university of groningen

university college groningen

Proof of Concept for Rule-based Automated De Novo Tandem Peptide Spectrum Interpretation

David Sebastian Bebensee[1], Marcello Goulart[1], Kristoffer Basse[2], Peter Horvatovitch[2], and Volker Nannen[1]

[1]University College Groningen
[2]Department of Analytical Chemistry, University of Groningen

July 2025

Abstract

Abstract

De Novo sequencing of peptides is still essential for identifying novel peptides using mass spectrometry. Advances in this field can significantly improve cost-effectiveness and speed of disease diagnostics, while also strengthening our ability to combat drug threats. This study explored the development of an automated de novo interpretation program based on a rule-based expert system. The system constructs a distance matrix from peak-to-peak distances in mass spectra to identify novel peptides. The work is intended to be incorporated into a larger AI-driven system that uses large language models to enhance the annotation and identification of peptide spectra.

The program interpreted 487 tandem mass spectra and attempted to reconstruct their peptide sequence. It fully identified 2.25% of the peptides, and achieved numerous partial matches. The approach demonstrates potential for substantial improvement. With further development and implementation of suggested enhancements, this system could increase its accuracy. Additionally, this approach has the potential to provide novel heuristics in the field of mass spectrometry.

Acknowledgments

I thank Volker Nannen for his guidance and support, and for helping through the force majeure issues that arose during my work. I thank Peter Horvatovitch and Kristoffer Basse for their continuous availability and support during the development process. David Tabb deserves thanks for his availability and suggestions which have greatly improved my final thesis. I thank Marcello Goulart for co-assessing my thesis. I also thank Victor Guistini Perez and Julia St.Germain for helping me with the Introduction and Background. I thank Gina Pinas for assisting with organization and accountability while I was part of the Thesis support group. I thank Adriana Mattos for her assistance through my forced topic change. Finally, I thank Mr. Nicholas Wedd for allowing me to use his graphics.

Contents

Abbreviations

BUP bottom-up proteomics, a field within proteomics where proteins are digested into peptides before analysis. 2

Da dalton, unit of mass. 9, 10, 13, 30

DDA data dependent approach. 3, 5, 9, 11

DIA data independent approach. 3, 5

ESI electrospray ionization. 1

HCD higher energy collisional dissociation. 2, 4, 9

IRS intensity rank sum. 14

LC liquid chromatography. 2

m/z mass-to-charge ratio. 1, 5, 8–14, 17, 25, 27, 28, 30

MS mass spectrometry, analytical technique to measure mass to charge ratio of molecules. 1

MS/MS also called tandem mass spectrum or product ion spectrum, a mass spectrum of a fragmented molecule. 3–6, 8–13, 17–19, 21, 22, 26–28, 30

MS1 also called a precurser ion spectrum, a mass spectrum generated without fragmentation. 2, 3, 5, 10, 11

PTM post translational modifications. 2

SSE sum squared error. 14, 15, 27

TDP top-down proteomics, a field within proteomics where proteins get analysed without being digested. 2

Chapter 1

Introduction

1.1 What is mass spectrometry

Mass spectrometry (MS) is an analytical technique used to measure the mass-to-charge ratio of molecules. It enables accurate identification of peptides and proteins in complex mixtures and is a cornerstone of proteomics, a field of study which investigates proteins, their composition, and functions in biological systems (Sinha & Mann, 2020). In healthcare, MS provides value through precise diagnostics and screening, from the detection of metabolic disorders in newborns to the identification of biomarkers for diabetes, Alzheimer's disease, or Parkinson's disease (Jannetto & Fitzgerald, 2016; Loponte et al., 2025). It also holds promise of being an early predictive measure of many other diseases, including cancers, providing a cost-effective and time-efficient alternative to traditional screening methods (Trifonova et al., 2021; Tully et al., 2019; Urbiola-Salvador et al., 2024). Beyond diagnostics, the uses of MS also extend to the detection of novel opioids or analogs to drugs such as fentanyl, enabling faster regulatory responses to emerging drug threats (Armenian et al., 2018) .

1.2 The basic workflow of mass spectrometry instruments

Although varied, mass spectrometry instruments generally follow the same logic and contain the same three components: an analyzer, an ion source, and a detector (Al-Amrani et al., 2021; Sinha & Mann, 2020). A sample of molecules fed into the machine is ionized, using techniques such as electrospray ionization (ESI), and these ions are separated by their mass-to-charge (m/z) ratio. The abundance of ions at the same m/z is then detected and represented as an intensity

