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

Dieses Kapitel führt in die Thematik der Bachelor-Thesis ein.

1.1 Motivation

Beschreiben Sie hier die Motivation für Ihre Arbeit. Warum ist das Thema relevant? Welche Probleme werden adressiert?

1.2 Ziel der Arbeit

Formulieren Sie die Ziele Ihrer Thesis klar und präzise.

1.3 Aufbau der Arbeit

Geben Sie einen kurzen Überblick über die Struktur der Thesis.
Dies ist ein Mock-Beispiel. Passen Sie den Inhalt an Ihre Einführung an.

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 [4]. 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 [1].

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 [10]. 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

[9]. 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. 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 [2]. 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) [3]. 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 [7].

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 [11]. 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 [16]. Because de novo sequencing does not de-

pend 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 [12]. 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. This significantly reduces instrument time and eliminates the missing value problem often encountered in label-free quantification [15].

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 [6]. 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 [5].

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 [7].

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 [13]. 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 [14]. 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 [16]. 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 previously predicted residues and the encoded spectrum.

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 [**Qiao21**]. This prevents the model from being trapped by a single high-probability amino acid that might lead to an invalid total mass, ensuring that the final sequence is globally consistent with the precursor mass.

3 Datasets

3.1 Datasets

The performance and generalizability of deep learning models in proteomics are fundamentally determined by the quality and composition of the underlying training data. For this project, a curated dataset was designed to bridge the gap between unlabeled high-resolution spectra and Tandem Mass Tag (TMT) multiplexed data, while maintaining the model's ability to identify diverse post-translational modifications (PTMs).

3.1.1 Fine-Tuning Data

The fine-tuning strategy aims to adapt a pre-trained transformer model to the specific fragmentation patterns and mass shifts introduced by TMT labeling. To ensure a robust transition without losing prior knowledge, the dataset follows a specific composition and rigorous filtering pipeline.

Data Selection and Preprocessing

The raw data, initially available in .mzML and .csv formats, was converted into the Mascot Generic Format (.mgf) to facilitate efficient processing. The training corpus was constructed using an 80/20 balance between TMT-labeled and unlabeled spectra. This ratio was chosen to provide sufficient exposure to the TMT-specific reporter ions and modified N-termini/lysines while retaining the general features of peptide fragmentation learned during pre-training.

To ensure unbiased evaluation, the data was divided into 80/10/10 splits (training, validation, and testing). Several measures were implemented to address common pitfalls in proteomics machine learning:

- **Addressing Peptide Imbalance:** To prevent the model from overfitting to highly abundant "housekeeping" proteins, a maximum of 229 Peptide-Spectrum Matches (PSMs) per peptide sequence was enforced. This sampling strategy ensures that the model learns sequence-independent fragmentation features rather than memorizing frequent sequences.

- **Data Leakage Prevention:** Special care was taken to ensure that modified versions of the same peptide (e.g., a phosphorylated vs. a non-modified version) were assigned to the same split. This prevents the model from "cheating" by recognizing the backbone sequence in the validation set that it had already seen in the training set.
- **Quality Filtering:** Spectra were subjected to quality control where empty peaks were removed, and the distributions of peptide length and charge states were monitored. This step ensures that no systematic bias (e.g., only short peptides being TMT-labeled) is introduced into the model [**Yue2024**].

TMT-labeled Dataset

The primary source for TMT-specific training is the Multi-PTM PROSPECT dataset. This dataset is particularly valuable as it contains a high density of annotated PTMs and TMT-labeled spectra, providing the necessary complexity for modern proteomics workflows [**Yue2024**].

The dataset comprises:

- **Training Set:** 3,683,888 PSMs
- **Validation Set:** 462,008 PSMs

During preprocessing, all non-TMT spectra originating from this source were excluded to maintain the integrity of the TMT-specific training phase. Furthermore, the peptide annotations were harmonized by translating Unimod syntax into a mass shift-based syntax. This allows the transformer model to treat modifications as continuous mass offsets rather than discrete categorical labels, which is crucial for identifying rare or novel PTMs.

