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 (DBIS) remain the gold standard, they are inherently limited by a predefined search space, often overlooking rare post-translational modifications (PTMs) and single nucleotide polymorphisms (SNPs). 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 by TMT-induced features. Iterative hyperparameter optimization, specifically expanding the isotope error window and utilizing a beam search width 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 genomically 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

Hier kommt eine kurze Zusammenfassung der Thesis...
Dies ist ein Mock-Beispiel. Passen Sie den Inhalt an.

Contents

List of Figures

List of Tables

1 Introduction

1.1 The Challenge of the "Dark Proteome" and PTMs

Mass spectrometry (MS)-based proteomics has revolutionized our understanding of cellular processes by allowing the large-scale identification and quantification of proteins. However, a significant portion of the acquired MS/MS spectra—often referred to as the "dark proteome"—remains unassigned by traditional database-driven identification strategies (DBIS) [**Aebersold2016**]. One major reason for this gap is the immense complexity of post-translational modifications (PTMs). PTMs, such as phosphorylation, ubiquitination, and methylation, are crucial regulators of protein function, localization, and signaling pathways, especially in diseases like cancer [**Mertins2016**].

Traditional search engines like MaxQuant or Sequest rely on predefined databases. Including a wide array of PTMs in these searches leads to a combinatorial explosion of the search space, which not only increases computational time but also drastically reduces statistical power and increases the false discovery rate (FDR) [**Chick2015**]. Consequently, DBIS are often "blind" to unexpected or multiple PTMs occurring on the same peptide, leaving biologically critical information hidden in the unassigned data.

1.2 De Novo Sequencing and the TMT Gap

De novo peptide sequencing offers a powerful alternative by predicting the amino acid sequence directly from the fragment spectra without a reference database. Recent advances in deep learning, particularly transformer-based models like *Casanovo* and its PTM-specialized derivative *Modanovo*, have pushed the boundaries of accuracy in this field [**Klaproth2024**]. These models have shown great potential in uncovering novel PTM patterns and even single nucleotide polymorphisms (SNPs) in highly mutated samples, such as those found in glioma research [**Smith2019**].

Despite these advances, a significant barrier remains: Tandem Mass Tag (TMT) labeling. TMT is the gold standard for multiplexed quantitative pro-

teomics, enabling the simultaneous analysis of multiple clinical samples. However, TMT labeling introduces systematic mass shifts and significantly alters the fragmentation patterns (e.g., favoring b-ions over y-ions) [**Shen2018**]. Current *de novo* models, including *Modanovo*, were primarily trained on label-free data and therefore fail to accurately interpret TMT-labeled spectra. This creates a critical bottleneck for clinical proteogenomics, where TMT is the preferred workflow.

1.3 Objectives of this Thesis

The objective of this thesis is to bridge the gap between high-performance *de novo* sequencing and TMT-based quantitative proteomics. We propose an adaptation of the *Modanovo* framework specifically designed to handle the chemical signatures of TMT labeling. The focus lies on:

- **Fine-Tuning:** Adapting the model weights using a large-scale, TMT-labeled dataset to learn specific mass shifts and fragmentation biases.
- **Architectural Enhancements:** Expanding the token vocabulary to include TMT-specific modifications.
- **Conditioning:** Implementing covariate embeddings to explicitly inform the model about the TMT status of a spectrum.
- **Evaluation:** Validating the model on a complex glioma dataset to demonstrate its ability to recover PTMs and SNPs where standard DBIS like MaxQuant reach their limits.

By unlocking the sequence information in TMT-labeled "dark" spectra, this work aims to provide deeper biological insights into the regulatory networks of cancer.

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

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**]. 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 limited by the predefined "search space." Although open search strategies exist to identify PTMs without *a priori* definition, traditional closed searches can only identify sequences explicitly included in the database. Consequently, peptides from non-model organisms or those with unexpected biological variants are frequently missed. Including all possible modifications in a 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 adjacent 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 isomers 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 seq2seq 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 (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 [**Qiao2021**]. 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

Most modern *de novo* sequencing models generate peptide sequences autoregressively, predicting one amino acid at a time conditioned on previously predicted residues and the encoded spectrum. Beam search is commonly applied during inference to explore multiple high-probability candidate sequences simultaneously.

This strategy allows the model to balance local 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. DeepNovo pioneered the use of CNNs for feature extraction [**Tran2017**], while newer architectures like InstaNovo and π-Prime focused on

increasing 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**].

Despite these advances, a significant gap remains regarding Tandem Mass Tag (TMT) proteomics.

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.

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**].

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.

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

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

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

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

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.

3.1.1 Multi-task Learning Architecture

Building on the modularity of Modanovo, a multi-task learning head was evaluated. This architectural extension aims to decouple the prediction of the amino acid backbone from the specific PTM state by utilizing a dedicated PTM prediction head [**KlaprothAndrade2025**].

