Acknowledgments

If someone contributed to the thesis... might be good to thank them here.

Abstract

An abstract abstracts the thesis...

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 Introduction

1.1.1 The Role of Proteomics in Clinical Cancer Research

The advent of precision medicine has shifted the focus from broad therapeutic approaches to individualized treatment strategies, particularly in oncology. Proteomics, the large-scale study of proteins and their functions, provides a functional snapshot of the cell that genomics alone cannot capture [**Aebersold2016**]. In clinical cancer research, specifically in the study of complex malignancies such as gliomas, mass spectrometry (MS)-based proteomics is indispensable for identifying biomarkers and therapeutic targets [**Mertins2016**]. By analyzing the proteome, researchers can observe the actual effectors of biological processes, making it possible to understand tumor heterogeneity and drug resistance mechanisms at a molecular level.

1.1.2 Limitations of Database-Informed Search (DBIS) Strategies

The standard paradigm for peptide identification is the Database-Informed Search (DBIS). While effective for well-characterized organisms, DBIS faces the "search space challenge." As the number of considered post-translational modifications (PTMs) and single nucleotide polymorphisms (SNPs) increases, the theoretical search space expands exponentially, leading to a significant loss in statistical power and increased false discovery rates [**Chick2015**]. Consequently, rare but biologically significant modifications—often crucial in cancer signaling—remain "hidden" because they were not explicitly included in the search parameters.

1.1.3 De Novo Sequencing: Unbiased Identification of Peptides and PTMs

To overcome the constraints of predefined databases, *de novo* peptide sequencing has emerged as a powerful alternative. This approach predicts the amino acid sequence directly from the MS/MS fragment ions without prior knowledge of the proteome. Recent breakthroughs in deep learning, particularly transformer-based architectures like ModaNovo or Casanovo, have significantly increased the accuracy of these predictions [**Yilmaz2022**, **Rappsilber2024**]. These models can theoretically identify any peptide sequence, including those with unexpected modifications, making them ideal for discovering novel proteoforms in clinical samples.

1.1.4 Challenges of Multiplexing in De Novo Sequencing

Quantitative proteomics often relies on multiplexing techniques such as Tandem Mass Tags (TMT) to increase throughput and reduce technical variability across clinical cohorts. However, TMT labeling introduces significant complexity into MS/MS spectra. The chemical tags add substantial mass to the N-terminus and lysine residues, and the fragmentation process produces high-intensity reporter ions in the low m/z range [**Thompson2003**]. These systematic shifts and altered intensity patterns are not well-represented in standard training datasets for *de novo* algorithms, leading to a drop in performance when analyzing multiplexed data.

1.1.5 Problem Statement and Research Gap

Despite the progress in deep learning-based *de novo* sequencing, a significant gap remains: most state-of-the-art models are trained on unlabeled data. When applied to TMT-labeled samples, models like ModaNovo often fail to correctly interpret the mass shifts and the altered fragmentation behavior. There is currently a lack of specialized *de novo* sequencing architectures or fine-tuning strategies that can handle the unique chemical footprint of TMT labeling while simultaneously identifying diverse PTMs. This limitation prevents the full utilization of multiplexed datasets for discovering novel biological insights beyond the reach of database searches.

1.1.6 Objectives and Contributions

The objective of this thesis is to bridge this gap by expanding *de novo* peptide sequencing capabilities to TMT-labeled proteomics data. We propose to evaluate and adapt transformer-based models—specifically focusing on fine-tuning and conditioning strategies—to account for the systematic mass shifts introduced by TMT. By integrating diverse PTMs into this framework, this work aims to provide a robust tool for the unbiased identification of modified peptides in multiplexed experiments, ultimately uncovering regulatory networks in cancer biology that are overlooked by traditional methods.

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 separated by liquid chromatography (LC) and ionized (e.g., via Electrospray Ionization, ESI) to be transferred into the gas phase for mass spectrometric analysis [**Aebersold2016**].

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). In this process, a specific precursor ion is isolated based on its mass-to-charge ratio (m/z) and subsequently subjected to fragmentation [**Steen2004**]. In high-resolution instruments like the Orbitrap, Higher-energy Collisional Dissociation (HCD) is the preferred method, producing a predictable pattern of fragment ions.