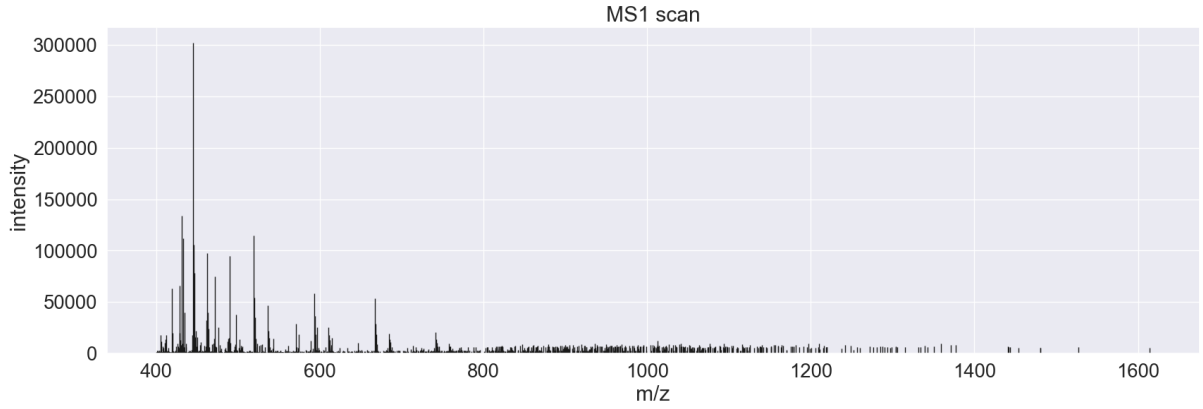


Figure 1.1: MS1 scan of m/z ratios observed at a single retention time. Subset of the MRC-5a dataset. Every intensity peak represents a unfragmented peptide ion

on a mass spectrum (Al-Amrani et al., 2021; Sinha & Mann, 2020). Figure 1.1 depicts a mass spectrum called MS1 or precursor ion spectrum.

In the case of proteomics, the primary molecules that are being analyzed are peptides and proteins. A peptide is a short sequence of up to 50 amino acids. Where each amino acid can be represented by a letter (Frank & Pevzner, 2005). A sequence of peptides strung together makes a protein. There exist two distinct approaches to analysis within proteomics. Top-down proteomics (TDP) feeds undigested proteins into the mass spectrometer. This allows for interpretation of the full protein when analysing data and enables the possibility of locating post-translational modifications (PTM). Bottom-up proteomics (BUP) enzymatically digests a protein or several proteins into smaller peptides, generally using trypsin, before using the mass spectrometer. This more common approach allows for more detailed data and has more established pipelines (Toby et al., 2016; Wehr, 2006). In most cases, both of these approaches add an additional dimension of analysis before processing by employing methods such as liquid chromatography (LC), which make use of molecular characteristics like hydrophobicity to further differentiate the peptides and proteins (Sinha & Mann, 2020). Visible in figure 1.2 is a representation of a MS1 scan using OpenMS TOPPView software, it includes the additional dimension of retention time obtained through LC from unfragmented peptides (Rost et al., 2016; Sturm & Kohlbacher, 2009).

However, in proteomics, it is typical that these peptide ions are fragmented to learn more about their composition. This is achieved through higher-energy collisional dissociation (HCD) or other methods, turning peptide ions into fragment ions. These fragment ions can be reinserted into the mass spectrometer to return a mass spectrum called MS2, product, tandem mass spectra