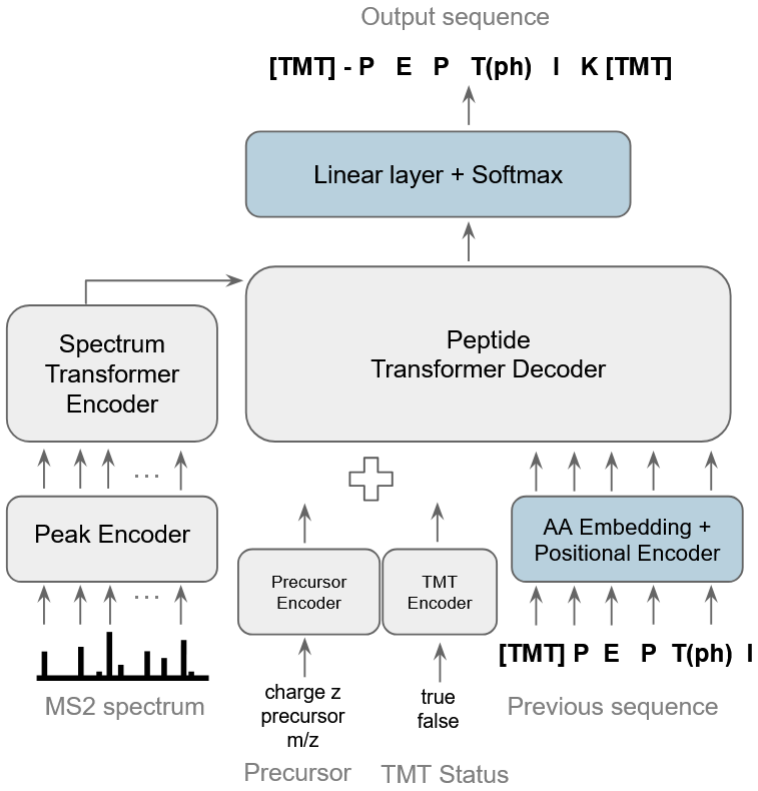

Figure 3.1: Schematic representation of the adapted transformer architecture. The TMT status is fed as a covariate embedding into the decoder alongside the precursor and previous predicted residues. Moreover vocabulary is expanded with TMT tokens.

3.2 Data Selection and Training Protocol

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 ground truth sequences associated with high-resolution fragment spectra. Data were integrated from various sources, initially stored in CSV and mzML 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**].

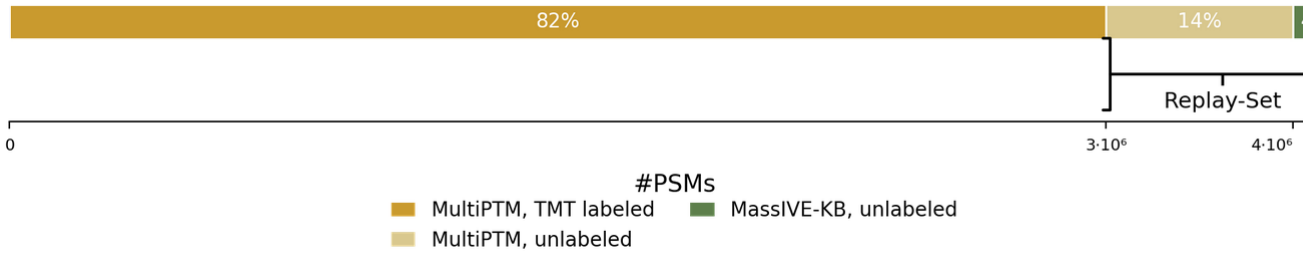

Figure 3.2: Composition of the fine-tuning dataset showing the distribution of Peptide-Spectrum Matches (PSMs). The majority (82%) consists of TMT-labeled MultiPTM data, complemented by a 18% Replay-Set to maintain performance on unlabeled spectra.

3.3 Fine-Tuning Data

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).

3.3.1 TMT-labeled Dataset

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 ϵ-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.

3.3.2 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.

3.3.3 Modification Distribution and Heatmap Analysis

To evaluate the coverage of the training data, we visualized the modification density across different residues. This approach was closely oriented towards the *ModaNovo* framework to ensure comparability and systematic evaluation of PTM-residue pairs [**KlaprothAndrade2025**].

As shown in Figure 3.5, the current data selection strategy provides high coverage for common PTMs such as Oxidation (M) and Phosphorylation (S, T, Y). However, there remains significant optimization potential. Specifically, certain rare PTM-residue combinations or complex multiplexed modifications (e.g., Ubiquitinylation + TMT) exhibit lower spectral counts, which might limit the model's ability to generalize on extremely sparse clinical samples.

3.3.4 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.

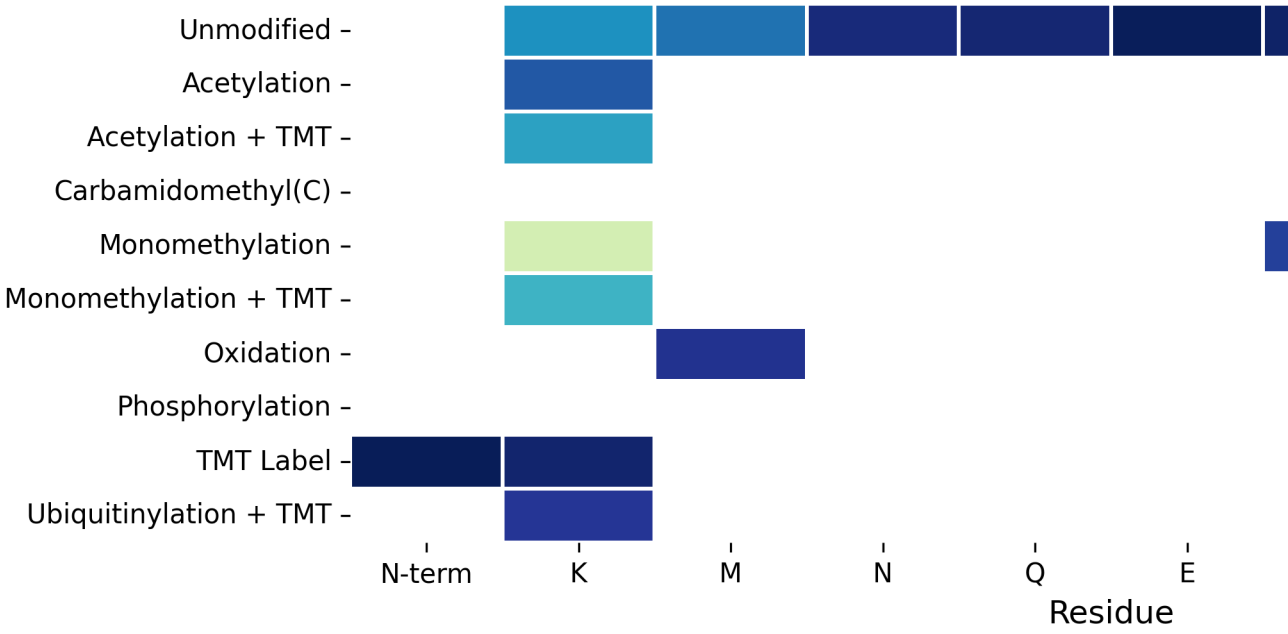

Figure 3.3: Modification distribution for the TMT-labeled dataset.