According to the established peptide fragment ion theory, the fragmentation of the peptide backbone occurs primarily at the amide 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**]. By measuring the mass difference between consecutive ions in a series, the corresponding amino acid can be inferred, as each amino acid (except for the isomers Leucine and Isoleucine) possesses a unique residual mass [**Steen2004**]. However, the presence of post-translational modifications (PTMs) or chemical labels like Tandem Mass Tags (TMT) shifts these masses, increasing the complexity of the spectra and necessitating advanced computational strategies for identification.

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 approaches.

2.2 Peptide Identification Strategies

2.2.1 Database Search Engines (DBIS)

The most prevalent method for peptide identification is database searching. This strategy relies on a predefined protein sequence database (e.g., UniProt). Computational search engines, such as 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, often referred to as a Peptide-Spectrum Match (PSM) [**Eng1994**]. While highly robust, DBIS is inherently limited by the "search space" problem: it can only identify peptides and modifications that are explicitly included in the database. Consequently, rare or novel post-translational modifications (PTMs) are frequently missed because including all possible modifications would lead to a combinatorial explosion, drastically increasing false discovery rates and computational costs [**Nesvizhskii2010**].

2.2.2 Principles of 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 treats the spectrum as a puzzle where the mass differences between adjacent peaks are mapped to the masses of amino acids.

Historically, de novo sequencing was limited by spectral noise and incomplete fragmentation, which often led to gaps in the predicted sequence. However, modern approaches utilize deep learning architectures—specifically

Transformer-based models—to capture long-range dependencies between fragment ions and their intensities [**Yilmaz2023**]. Because de novo sequencing does not depend on a database, it is uniquely suited for discovering novel PTMs, identifying peptides from non-model organisms, and uncovering biological variants that remain "dark" to traditional search engines.

However, the chemical environment of the peptide significantly influences its fragmentation, which is particularly evident when using isobaric tags.

2.3 Tandem Mass Tag (TMT) Labeling

2.3.1 Isobaric Labeling Chemistry and Multiplexed Quantitative Proteomics

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 inherently minimizes technical batch effects and ensures consistent quantification across all multiplexed channels, effectively overcoming the 'missing value' challenge [**Rauniyar2014**]. Furthermore, the use of a 'carrier channel'—an isobaric spike-in of a high-abundance proteome—boosts the precursor signal of low-input samples, significantly enhancing the identification and de novo sequencing depth of low-abundance peptides in single-cell or limited-material applications [**Budnik2018**].

2.3.2 Impact of TMT on Fragmentation Patterns

The use of TMT tags introduces systematic changes to the peptide fragment spectra. Upon fragmentation (typically 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, which are used for quantification [**McAlister2014**]. However, for de novo sequencing, TMT labeling presents a challenge: the tag adds a significant, constant mass shift to the N-terminus and lysine residues. Furthermore, the presence of the bulky tag can alter the gas-phase basicity and the fragmentation efficiency of the peptide backbone, often leading to different relative intensities of b- and y-ions compared to unlabeled peptides [**Hogrebe2018**].

While TMT tags provide a predictable mass shift, naturally occurring modifications are far more diverse.

2.4 Post-Translational Modifications (PTMs)

2.4.1 Biological Significance and Diversity

PTMs, such as phosphorylation, acetylation, and ubiquitination, exponentially increase the proteome's complexity by altering protein function, localization, and stability. They are critical for cellular signaling but are often present in low substoichiometric amounts, making their detection challenging.

2.4.2 Mass Shifts, Diagnostic Ions, and Limitations

Each PTM induces a specific mass shift in the precursor and fragment ions (e.g., $+79.966$ Da for phosphorylation). Some modifications also produce "diagnostic ions"—specific fragment peaks that indicate the presence of a PTM but do not provide sequence information. Standard DBIS methods fail when the modification is not predefined in the search space or when multiple modifications occur on the same peptide, leading to a high "dark proteome" fraction that only de novo sequencing can resolve [**Nesvizhskii2010**].

To address these challenges, modern computational biology has turned to advanced machine learning.

2.5 Deep Learning and Transformer Models

2.5.1 Neural Networks for Sequence Modeling

The identification of peptides from MS/MS spectra can be framed as a sequence-to-sequence (Seq2Seq) translation task, where a sequence of mass peaks is translated into a sequence of amino acids. Early deep learning approaches in this field utilized Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) units to handle the sequential nature of peptides [**Tran2019**]. However, RNNs suffer from vanishing gradients and struggle to capture long-range dependencies between distant fragment ions, which is critical for resolving complex PTM patterns.

