stringent false discovery rate (FDR) filtering for *de novo* PTMs is still an evolving area of research.

We took a quick look into the top 5 candidate cancer-related proteins frequently identified in our p-site pipeline:

- **SRCIN1 (Q9C0H9):** This protein acts as a negative regulator of SRC kinase by activating CSK, thereby inhibiting downstream signaling pathways involved in cell migration [[UniprotQ9C0H9](#)]. While databases like PhosphoSitePlus indicate potential regulation via phosphorylation, our model identified a high density of p-sites (e.g., over 80 sites in certain contexts), suggesting a complex regulatory “p-code” that warrants further biochemical validation.
- **NCOR1 (O75376):** A nuclear receptor corepressor that recruits histone deacetylases to mediate gene repression. Phosphorylation (e.g., via Akt) is known to regulate its dissociation from receptors like PPAR $\alpha$  [[NCOR1\\_PMC](#)]. Our discovery of additional sites suggests a more nuanced control of metabolic gene activation than previously documented.
- **UBR4 (Q5T4S7):** An E3 ubiquitin-protein ligase involved in the N-degron pathway. Our data shows numerous previously uncharacterized p-sites alongside known ubiquitination sites, likely reflecting its role in orchestrating complex stress responses and protein turnover [[UniprotQ5T4S7](#)].
- **PML (P29590):** Crucial for the formation of PML-nuclear bodies (PML-NBs), this protein is a central hub for tumor suppression. PTMs are known to regulate its scaffolding function and antiviral responses [[PML\\_PMC](#)]. The expanded p-site map provided by our model could clarify the dynamics of PML-NB assembly in cancer cells.
- **EGFR (P00533):** As a major therapeutic target in oncology, the phosphorylation of EGFR is well-studied, with over 100 known sites in specialized databases [[EGFR\\_PubChem](#)]. Our model successfully recovered known regulatory tyrosines (e.g., Y1173) while proposing novel threonine and serine sites that may contribute to signaling crosstalk.

Despite these findings, the high number of predicted sites (particularly the 80 sites on SRCIN1) suggests that the current *de novo* PTM output requires further structural filtering. The potential for false positives due to spectral noise or misassignment of mass shifts remains a challenge, necessitating more refined localized scoring in future iterations of the pipeline.

While the high density of predicted sites in proteins like SRCIN1 underscores the need for further structural filtering, our pipeline successfully identified several high-confidence regulatory markers with profound biological implications. The most compelling example is the detection of phosphorylation at **Serine 15 (p-Ser15)** on the liver isoform of glycogen phosphorylase, **PYGL** (P06737).

## 4 Results

---

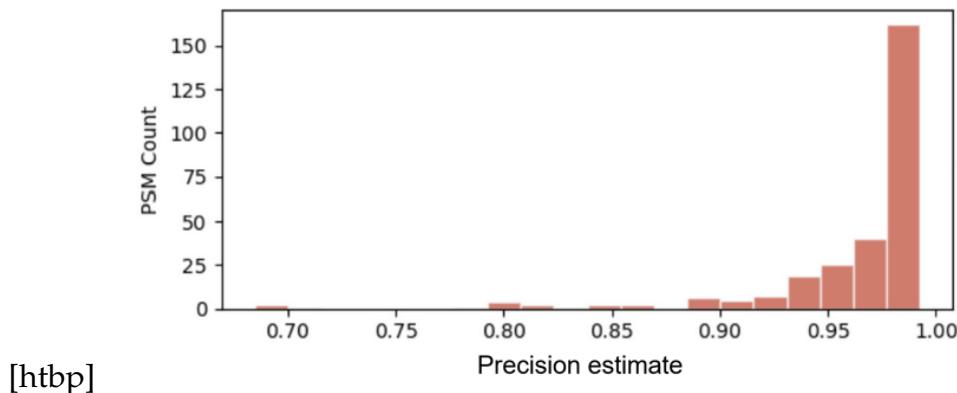


Figure 4.8: Confidence distribution for the predictions validating p-Ser15 on PYGL. The identification is supported by 279 PSMs, indicating high reproducibility.