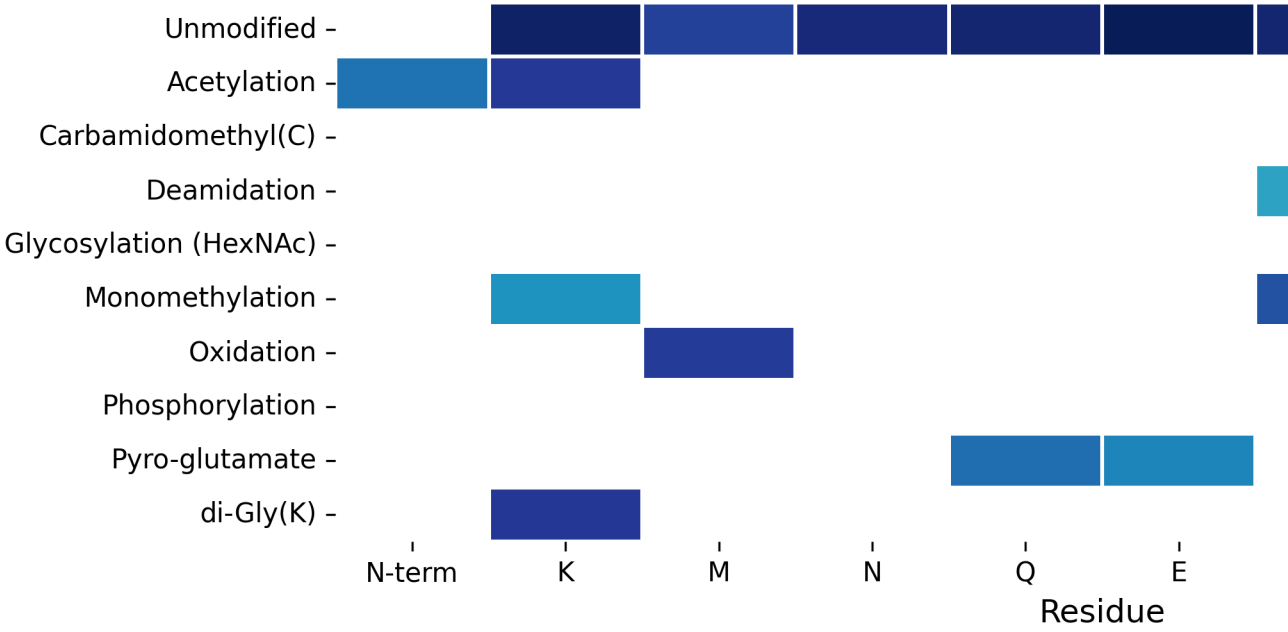

Figure 3.4: Modification distribution for the non-TMT Replay Set.

Figure 3.5: Heatmaps showing the log-scaled spectral counts for various PTMs across amino acid residues. While the TMT dataset (a) shows high density for N-terminal and Lysine labeling, the Replay Set (b) provides a broader PTM diversity including Phosphorylation and Glycosylation.

- **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.

3.3.5 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 1×10^{-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 1×10^{-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 Casanovo 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 physical 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:

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

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

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 amino acids such as Leucine and Isoleucine as equivalent **[KlaprothAndrade2025]**.

3.4.3 Modification-Specific Evaluation

To uncover biological insights beyond standard searches, 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.

3.5 Downstream Validation and Biological Integration

To evaluate the practical utility of the model, it was applied to an independent Glioma cancer dataset.

3.5.1 Data Acquisition and Preprocessing

The primary data source consisted of Orbitrap-based mass spectrometry raw files (.raw) from cancer proteomic studies. To ensure compatibility with the deep learning framework, raw files were converted into .mgf format using the `ThermoRawFileParser` (v2.0.0), which facilitates the extraction of metadata and spectral information **[Hulstaert2020]**. All MS1 (precursor scans) and MS3 scans were removed for the analysis.

De Novo Sequencing Configuration (finetuned ModaNovo) 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 datasets. The isotope error range was restricted to $[0, 3]$.
- **Spectrum Processing:** To reduce noise, only the 150 most intense peaks per spectrum (`n_peaks`) within an m/z range of 50.0 to 2500.0 were retained. Peaks within a 2.0 Da window of the precursor m/z were removed to prevent interference.
- **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.
- **Quality Control:** Only spectra with a precursor charge of ≤ 10 and a relative intensity threshold of 0.01 were processed.

Database Search Configuration (MaxQuant) To provide a complementary ground-truth baseline, all Glioma datasets were processed using MaxQuant (version 2.1.3.0) [**Tyanova2016**]. The search was conducted against the human reference proteome (UniProt UP000005640).

The search parameters were harmonized with 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) was set as a static modification.
- **Variable Modifications:** To capture the regulatory landscape of the cancer samples, oxidation (M) and phosphorylation (S/T/Y) were included in the search space.

The results from this database-driven approach serve as the benchmark for calculating the precision and coverage of the de novo model's predictions.

3.5.2 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.

BLASTp Configuration and Ambiguity Handling Alignment was executed using `Protein-Protein BLAST` (version 2.17.0+) against the same human reference proteome (UniProt UP000005640) used for the MaxQuant search. To account for the inherent mass-spectrometric ambiguities where different amino acid compositions result in near-identical masses, a specialized encoding strategy was implemented:

- **I/L Equivalence:** Leucine and Isoleucine were treated as identical.
- **Mass-Based Ambiguitites:** Specific residues with overlapping mass shifts, such as the deamidation of Glutamine ($Q + 0.984$ Da) resulting in the same mass as Glutamic acid (E), or Asparagine deamidation matching Aspartic acid (D), were addressed.
- **Ambiguity Codes:** All unmodified Aspartic acid (D) residues in the query were replaced with the code `B` (representing D or N), and all unmodified Glutamic acid (E) residues were replaced with `Z` (representing E or Q). The `PAM30` substitution matrix was utilized, as it natively handles these ambiguity codes to allow perfect matches against either potential target residue.