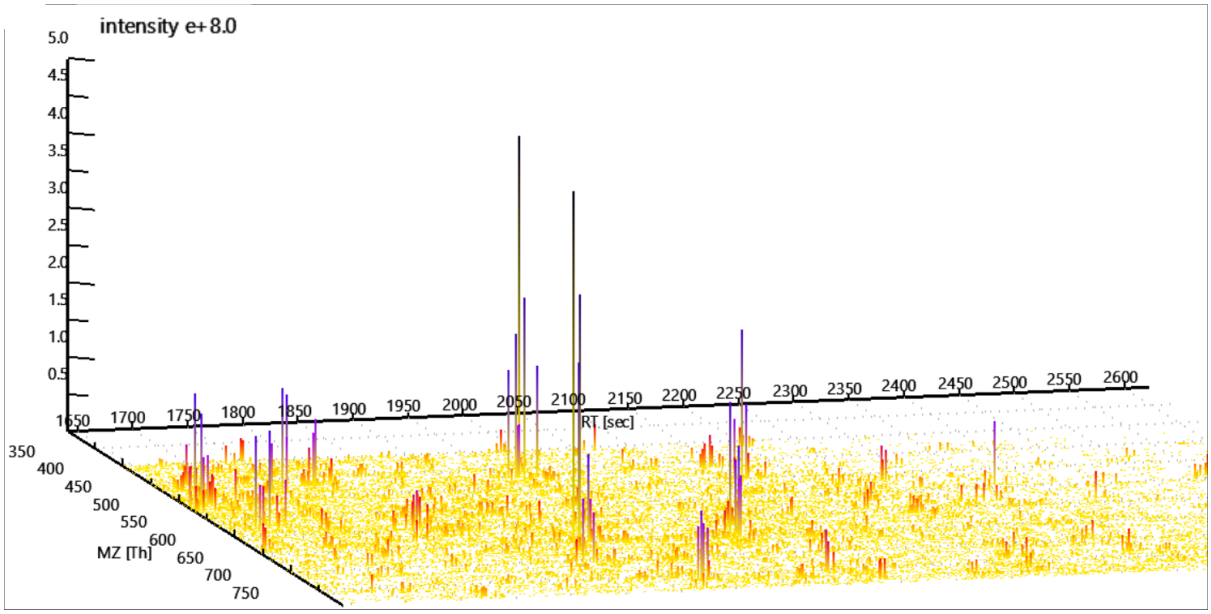


Figure 1.2: MS1 scan. A 3D view of the MRC-5-a dataset showing retention time (RT), mass-to-charge ratio (MZ) and intensity of the observed peptides in the sample.

or MS/MS (Figure 1.3) (Sinha & Mann, 2020). Although a peptide can fragment at any position along its chain, there are three possible fragment ion categories, as seen in Figure 1.4. The most common fragmentation occurs at the amide bond, resulting in a y-ion and a b-ion, where y-ions are observed more abundantly than b-ions (Paizs & Suhai, 2005; Shao et al., 2014). The y-ions are the fragment connected to the C-terminus of the peptide, and the b-ions are the fragment connected to the N-terminus. There are two ways to follow through on the second run through the mass spectrometer. A data-independent approach (DIA) fragments all peptide ions in a certain range together, while a data-dependent approach (DDA) chooses the most intense peaks and fragments them individually (Sinha & Mann, 2020). Figure 1.5 visualizes the different approaches. A DIA creates comprehensive but very complex datasets that require sophisticated deconvolution to handle, while a DDA can miss low-abundance ions but produce cleaner, more easily interpretable MS/MS. These MS/MS spectra can be seen as fingerprints for a certain peptide since, if it contains all possible y-ions, its amino acid sequence can be accurately inferred from the distances between its peaks. This process of identifying unknown peptides from a MS/MS is called *de novo* identification or *de novo* sequencing (Wedd, n.d.). Another common way to identify which peptide is represented by a MS/MS is to use database matches. Software such as MSFragger, or SEQUEST can simulate fragmentation and generate expected MS/MS for a given peptide and match them with the observed MS/MS to find a peptide spectrum match (Kong et al., 2017; Tabb, 2015).

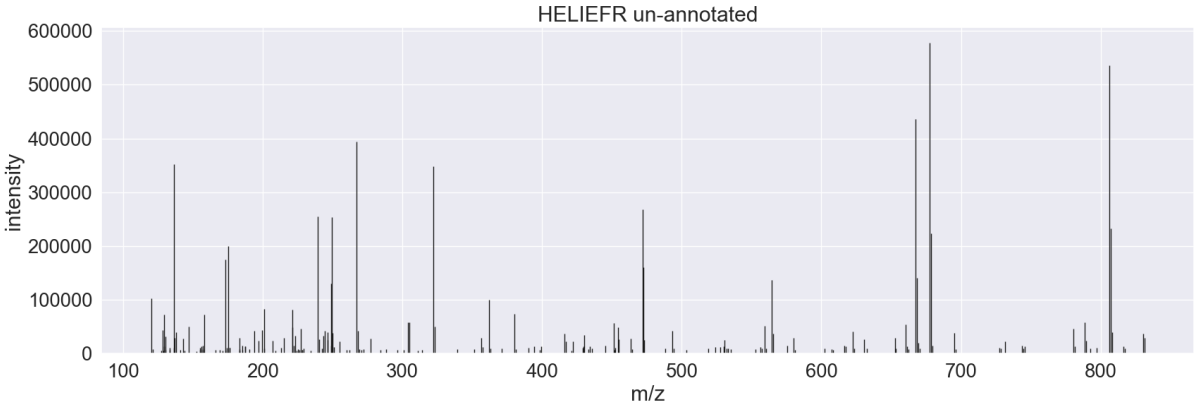


Figure 1.3: Example of an un-annotated MS/MSrepresenting the peptide *HELIEFR*. This raw data is enough for *de novo* sequencing. The y-axis depicts intensity, indicating the abundance at which a certain molecule, which is assumed to be a fragmented peptide ion, appears. The X-axis depicts the mass-to-charge (m/z) of the molecule.