The identification of p-Ser15 on PYGL is supported by substantial spectral evidence, comprising 279 Peptide-Spectrum Matches (PSMs) with a high mean spectral angle of 0.7 when comparing against the Probit-predicted spectrum. Moreover we see quite high model confidence across all observations (see Figure 4.8). In liver metabolism, Ser15 acts as the central “on/off” switch for the enzyme; its phosphorylation by phosphorylase kinase (PHK) converts the inactive *phosphorylase b* into the active *phosphorylase a*, thereby driving the rate-limiting step of glycogenolysis—the breakdown of glycogen into glucose-1-phosphate [Zois2022].

Although PYGL is primarily known as the hepatic isoform, recent studies have highlighted its critical role in the “glycogen shunt” of cancer cells, particularly in glioblastoma (GBM). Under hypoxic conditions, tumor cells frequently undergo “isoform switching” or upregulate PYGL to utilize glycogen stores as a survival mechanism [Favaro2012]. High expression of PYGL has been linked to poor prognosis in glioma patients and is strongly associated with hypoxia-inducible factor (HIF) signatures [Zois2022, PURE\_Ulster].

Our detection of p-Ser15-PYGL in these samples serves as a functional marker for the metabolic state of the tumor. The presence of the active *phosphorylase a* form suggests that these cells are actively mobilizing glycogen to maintain energy homeostasis under metabolic stress. Furthermore, research in other cancer models (e.g., HCT116) has shown that p-Ser15 levels increase significantly under chemically induced hypoxia, paralleling other regulatory PTMs like O-GlcNAcylation [PMC9240045]. The ability of our *de novo* approach to confidently recover this specific regulatory site without prior database constraints demonstrates its potential to uncover metabolic drug targets and biomarkers that are central to cancer cell resilience.

# 5 Discussion

## 5.1 Key Findings

This work serves as a proof-of-concept that adapting transformer-based *de novo* sequencing models to TMT-labeled proteomics data is not only feasible but yields significant biological insights. By accounting for the non-trivial systematic mass shifts and complex fragmentation patterns of Tandem Mass Tags, the model successfully recovered a vast majority of database-identified peptides while doubling the number of unique sequences identified in complex glioma samples. The discovery of 694 genomic-validated SNPs and the identification of the metabolic “on-switch” p-Ser15 on PYGL demonstrate that the model can reliably move “beyond the database search engine” to uncover regulatory and structural variants that are invisible to traditional methods.

## 5.2 Limitations and Open Work

### 5.2.1 Data Challenges and Domain Adaptation Strategy

Despite the advancements in adapting *de novo* architectures for TMT data, several limitations persist. A primary challenge encountered during development was the relative sparsity of high-quality, large-scale TMT datasets that provide a “gold standard” ground truth. To address this, the **ProteomeTools** dataset was utilized. ProteomeTools consists of chemically synthesized peptides, providing an exceptionally reliable ground truth that is often unattainable in *in vivo* datasets, where peptide identification remains subject to search engine biases [Zolg2017].

The initial strategy to bridge the gap between unlabeled and labeled data involved the use of a “**replay set**”. In this context, a replay set refers to a balanced training mixture where a subset of previously learned, non-labeled HCD spectra is reintroduced alongside the new TMT-labeled data. The goal of this approach was to foster a stable knowledge transfer between domains and prevent catastrophic forgetting of the general peptide fragmentation patterns. However, subsequent **linear probing**—a diagnostic method where a simple linear

classifier is trained on top of the frozen encoder—revealed a distinct separation in the model’s internal representations. The results indicated that the encoder inherently learns to distinguish between TMT-labeled and non-labeled spectra within the embedding space, suggesting that the chemical shift introduced by the TMT tag creates a significant domain offset that the model must actively reconcile. This distinction suggests that the decoder’s predictive capacity remains somewhat restricted to the specific modification patterns present in the TMT training data, potentially hindering true cross-domain transfer learning. To overcome this, a domain-agnostic preprocessing step is one potential direction to generate a universal embedding space. Such a “foundation model” approach would allow the decoder to apply knowledge learned from vast non-TMT datasets to the specific challenges of multiplexed proteomics.

Additionally future refinement of the finetuning dataset could benefit from including unmodified TMT data to provide a more balanced representation of the peptide space. Finally, extending the model architecture with specialized “heads” for downstream tasks—such as PTM localization scoring or direct PTM classification—would further enhance the interpretability and confidence of the sequencing results [Zhu2017] or could be one task for itself.