2.5.2 The Transformer Architecture and Attention Mechanism

The introduction of the Transformer architecture revolutionized sequence modeling by replacing recursion with Self-Attention [**Vaswani2017**]. The core innovation is the Scaled Dot-Product Attention, which allows the model to weigh the importance of different peaks in a spectrum simultaneously, regardless of their distance. In a proteomic context, this means the model can correlate a low-intensity b-ion at the beginning of the spectrum with a corresponding y-ion at the end, significantly improving the reconstruction of the peptide backbone.

2.6 Transformer-based De Novo Framework

2.6.1 Spectrum Encoding and Embedding

The first step in a Transformer-based de novo framework is the transformation of raw MS/MS data into a high-dimensional representation. Unlike text, where tokens are discrete, MS/MS peaks are defined by continuous m/z values and intensities. Modern models use Point-based Encoding or Binning strategies to embed these values into a latent space [**Yilmaz2023**]. This embedding allows the Transformer's encoder to extract structural features from the fragmentation pattern, even when shifted by TMT labels or PTMs.

2.6.2 Autoregressive Sequence Generation and Beam Search

The decoder of the Transformer predicts the peptide sequence amino acid by amino acid in an autoregressive manner. At each step, the model calculates a probability distribution over the possible amino acids (the "vocabulary") based on the precursor m/z, previously predicted residues and the encoded spectrum [**Yilmaz2023**].

To optimize this process, Beam Search is employed instead of a simple greedy search. Beam Search maintains a set of k most likely sequences (the 'beam width') at each step, exploring multiple paths simultaneously [**Tran2017**]. This approach effectively balances local amino acid probabilities with the global constraint of the precursor mass, preventing the model from being trapped in suboptimal paths that would otherwise lead to mass discrepancies or invalid peptide sequences.

3 Datasets

This chapter describes the data resources used for the fine-tuning and evaluation of the model. All spectral data were acquired using high-resolution mass spectrometry and represent a diverse range of peptide sequences and modifications.

3.1 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.

3.1.1 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. The raw data are hosted via the PRIDE archive and ProteomeXchange [**Perez-Riverol2022**].

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 (Lys) 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 into training, validation, and test sets with the following dimensions:

- **Training Set:** 3,683,888 Peptide-Spectrum Matches (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.

3.1.2 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**].

Technical Characteristics The instrumentation and fragmentation settings (Orbitrap HCD) are consistent with the TMT-labeled dataset to ensure technical compatibility. This set includes a wide variety of post-translational modifications and biological sequences.

Dataset Statistics The non-TMT reference data is partitioned into training, validation, and test sets with the following dimensions:

- **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.

4 Methods

4.1 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.

Figure 4.1: Schematic representation of the adapted transformer architecture. The TMT status is fed as a covariate embedding into the decoder alongside the spectral features, as adapted from the Modanovo framework.

4.1.1 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**].

4.2 Data Selection and Training Protocol

4.2.1 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**].

4.2.2 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 to maintain data density.
- **Distribution Analysis:** The dataset was screened for biases in peptide length and charge state distributions. Ensuring a representative spread across these parameters is crucial for the generalization of transformer-based models [**KlaprothAndrade2025**].
- **Bias Mitigation:** To prevent the model from overfitting to hyper-abundant peptides, a threshold of 229 Peptide-Spectrum Matches (PSMs) per unique peptide sequence was enforced.
- **Data Leakage Prevention:** Following the rigorous validation protocols of *ModaNovo*, 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 that the model learns the chemical principles of modifications rather than memorizing specific 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.

4.2.3 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 1×10^{-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.

4.3 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.

4.3.1 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.

4.3.2 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:

$$P(t) = \frac{|\text{Correct PSMs with score} \geq t|}{|\text{Total predictions with score} \geq t|} \tag{4.1}$$

$$C(t) = \frac{|\text{Predictions with score} \geq t|}{|\text{Total ground truth identifications}|} \tag{4.2}$$

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**].

4.3.3 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.