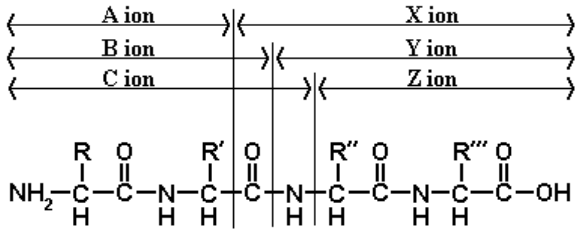


Figure 1.4: Peptide Backbone and its possible fragment points. Most commonly, when HCD is used, the peptide breaks into B/Y Ions. It rarely breaks into A/X or C/Z Ions. The Rṫs represent different amino acids held together by the backbone and bonded together. The bonds seen here repeat between each amino acid in a peptide sequence. Figure reproduced from (Wedd, n.d.) with permission from the Author.

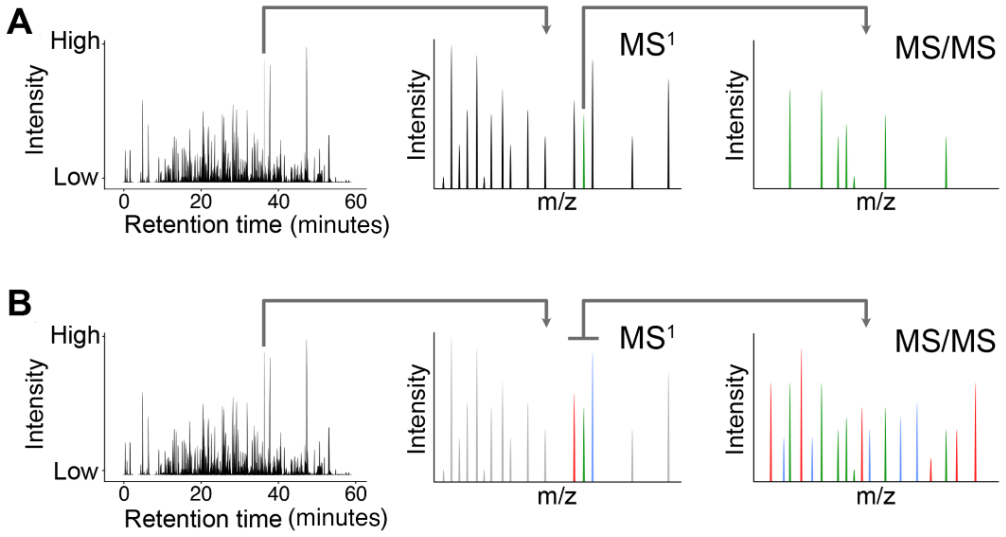


Figure 1.5: "DDA and DIA are the common data acquisition strategies in shotgun proteomics. (A) In DDA, a peptide ion is selected from many ions available in the MS1 data at a given retention (chromatography) time. The peptide ion is fragmented, and the data are recorded as MS/MS. (B) In DIA, multiple peptide ions are selected based on m/z window at a given retention (chromatography) time. The peptide ions are fragmented, and they inherently produce a complex MS/MSspectra." Figure reproduced from (Sinha & Mann, 2020) under the Creative Commons Attribution License.

1.3 Problem Statement

However, while MS obtains large amounts of data and there are possibilities to interpret it, the process is often challenging. When a MS/MS is being generated, it rarely perfectly represents a single peptide but, even in a DDA, often contains peaks from other peptides, instrument noise, and heavier or lighter variations of ions. These variations can be due to isotopes as well as the loss of a water or ammonia molecule, and contaminate the original spectrum(Frank & Pevzner, 2005; Wedd, n.d.). Additionally, peaks generated by real y-ions can be at such low intensity that they are virtually indistinguishable from noise and cannot be measured at all in some cases. This is because some bonds between amino acids are stronger and less likely to break than others. Therefore, accurate identification of peptides from MS/MS remains a bottleneck in MS (Liu et al., 2023; Ma, 2015). Traditionally, a database search is used when attempting to identify a Peptide sequence in a new dataset, using the aforementioned software. This works well since databases for identified peptides are constantly growing and expanding. However, these databases cannot encompass all the peptides and proteins. They often lack sequences for specific species, do not account for taxonomic diversity, and are unable

to handle unanticipated modifications or sequence variants that can occur in tumors or other mutations. For these reasons, *de novo* sequencing of new unidentified peptides is still widely relied upon for unmatched peptides(Liu et al., 2023). While approaches to manual *de novo* sequencing exist, interpreting the large datasets provided by MS would require a lot of time and be highly inefficient (Wedd, n.d.). Subsequently, there is a market for software that can identify large amounts of MS/MS through the *de novo* methods (Urban et al., 2024), such as PepNet, PointNovo, and DeepNovo (Liu et al., 2023; Qiao et al., 2021; Tran et al., 2017). This type of software still sees rapid development and innovation expanding into using neural networks, deep learning, and other AI fields to solve the problem of effective *de novo* interpretation.