Non-TMT Reference Data (Replay Set)

To prevent "catastrophic forgetting"—a phenomenon where a model loses its ability to perform original tasks after being fine-tuned on new data—a "Replay Set" of non-TMT data was integrated. This set consists of a highly diverse subset of the MoDaNovo dataset and the Massive Knowledge Base (Massive-KB) [**Wang2022**].

The composition of this reference set includes:

- **Source:** 80% Multi-PTM (non-labeled subset) and 20% Massive-KB.
- **Volume:** 784,128 PSMs corresponding to 289,568 unique peptides.

- **Split:** The training subset contains 98,396 peptides, while the validation subset contains 23,004 peptides.

By mixing these high-confidence reference spectra into the fine-tuning process, the model retains its baseline accuracy for standard peptide identification while successfully learning the nuances of TMT-labeled fragmentation.

4 Methods

5 Results

6 Discussion

References

[1] R. Aebersold and M. Mann. "Mass-spectrometry-based proteomics." In: *Nature* 537.7620 (2016), pp. 347–355. DOI: 10.1038/nature19949.

[2] J. Cox and M. Mann. "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification." In: *Nature Biotechnology* 26.12 (2008), pp. 1367–1372. DOI: 10.1038/nbt.1511.

[3] J. K. Eng, A. L. McCormack, and J. R. Yates. "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database." In: *Journal of the American Society for Mass Spectrometry* 5.11 (1994), pp. 976–989. DOI: 10.1016/1044-0305(94)80016-2.

[4] S. A. Gevaert and J. Vandekerckhove. "Protein identification methods in proteomics." In: *Protein Science* 12.9 (2003), pp. 1913–1925. DOI: 10.1110/ps.0309203.

[5] J. P. Hogrebe, C. F. von Hagel, et al. "The impact of TMT labeling on the fragmentation of peptides." In: *Journal of Proteomics* 175 (2018), pp. 130–139. DOI: 10.1016/j.jprot.2017.11.012.

[6] G. C. McAlister, D. P. Nusinow, et al. "MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell lines." In: *Analytical Chemistry* 86.14 (2014), pp. 7150–7158. DOI: 10.1021/ac502040v.

[7] A. I. Nesvizhskii. "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics." In: *Journal of Proteomics* 73.11 (2010), pp. 2092–2123. DOI: 10.1016/j.jprot.2010.08.009.

[8] R. Qiao et al. *Computationally efficient de novo peptide sequencing via a transformer-based model*. arXiv:2104.14501. 2021.

[9] R. Roepstorff. "All beginnings are easy: The early steps of peptide fragmentation." In: *Journal of the American Society for Mass Spectrometry* 21.7 (2010), pp. 1085–1090. DOI: 10.1016/j.jasms.2010.04.001.

[10] H. Steen and M. Mann. "The ABC's (and XYZ's) of peptide sequencing." In: *Nature Reviews Molecular Cell Biology* 5.9 (2004), pp. 699–711. DOI: 10.1038/nrm1468.

[11] J. A. Taylor and R. S. Johnson. "Sequence databases: a new iterative method for predicting amino acid sequences from tandem mass spectra." In: *Rapid Communications in Mass Spectrometry* 11.10 (1997), pp. 1067–1075. DOI: 10.1002/(SICI)1097-0231(19970630)11:10<1067::AID-RCM946>3.0.CO;2-9.

[12] A. Thompson, J. Schäfer, et al. "Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS." In: *Analytical Chemistry* 75.8 (2003), pp. 1895–1904. DOI: 10.1021/ac0262560.

[13] N. H. Tran, R. Qiao, et al. "Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry." In: *Nature Methods* 16.1 (2019), pp. 63–66. DOI: 10.1038/s41592-018-0260-3.

[14] A. Vaswani, N. Shazeer, et al. "Attention is all you need." In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[15] T. Werner, I. Becher, et al. "High-resolution sampled 10-plex TMT for proteomics." In: *Analytical Chemistry* 86.14 (2014), pp. 7025–7031. DOI: 10.1021/ac501510y.

[16] M. Yilmaz, W. Fondrie, et al. "De novo peptide sequencing with deep learning." In: *Nature Methods* 20.2 (2023), pp. 276–282. DOI: 10.1038/s41592-022-01712-9.