The alignment parameters were set to an E-value of 2000, `qcov_hsp_perc` of 80, and `comp_based_stats` disabled to prevent score inflation for short peptide sequences.

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.

3.5.3 Genomic Evidence

3.5.4 Spectral Quality

4 Results

4.1 Model Training and Convergence

The model was trained until the validation loss showed early signs of stagnation, indicating the onset of overfitting. Convergence was reached after 13 epochs, corresponding to approximately 1,550,000 training steps. To assess the final performance, the model state at this checkpoint was utilized for inference on the complete, independent test dataset.

4.2 Performance Evaluation on the test set

To evaluate the predictive confidence, precision-coverage curves were generated for the different PTM categories and stratified by TMT vs non-TMT data. The performance is quantified using the Area Under the Precision-Coverage Curve (AUPCC).

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.

4.2.1 Comparative Performance: Non-TMT and Modanovo Baseline

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

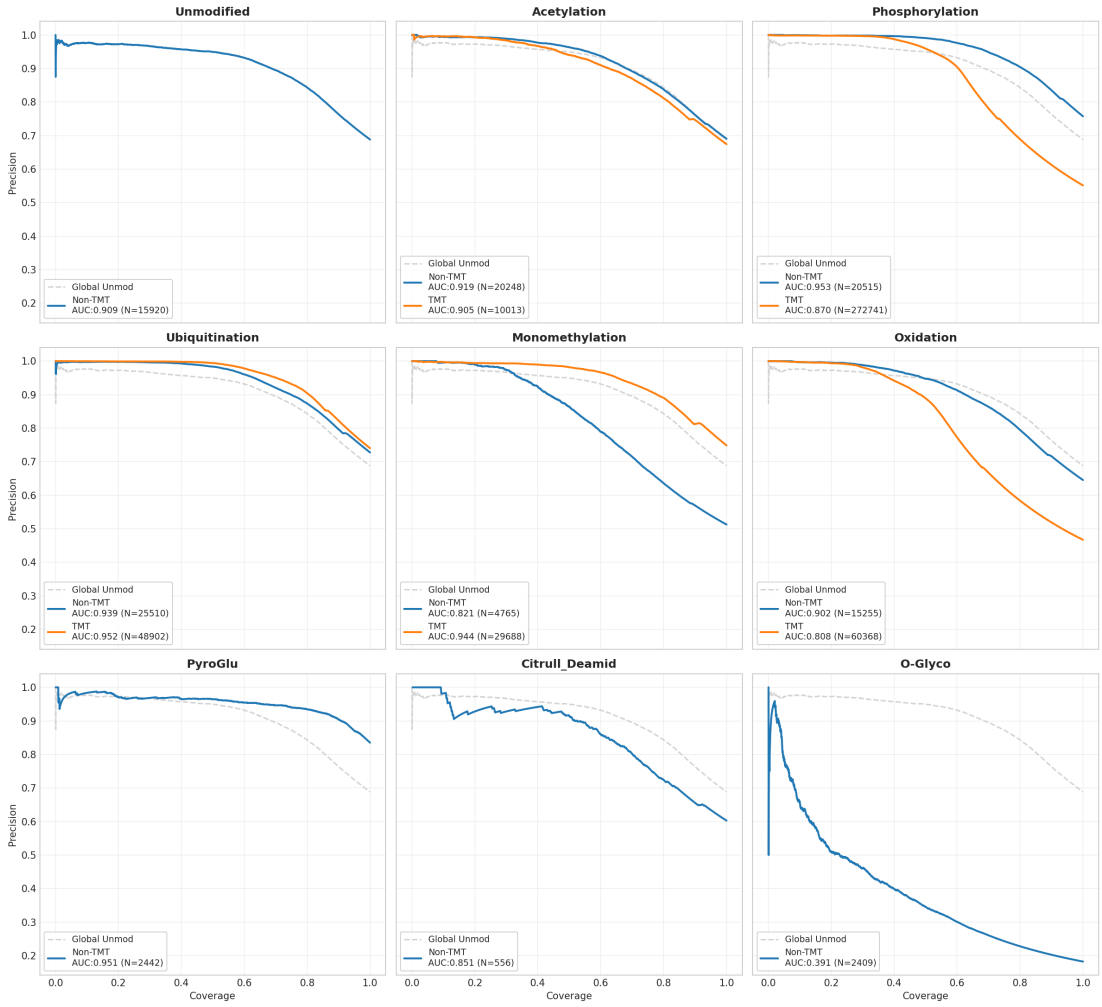

Figure 4.1: Precision-coverage curves across different PTMs. The performance of the final model on TMT-labeled (orange), non-TMT (blue) spectra is shown. Unmodified peptides are shown in light grey for comparison.

This stability is further reflected in the specific PTM performance. For instance, the model demonstrated similar accuracy for Phosphorylation on non-TMT data (AUPCC 0.95) staying aligned with the general Modanovo median for phosphorylation (0.93–0.96 depending on the residue) [**Yilmaz2022**]. Similarly, high performance was observed for Pyro-Glu (AUPCC 0.95), likely due to its restricted occurrence at the peptide N-terminus, a pattern the model successfully internalizes. In contrast, O-Glycosylation (S/T[+203.079]) remains a significant challenge with an AUPCC of 0.39, mirroring the difficulties reported in literature due to low training examples and complex fragmentation [**Yilmaz2022**].

4.2.2 Generalization on TMT-Labeled Data

The global performance on TMT data (AUPCC 0.89) is slightly lower than on non-TMT data (AUPCC 0.92), reflecting the increased spectral complexity introduced by the chemical labels.

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.
- **Oxidation:** A noticeable decrease in performance was observed for TMT-labeled oxidized peptides (AUPCC 0.81) compared to non-TMT (AUPCC 0.90), pointing towards potential interference between the TMT-label fragments and the neutral losses associated with methionine oxidation.