1.4 Hypothesis and Goals

This study is an exploratory approach to *de novo* sequencing software and aims to build a baseline *de novo* algorithm following the manual *de novo* instructions presented by Mr. Nicholas Wedd (Wedd, n.d.). It intends to fit into a brader research scope with the goal of eventually designing an AI system that integrates large language models to enhance MS/MS annotation utilizing expert knowledge. Such an AI system would have the advantage of not only being able to identify technically possible peptide sequences inferred from a MS/MS but also assess their biochemical validity, considering metadata such as patient history and sample preparation methods. This would increase time efficiency in peptide analysis and potentially increase accuracy compared to current methods. The final AI model would optimally provide more data than just a peptide spectrum match, but rather a readable annotation that can be interpreted and corrected by experts. This study integrates into that framework by providing a non-black box approach to *de novo* sequencing, where the rules and process are understandable by humans. The goal in this study is to set the groundwork for this final AI model by developing an expert system that can identify peptide sequences from a given MS/MS. The expert system is to be built in a manner such that it is able to be expanded upon and can ultimately supply adequate training data for the final AI model.

1.5 Research Question

How effectively can a rule-based expert system for *de novo* sequencing, using a distance-matrix approach, identify novel peptide sequences from tandem mass spectra, and to what extent can it contribute to the development of an AI-driven system for enhancing spectral annotation?

Chapter 2

Methods and Theory

This section will describe the attempt to build an expert system based *de novo* sequencing engine. The approach described by Nick Wedd will be followed as a general guideline (Wedd, n.d.). This was provided to me by my supervisor as an instruction to follow. Wedd's description of the *de novo* approach lays the groundwork for this program but was adapted to fit into a more programmatically applicable format.

2.1 Manual Identification Process

Using the *de novo* interpretation approach of MS/MS, an amino acid sequence can be deduced from a MS/MS through understanding and identifying the different peaks as fragmented peptide masses. Each peak potentially represents the same peptide, just fragmented at a different position along its chain. I.e. the peptide *HELIEFR* can break into *HEL* and *IEFR*, *HE* and *LIEFR* , or any other combination. Following this example, *HEL* would be the b-ion and *IEFR* the y-ion. Y-ions are the ions whose weight is typically measured in MS/MS and through which the sequence is identified. Starting with the full *HELIEFR* peptide as the original mass, the next step is to move downwards over the m/z values to find the peak associated with *ELIEFR*, then *LIEFR*, then *IEFR* , etc.. Each time, the distance between these peaks is noted, as it represents the weight of the amino acid that is missing between them. This way, without knowing that the true peptide is *HELIEFR* it is possible to deduce it from its MS/MS. Figure 2.1 shows the real application of this. Problems occur when introducing instrument noise and foreign peaks. The true difficulty lies in being able to differentiate a true peak from untrue peaks. One measure for this is the intensity of the peak, while an intense peak is not guaranteed to be part

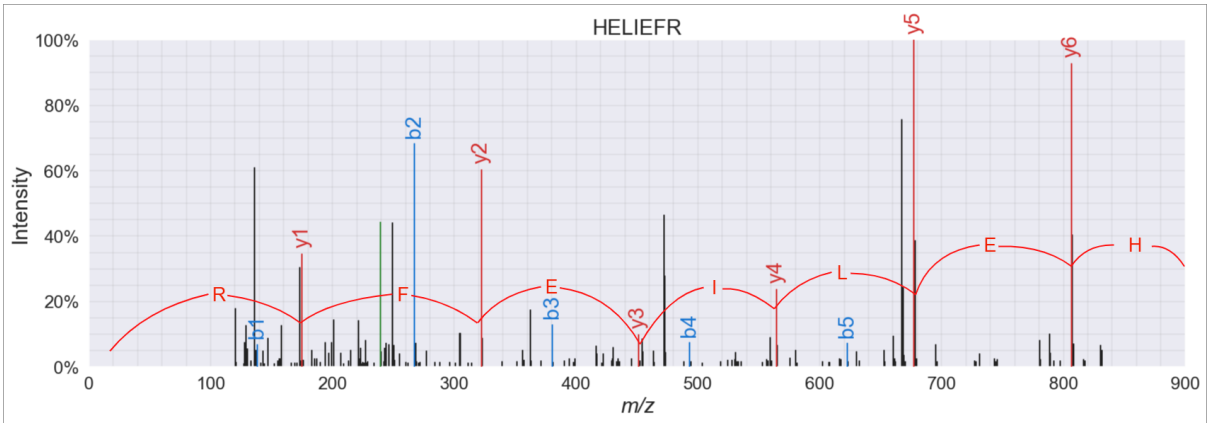


Figure 2.1: Example annotated MS/MS of *HELIEFR* used for analysis. Y-ions are marked in red, b-ions marked in blue, rarely occuring a-ions marked in green. The difference between each annotated y-ion peak represents the weight of the amino-acid that was lost between them. Thereby enabling the induction of the *HELIEFR* sequence

of the peptide, it is more likely to be real than a smaller peak. Another important measure is accuracy and knowing the weight of amino acids to their third or fourth decimal as well as the weight of different isotopes of amino acids to see whether the distance matches exactly. It is even possible that a peak is missing, meaning that none of the 22 amino acid weights match the distance to a peak within a certain threshold. Within manual *de novo* it is typical to look for the next peak whose distance matches the weight of a pair of amino-acids or potentially even a triplet of amino-acids. All these factors lead to some MS/MS being more easily interpretable than others.