Supplementary Material

Dieses Kapitel enthält zusätzliches Material zur Thesis.

6.1 Zusätzliche Abbildungen

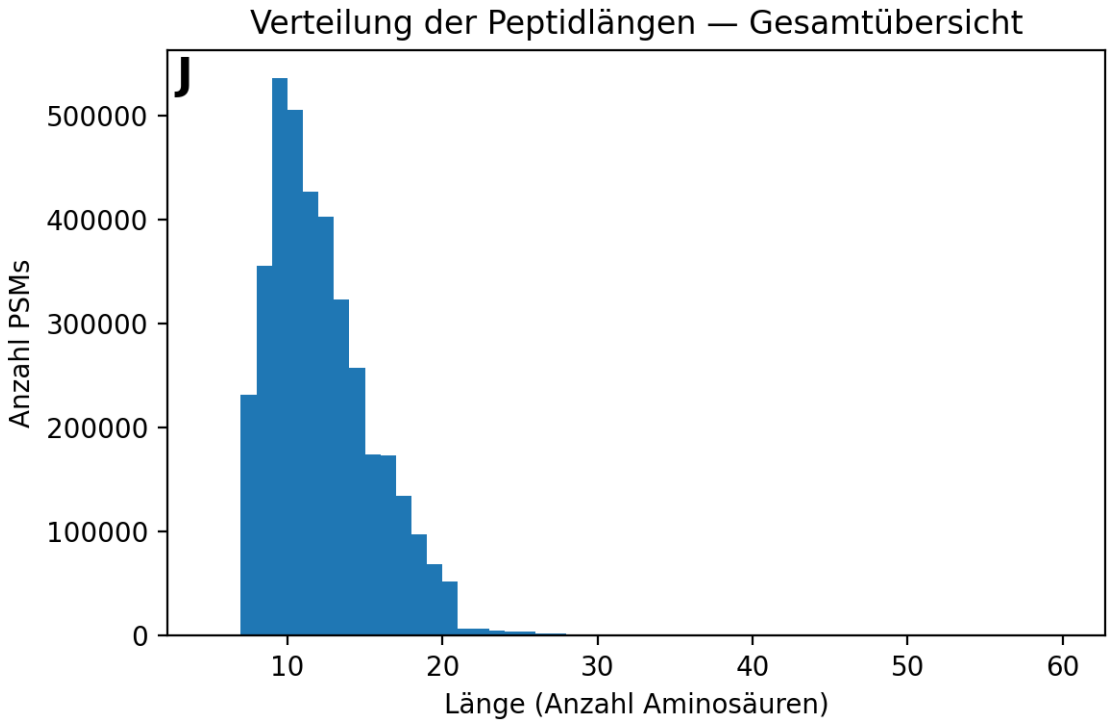

Figure S1: Histogram of peptide lengths for all data

6.2 Zusätzliche Tabellen