4.4 Biological Validation

4.4.1 Peptide Alignment

4.4.2 Genomic Evidence

4.4.3 Spectral Quality

5 Results

6 Discussion

Supplementary Material

Dieses Kapitel enthält zusätzliches Material zur Thesis.

6.1 Zusätzliche Abbildungen

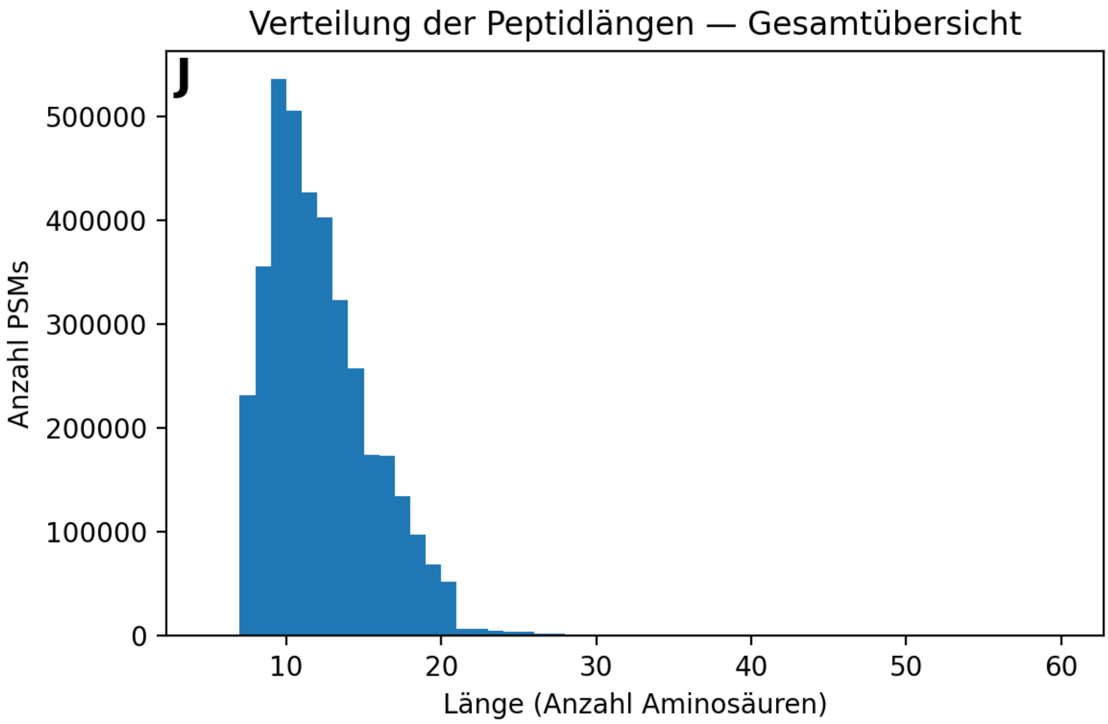

Figure S1: Histogram of peptide lengths for all data

6.2 Zusätzliche Tabellen

Weitere Tabellen...

Dies ist ein Mock-Beispiel.