### 5.2.2 Future Directions and Clinical Utility

The application of the developed model to complex datasets, such as those derived from glioma research, represents only the initial stage of its potential utility. While current collaborations provided a foundation, the depth of biological discovery can be significantly enhanced. Beyond identifying single amino acid variants, future iterations should integrate proteogenomic workflows. By incorporating six-frame translations or patient-specific transcriptomes, thousands of currently “high-confidence” but unassigned spectra could be systematically validated [Nesvizhskii2014]. This approach would facilitate the discovery of cryptic peptides, alternative splicing events, and non-canonical open reading frames (ORFs) that remain invisible to standard workflows.

Furthermore, the synergy between *de novo* sequencing and TMT multiplexing offers a scalable framework for precision medicine. By directly linking discovered variants to patient-specific MS3 channels, tumor heterogeneity can be characterized with unprecedented throughput. Future developments could focus on a real-time sequencing pipeline integrated with clinical metadata, transitioning mass spectrometry from a retrospective tool into a proactive platform for personalized oncology.

# Supplementary Material

Dieses Kapitel enthält zusätzliches Material zur Thesis.

## 5.3 Zusätzliche Abbildungen

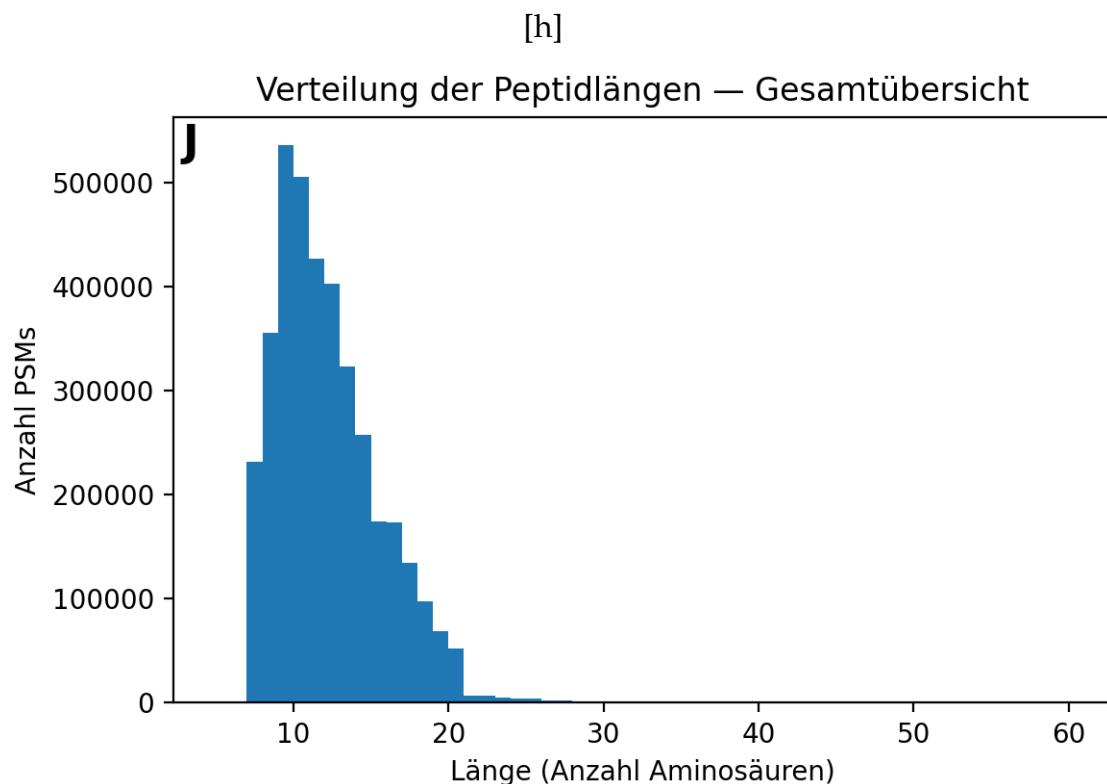


Figure S1: Histogram of peptide lengths for all data

## 5.4 Zusätzliche Tabellen

Weitere Tabellen...

Dies ist ein Mock-Beispiel.