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 Hyperparameter Optimization and Isotopic Sensitivity

A critical challenge identified during the evaluation of TMT-labeled data was the increased frequency of non-monoisotopic precursor selection. In TMT mul-

tiplexed 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.

4.3.1 Isotope Error Range and Beam Search Synergy

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 "off-by-one" or "off-by-two" Dalton errors.

Furthermore, while literature for models like Casanovo 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.

4.3.2 Quantitative Impact of Configuration Changes

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

4.3.3 Impact on Specific Modification Classes

The transition to Config FINAL particularly benefited complex PTMs:

- **Ubiquitination:** This category showed a massive 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.

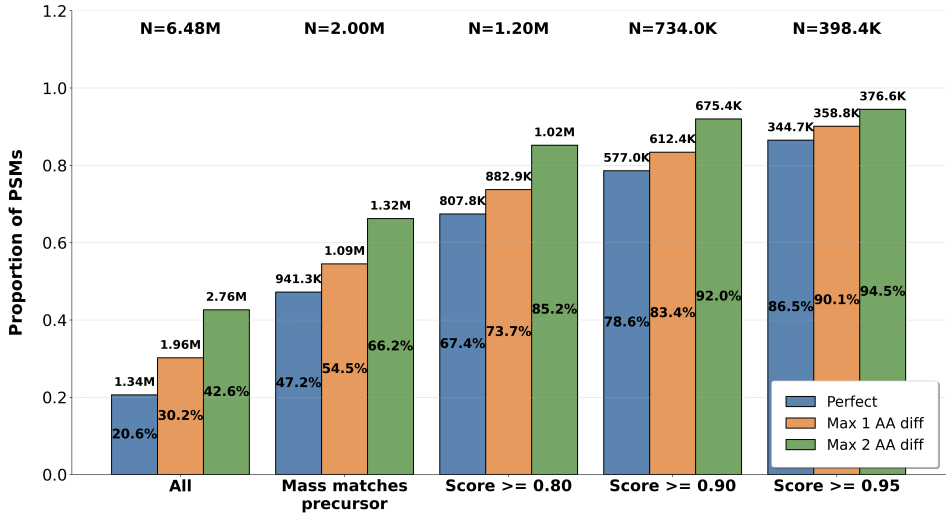

Figure 4.2: Alignment of *de novo* predictions to the human reference proteome. The proportion of perfect alignments (blue) increases significantly with higher confidence scores. At a score cutoff of > 0.95, 86.5% of the 398.4k predictions align perfectly.

The upward trend of the blue 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 already known to the genome, providing a solid baseline for the subsequent analysis of modified and novel peptides.

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.

As shown in Figure 4.3, 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

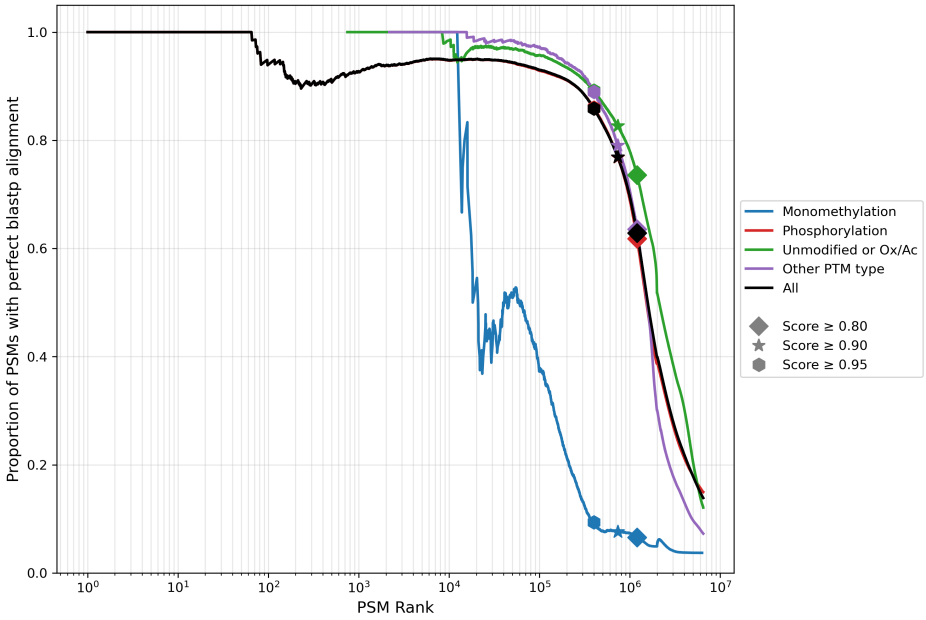

Figure 4.3: Proportion of perfectly aligned PSMs stratified by modification type and ranked by confidence score.

of the methyl group might still lead to mismatches during genomic alignment in real-world samples.

4.4.3 Summary of Validation

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 Uncovering the Dark Proteome: Comparison with 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 pro-

teome—spectra that contain high-quality information but remain unidentified by traditional search engines.

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 4.4, the *de novo* approach identified approximately 40,000 unique peptides that were also found by MaxQuant, demonstrating high consistency in the "known" space.