Figure 2.2 shows the peptide *IQASGILQLFASLLTPQSSCK* and its MS/MS. The peptide has a length of 22 amino acids, but its MS/MS shows only 27 peaks and m/z values up to around 1550 Da. At an average amino-acid weight of 110 Da, it is expected for a matching MS/MS to have m/z values up to around 2300 m/z. It is an example of a MS/MS where it is unlikely that the right peptide sequence is inferred. Comparably *HELIEFR* (figure 2.1) has a length of 7 amino-acids and m/z values up to 831 and 195 peaks.

2.2 Data Acquisition and Preprocessing

The Dataset used to test this model was obtained from an MRC-5 fibroblast cell line found in lung tissue of a 14-week-old male fetus. The cell line was digested with trypsin and filtered using high-performance liquid chromatography. A DDA approach was used, and the peptide ions were fragmented in a Q Exactive Plus Orbitrap with HCD. This Orbitrap machine achieved a mass

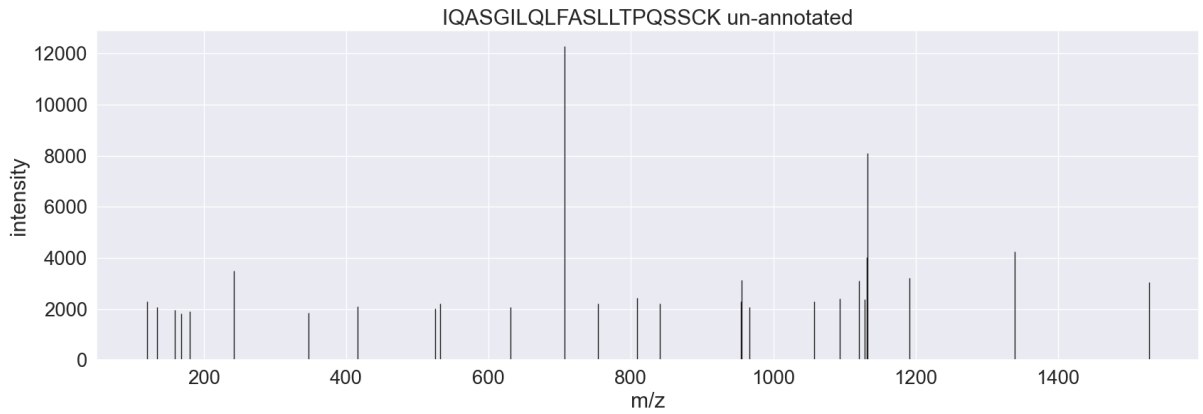


Figure 2.2: *IQASGILQLFASLLTPQSSCK* is a 22 amino acid peptide, but the MS/MS shows only 27 peaks and a max m/z value of around 1550 Da. The average amino-acid weight is 110 Da, meaning you would expect a max m/z value of around 2300 m/z for a full MS/MS. Making this is an example that would likely not be discovered easily using automated *de novo* methods

resolving power of 70000 for MS1 and 17500 for MS/MS, suggesting accuracy of measurement up to 0.002 Da in the final MS/MS. The output file received after running the data through the mass spectrometer is a *.raw* format that was converted into an *.mzML* file using ProteoWizard's MsConvert (Chambers et al., 2012). To improve ease of analysis, the Vendor algorithm for peak-picking was applied as a filter during the conversion process. The resulting *.mzML* file can then be read by a Python program using the pyteomics library. This *.mzML* file contains a subset of 30.000 mass spectra that were part of the MRC-5 analysis. Secondly, a *.csv* containing the likely peptide matches for 10.000 of the spectra, as well as metrics describing the accuracy of the matches, was generated using MSFragger and used to compare against the program's predictions (Kong et al., 2017).

2.3 Expert Systems

The program lends its fundamental design to the structure of an expert system. Consisting of a knowledge base and an inference engine. The knowledge base is built upon Wedd's description of *de novo* sequencing and uses its fundamental process to provide a rule system. The inference engine can then use these rules to make decisions in interpreting the given MS/MS(Tan et al., 2016).

