

## **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

<b>Abstract</b>	<b>i</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Ziel der Arbeit . . . . .	1
1.3 Aufbau der Arbeit . . . . .	1
<b>2 Background</b>	<b>3</b>
2.1 Mass Spectrometry-Based Proteomics . . . . .	3
2.1.1 Bottom-Up Proteomics Workflow . . . . .	3
2.1.2 Tandem Mass Spectrometry (MS/MS) and Peptide Fragment Ion Theory . . . . .	3
2.2 Peptide Identification Strategies . . . . .	4
2.2.1 Database Search Engines (DBIS) . . . . .	4
2.2.2 Principles of De Novo Peptide Sequencing . . . . .	4
2.3 Tandem Mass Tag (TMT) Labeling . . . . .	4
2.3.1 Isobaric Labeling Chemistry . . . . .	4
2.3.2 Impact on Fragmentation Patterns . . . . .	4
2.4 Deep Learning and Transformer Models . . . . .	4
<b>3 Datasets</b>	<b>7</b>
3.1 Überblick über die Datensätze . . . . .	7
3.2 Datenquellen . . . . .	7
3.3 Datenverarbeitung . . . . .	7
3.4 Statistische Eigenschaften . . . . .	7
<b>4 Materials and Methods</b>	<b>9</b>
<b>5 Results and Discussion</b>	<b>11</b>
<b>6 Discussion</b>	<b>13</b>
<b>References</b>	<b>15</b>



# **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.

## *1 Introduction*

---

## 2 Background

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

### 2.1 Mass Spectrometry-Based Proteomics

#### 2.1.1 Bottom-Up Proteomics Workflow

Mass spectrometry (MS)-based proteomics is the gold standard for large-scale protein analysis. The “bottom-up” approach is the most widely adopted strategy, where proteins are extracted and enzymatically digested—typically using trypsin—into smaller peptides before analysis [4]. This is essential as peptides are more easily fractionated and ionized than intact proteins. The resulting mixture is separated via liquid chromatography (LC) and ionized using Electrospray Ionization (ESI) [1].

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

Peptide sequences are identified using Tandem Mass Spectrometry (MS/MS). A precursor ion is isolated by its mass-to-charge ratio ( $m/z$ ) and fragmented, often using Higher-energy Collisional Dissociation (HCD) [10].

According to fragment ion theory, the peptide backbone fragments primarily at amide bonds, resulting in *b*-ions (N-terminal) and *y*-ions (C-terminal) [9]. The mass difference between consecutive ions in a series corresponds to specific amino acids. However, post-translational modifications (PTMs) or chemical labels like Tandem Mass Tags (TMT) shift these masses, requiring advanced computational identification.

## 2.2 Peptide Identification Strategies

### 2.2.1 Database Search Engines (DBIS)

The most common identification method is database searching. Engines like MaxQuant or SEQUEST compare experimental MS/MS spectra against *in silico* digested sequences from databases like UniProt [2, 3]. While robust, DBIS is limited by the “search space” problem: it cannot identify modifications not explicitly included in the database, leading to missed novel PTMs [7].

### 2.2.2 Principles of De Novo Peptide Sequencing

In contrast, *de novo* sequencing reconstructs sequences directly from fragment ion peaks without a reference database [11]. While historically limited by noise, modern Transformer-based models now capture long-range dependencies between ions, making this approach ideal for discovering novel PTMs and variants in the “dark proteome” [13].

## 2.3 Tandem Mass Tag (TMT) Labeling

### 2.3.1 Isobaric Labeling Chemistry

Tandem Mass Tag (TMT) labeling is used for high-throughput multiplexed quantification. TMT tags are isobaric, meaning labeled peptides from different samples appear as a single peak in MS1 scans, reducing instrument time and missing values [12, 15].

### 2.3.2 Impact on Fragmentation Patterns

TMT tags introduce systematic mass shifts to the N-terminus and lysine side chains. Upon fragmentation, they release reporter ions ( $m/z$  126–135) for quantification [6]. For *de novo* sequencing, these tags are challenging because they alter fragmentation efficiency and shift *b*- and *y*-ion series significantly [5].

## 2.4 Deep Learning and Transformer Models

The identification of peptides is a sequence-to-sequence (Seq2Seq) task. While early models used LSTMs [13], the Transformer architecture revolutionized the field with the Self-Attention mechanism [14].

In a proteomic context, the Transformer’s encoder extracts structural features from continuous  $m/z$  and intensity values through point-based encoding [16]. The decoder then uses Beam Search to maintain a set of the  $k$  most likely sequences, ensuring the final result is globally consistent with the precursor mass [8].

## *2 Background*

---

# **3 Datasets**

Dieses Kapitel beschreibt die verwendeten Datensätze in der vorliegenden Arbeit.

## **3.1 Überblick über die Datensätze**

Hier können Sie einen Überblick über die Datensätze geben, die in Ihrer Bachelor-Thesis verwendet werden. Zum Beispiel:

- Datensatz 1: Beschreibung, Quelle, Größe, etc.
- Datensatz 2: Beschreibung, Quelle, Größe, etc.

## **3.2 Datenquellen**

Beschreiben Sie, woher die Daten stammen. Zum Beispiel:

- Öffentliche Datenbanken wie NCBI, Ensembl, etc.
- Eigene Experimente oder Simulationen.

## **3.3 Datenverarbeitung**

Erklären Sie, wie die Daten vorverarbeitet wurden:

- Filterung, Normalisierung, etc.
- Tools oder Skripte, die verwendet wurden.

## **3.4 Statistische Eigenschaften**

Fügen Sie Tabellen oder Abbildungen hinzu, die die Eigenschaften der Datensätze zeigen.

Dies ist ein Mock-Beispiel. Passen Sie den Inhalt an Ihre tatsächlichen Datensätze an.

### *3 Datasets*

---

Table 3.1: Übersicht der Datensätze

Datensatz	Anzahl Einträge	Quelle
Datensatz A	1000	NCBI
Datensatz B	500	Eigenes Experiment

## **4 Materials and Methods**

#### *4 Materials and Methods*

---

## **5 Results and Discussion**

## *5 Results and Discussion*

---

## **6 Discussion**

## *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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1016/j.jasms.2010.04.001).

## References

---

- [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**

Hier können zusätzliche Abbildungen oder Daten eingefügt werden.

## **6.2 Zusätzliche Tabellen**

Weitere Tabellen...

Dies ist ein Mock-Beispiel.