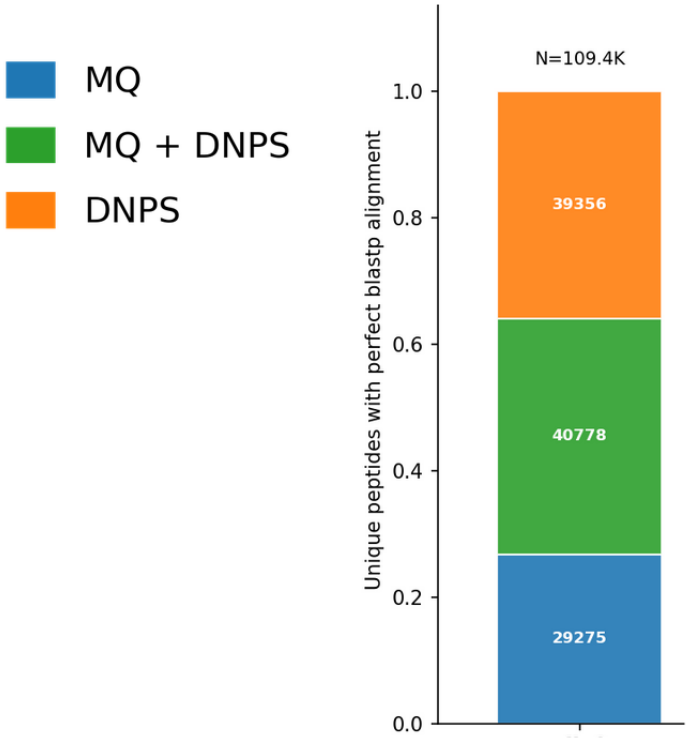

Figure 4.4: Overlap of unique peptide identifications between the proposed *de novo* model and MaxQuant. The model successfully recovers the majority of MQ identifications while contributing 40,000 additional unique sequences.

Crucially, the model discovered an additional 40,000 unique peptides that were entirely absent from the MQ results. Conversely, only a negligible fraction (approximately 30 peptides) was identified by MQ but missed by our model, indicating that the *de novo* approach covers nearly the entire search space of the database-driven method while doubling the number of unique sequences.

4.5.2 Identification of Previously Unidentified Spectra

The true potential of the model is reflected 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. These spectra likely represent peptides with unexpected modifications or amino acid substitutions (SNPs) that shift the precursor mass beyond the tolerance of the theoretical database entry, yet yield clear fragmentation patterns.

Summing these categories, the *de novo* approach provides plausible sequence candidates for over 670,000 spectra that were previously discarded in the MQ workflow. This substantial increase in spectral utilization highlights the model's ability to move beyond the "closed-search" paradigm.

4.5.3 From Global Discovery to Specific 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. In the following sections, we will categorize these findings into specific biological phenomena, namely novel phosphorylation sites (p-sites) and single nucleotide polymorphisms (SNPs).

4.5.4 Discovery of SNP-Related Variants

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. 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 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 $\rightarrow$ GCG).

These findings demonstrate that the *de novo* model effectively captures protein-level manifestations of genetic variability that remain invisible to standard "closed-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**].

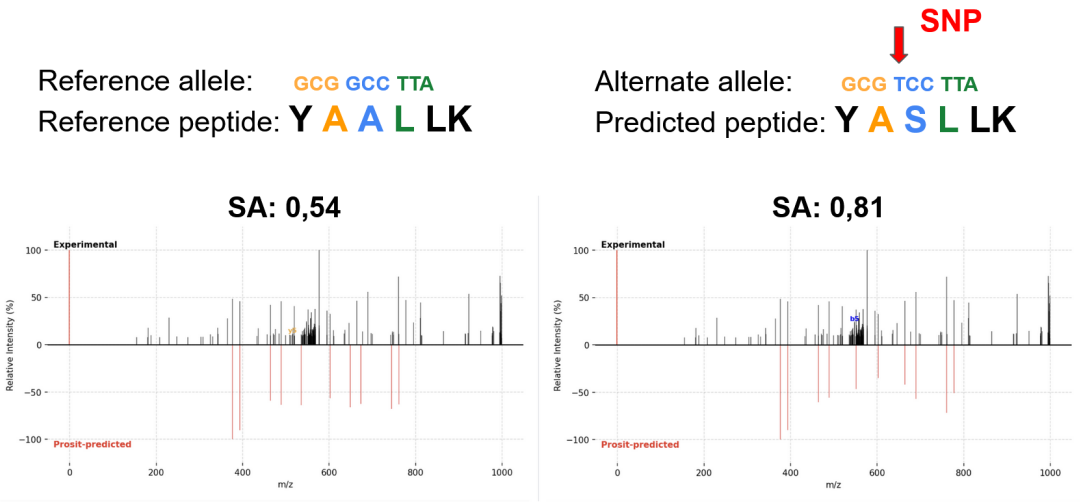

Figure 4.5: Spectral validation of identified SNPs using Prosit-predicted theoretical spectra. The high degree of overlap between experimental and theoretical fragment ions confirms the sequence accuracy of the identified variants.

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.

4.5.5 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.

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.

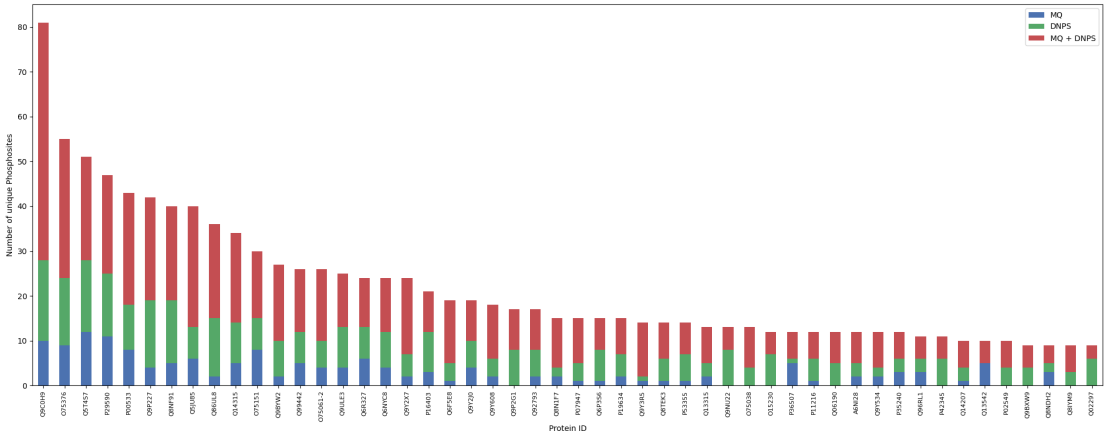

Figure 4.6: Distribution of unique phosphorylation sites identified by MaxQuant (MQ) and the proposed *de novo* peptide sequencing (DNPS) model for selected cancer-related proteins. The results demonstrate a significant expansion of the detectable phospho-landscape.