2.4 Design of Programmatic Approach

After receiving a *.raw* file from the Orbitrap machine, MSConvert is used to convert the file type and also apply a low-aggression peak picking algorithm. This converted data is what was interpreted by the program. This large dataset was sorted and cleaned to obtain an array of intensities and an array of m/z values, which together build the MS/MS. Additionally, the precursor mass, a doubly-charged parent-ion, was calculated, which in a DDA is known since it was picked during the MS1 scan. Beyond this data and the mono-isotopic weight of the 22 proteinogenic amino acids, the program has no further outside data on which it can make inferences about the final peptide sequence. While a peak picking algorithm was already applied through MSConvert, a second low-aggression peak picking algorithm using SciPy's find peaks function was applied, as during testing, an overflow of noise was still detected (Virtanen et al., 2020) . The program used the m/z values of the given MS/MS to create a distance matrix containing all the m/z distances from each peak to every other peak. These m/z values span from 19 Da, the weight of a water molecule with an extra proton, to the original precursor mass. In an optimal circumstance, where every possible y-ion was observed, and its m/z value is part of the m/z array, the entire peptide sequence can be interpreted.

This distance matrix contains the m/z array as both axes and computes the distance between each peak in the MS/MS. After it calculated the distances, it recognized which m/z distances (in Da) match the weight of an amino-acid or pair of amino-acids. The weight of this amino acid or pair of amino acids needed to match this distance within 0.05 Da. Even though the possible accuracy of the Orbitrap instrument measures up to 0.002, this threshold was chosen at the cost of a little more runtime to capture more possible results. The 0.002 threshold also performed less well during testing, missing real peaks. Additionally, amino-acid hits caught at a too high threshold are likely to be eliminated during confidence scoring. The addition of matching for amino-acid pair weights was done to be able to account for missing peaks in the MS/MS. This, however, increases the amount of possible distances that could be matched in the distance matrix from 22 (all original amino acids) to 506 (all possible combinations of amino acids + original amino acids), impacting runtime. In addition to including potentially missing peaks, these amino-acid pairs could increase confidence in a specific peptide sequence by introducing different versions of it into the dataset. Triplets of amino acids were ignored because they would increase run-time beyond a reasonable amount. The amino acids leucine

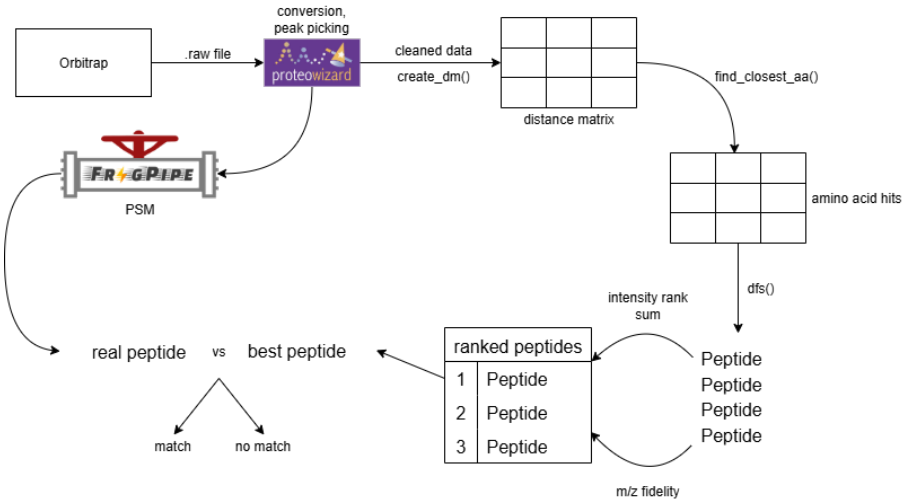


Figure 2.3: General pipeline of the programmatic approach. Data obtained from the Orbitrap in the form of a *.raw* file is converted into a *.mzML* file. This file is cleaned and filtered for peaks before being transformed into a distance matrix. Creating a table where the rows and columns are both the m/z values of the MS/MS. Each cell in the table contains the distance between the peak in its row index and the peak in its column index. Each distance is then matched to the weight of an amino acid or pair of amino acids. A depth-first search path-finding algorithm is then applied to every single path through the distance matrix. Here, the peaks are seen as individual nodes, and the edges connecting them are the amino acids. This results in a list of possible peptides that are then ranked through a confidence score. The peptide sequence with the highest confidence score is then compared to the real peptide sequence obtained through a database search using MSFragger. Thereby, we can measure the accuracy of the model.

(L) and isoleucine (I) are not distinguishable since they have the same mass. To resolve this, both are identified as L, and adjustments are made during scoring. Figure 2.4 visualizes this transformed matrix, where every circle is a successfully identified amino-acid or pair that "fits" into the distance between the 2 peaks. Figure 2.5 shows a part of the real distance matrix file for *HELIEFR*. Given this information, the program used a depth-first search algorithm, which interpreted the peaks as nodes and amino acids as edges to find every possible pathway through the matrix. While the graph-based distance matrix approach proposed here has been discovered independently, similar approaches to *de novo* identification have been described in the literature. The closest approach is the sequence graph described in the Lutefisk program (Taylor & Johnson, 2001). The programs pNovo+ and UniNovo use a similar spectrum graph for evaluating individual peptide sequences (Chi et al., 2013; Jeong et al., 2013) . PepNovo uses a graph-based approach to predict peptide fragmentation under specific restrictions (Frank & Pevzner, 2005).