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

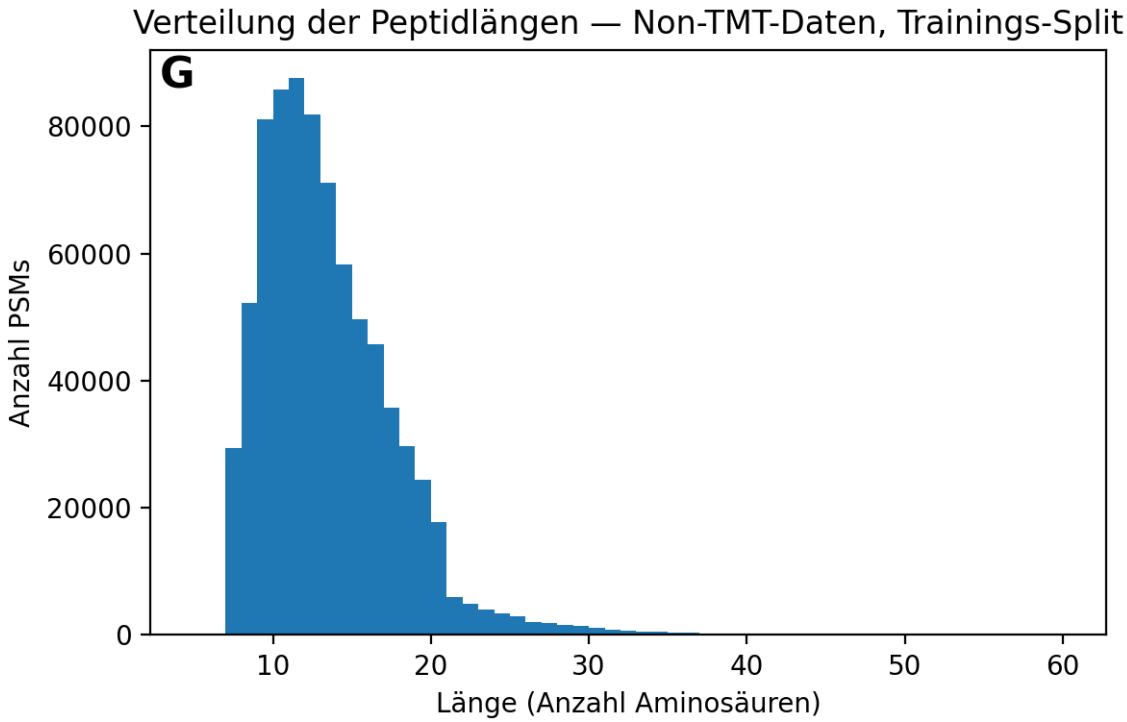

Figure S2: Histogram of peptide lengths for training data (TMT false)

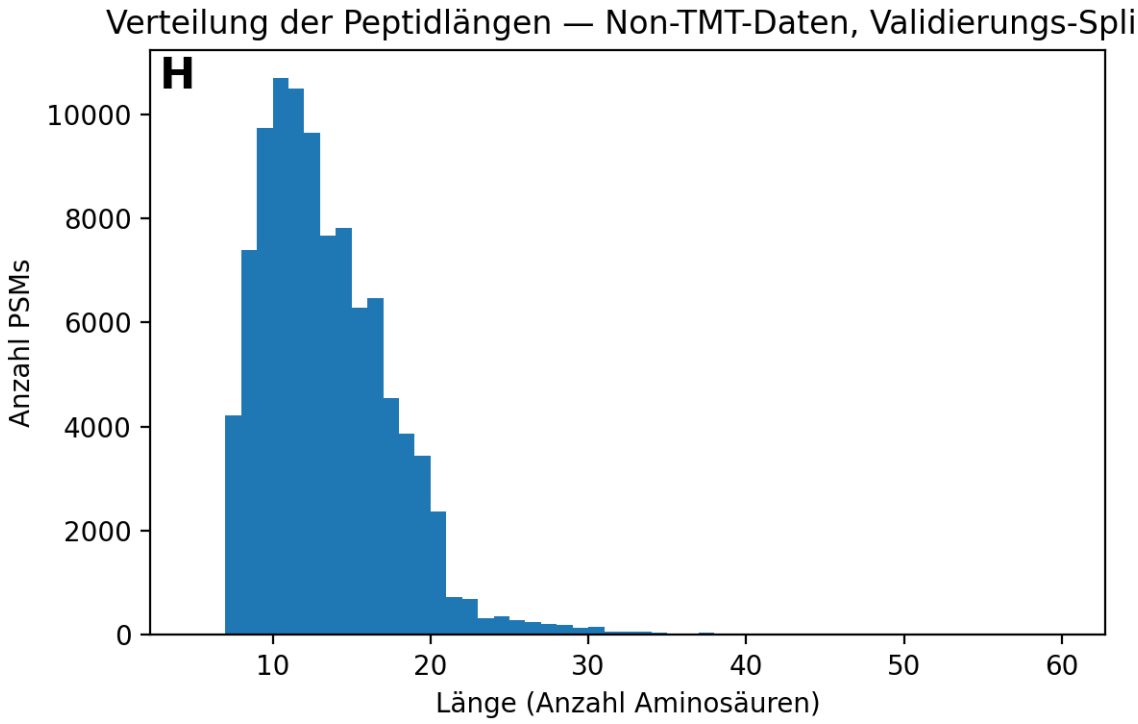

Figure S3: Histogram of peptide lengths for validation data (TMT false)