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 4.6). 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 PPARα [**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).

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 across all observations (see Figure 4.7). In liver metabolism, Ser15 acts as the central "on/off" switch for the enzyme; its phosphorylation by

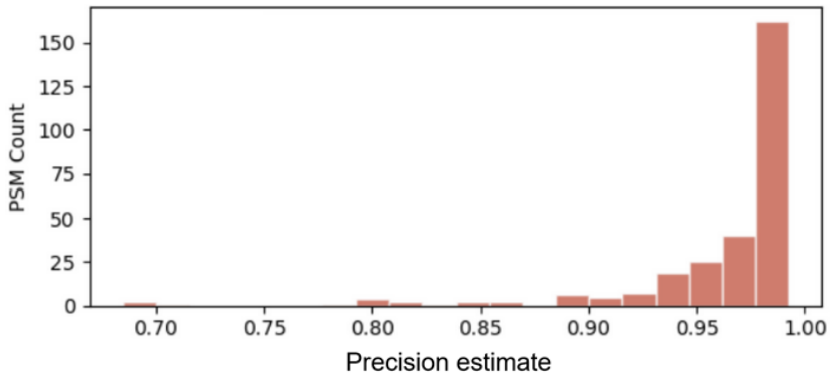

Figure 4.7: Confidence distribution and spectral evidence for p-Ser15 on PYGL. The identification is supported by 279 PSMs with a mean spectral angle (SA) of 0.7, indicating high reproducibility and spectral fidelity.

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 Discussion and Future Directions

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 systematic mass shifts 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" to uncover regulatory and structural variants that are invisible to traditional search engines.

5.1.1 Addressing Current Limitations and Model Refinement

Despite these promising results, the analysis of high-density modification sites in proteins like SRCIN1 highlights the need for more sophisticated scoring mechanisms. Currently, the model's confidence scores do not always distinguish between a correct sequence with a misplaced modification and a truly correct modified peptide. Future iterations must implement a dedicated localized scoring system to improve PTM site assignment and reduce false-positive rates in the "dark proteome" space [**Lim2021**].

Furthermore, the adaptation process revealed that maintaining a "replay set" of non-TMT data is likely unnecessary. Since TMT-based workflows represent a distinct experimental paradigm, the model should prioritize mastering the altered fragmentation patterns and reporter ion interferences inherent to multiplexing. To overcome the scarcity of high-quality, labeled training data, a targeted transfer learning approach should be employed. By fine-tuning the model on large-scale, unlabeled TMT datasets and incorporating un-mutated TMT-sequences as a baseline, the model can achieve a deeper understanding of the TMT-specific search space without being diluted by irrelevant non-labeled spectral features [**Zhu2023**].

5.1.2 Expanding the Horizon of Discovery

The potential for discovery extends beyond single amino acid variants. Future applications should integrate 6-frame translations of the genome or patient-specific transcriptomes to systematically validate the thousands of currently unassigned "high-confidence" spectra identified by the model. Such an interdisciplinary "proteogenomic" pipeline would allow for the discovery of cryptic peptides, alternative splicing events, and non-canonical open reading frames (ORFs) [**Nesvizhskii2014**].

Ultimately, the synergy between *de novo* sequencing and TMT multiplexing provides a scalable framework for precision medicine. Linking discovered variants directly to patient-specific MS3 channels enables a high-throughput characterization of tumor heterogeneity. Future work will focus on tighter integration with clinical metadata and the development of a real-time sequencing pipeline, transforming mass spectrometry from a retrospective analysis tool into a proactive discovery platform for personalized cancer therapy.

Supplementary Material

Dieses Kapitel enthält zusätzliches Material zur Thesis.

5.2 Zusätzliche Abbildungen

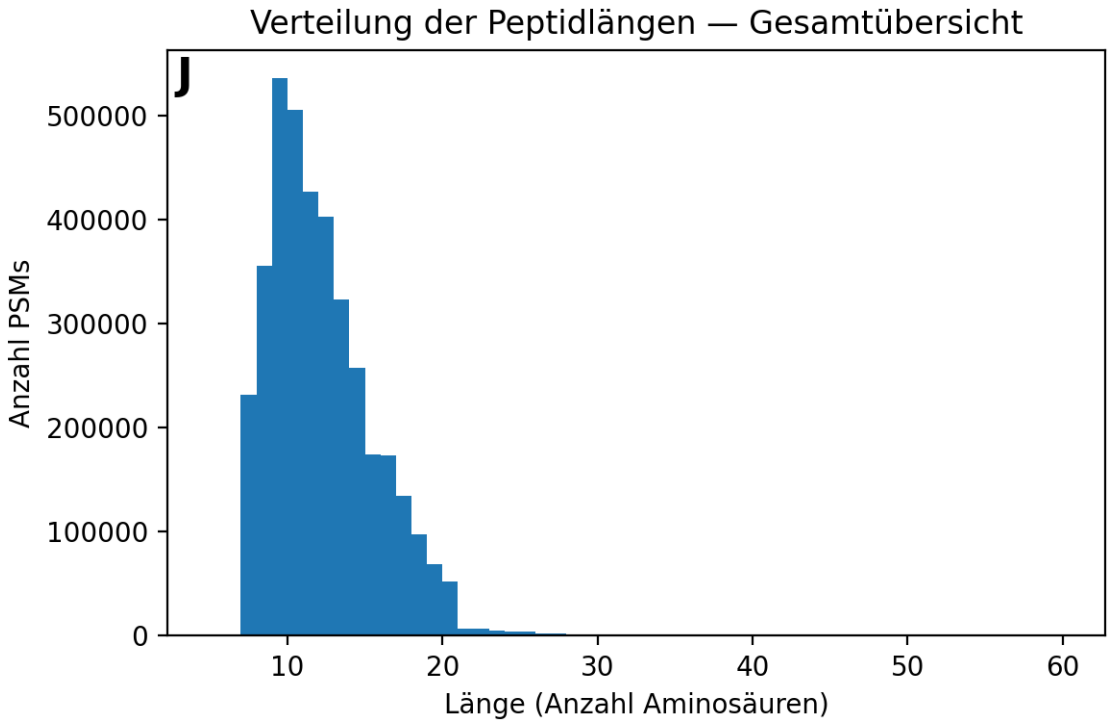

Figure S1: Histogram of peptide lengths for all data

5.3 Zusätzliche Tabellen

Weitere Tabellen...
Dies ist ein Mock-Beispiel.