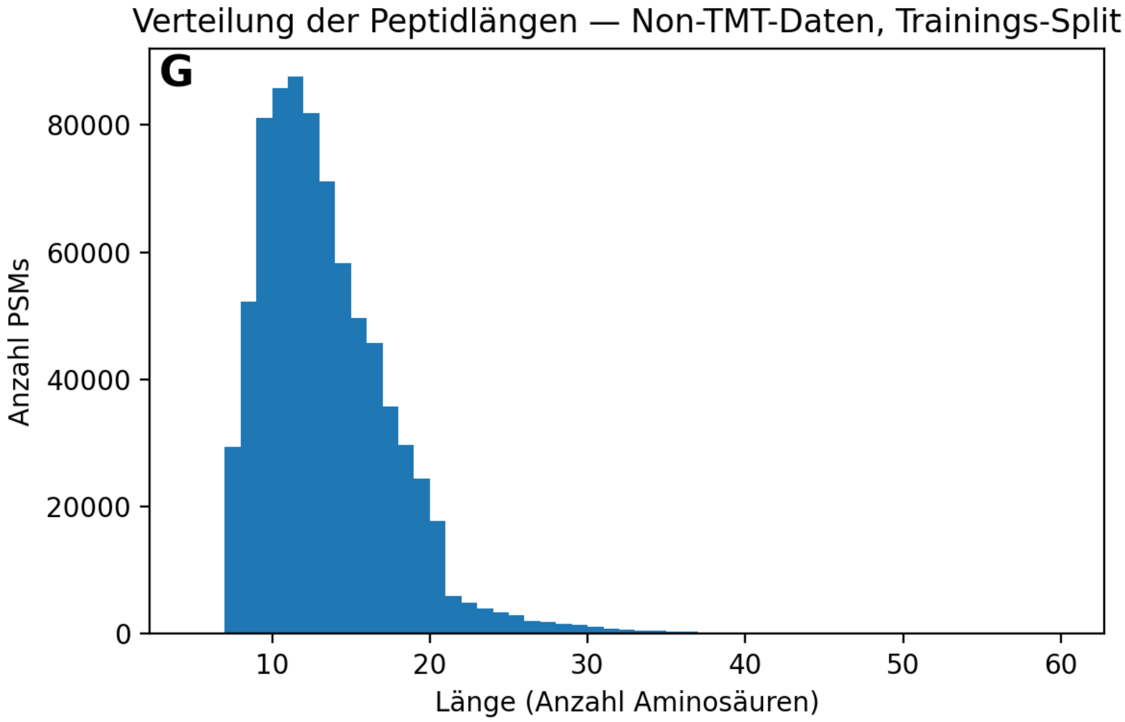

Figure S2: Histogram of peptide lengths for training data (TMT false)

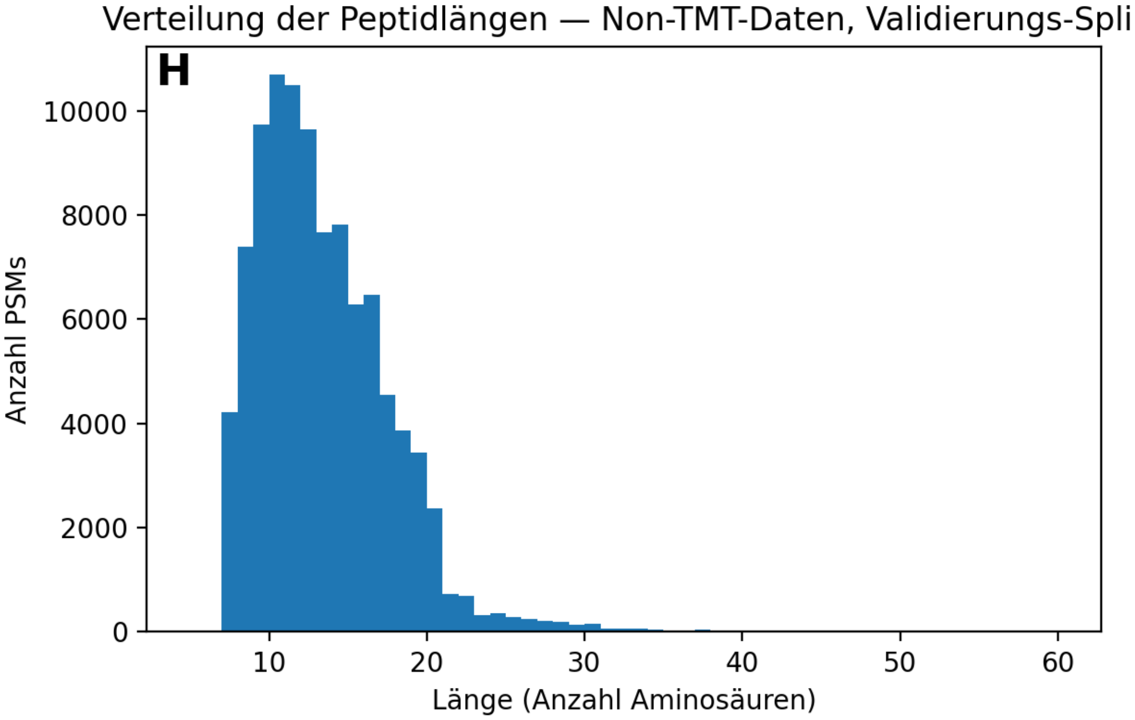

Figure S3: Histogram of peptide lengths for validation data (TMT false)