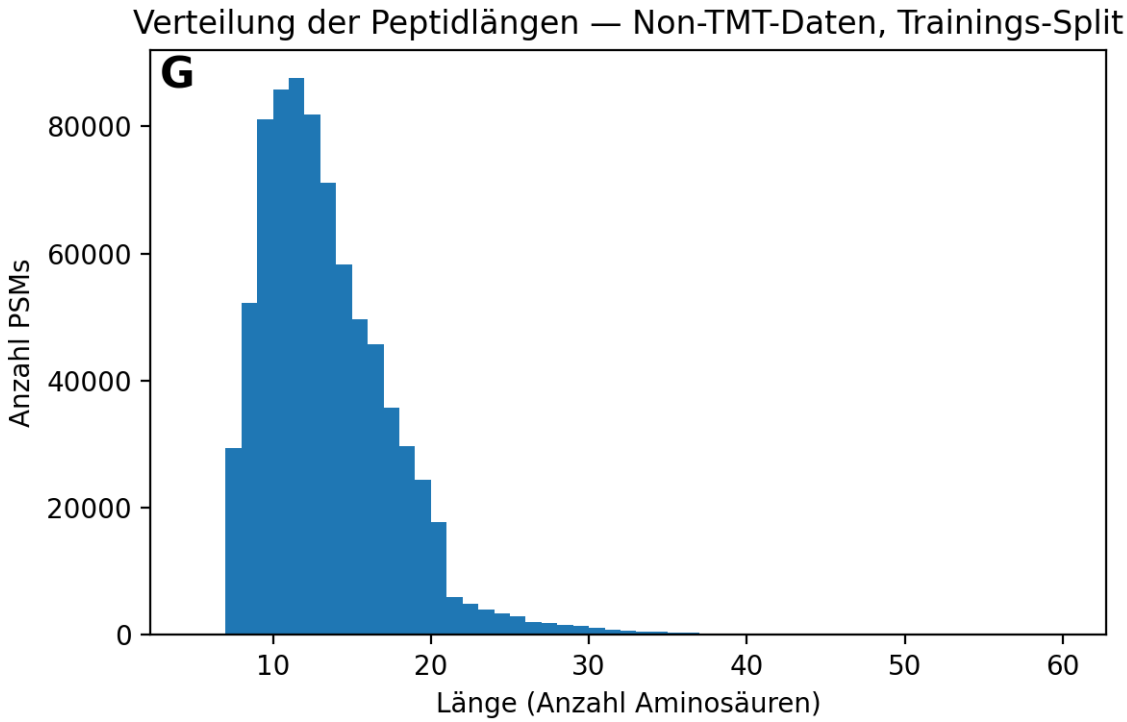

Figure S2: Histogram of peptide lengths for training data (TMT false)

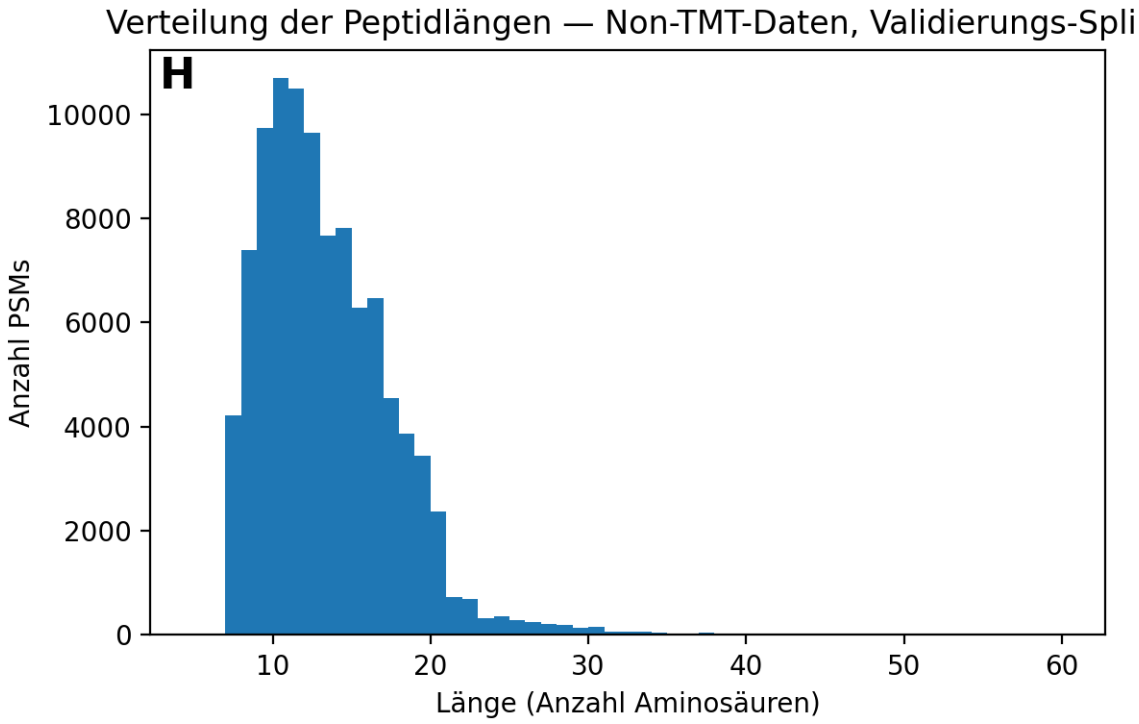

Figure S3: Histogram of peptide lengths for validation data (TMT false)