After pathfinding, the resulting amino-acid sequences were ranked based on a confidence score metric. This metric was adapted from the intensity rank sum and m/z fidelity concepts

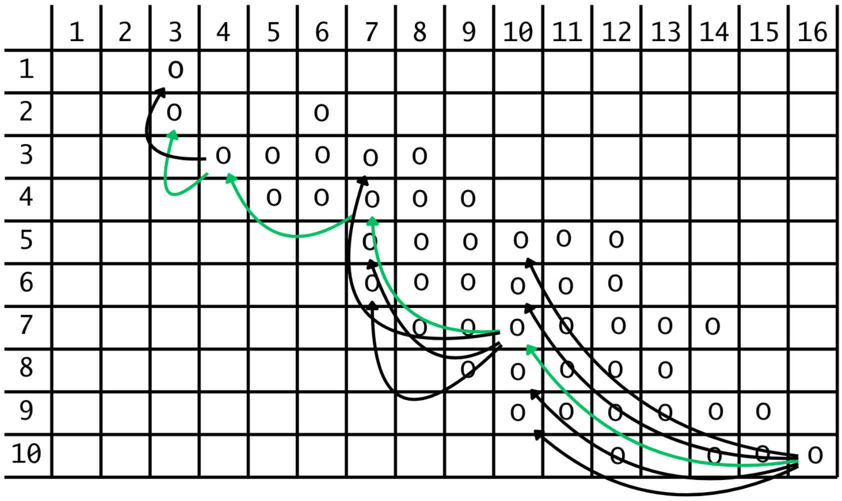
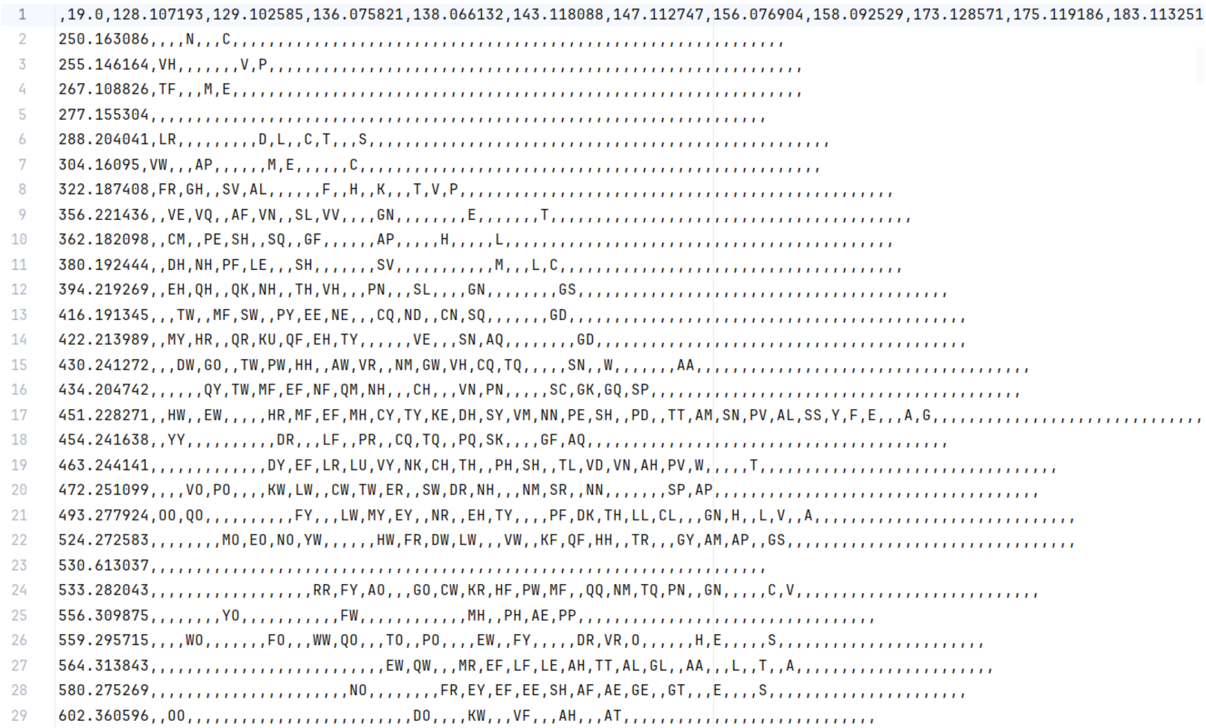


Figure 2.4: Visual representation of a single run through the distance matrix. This distance matrix contains the m/z array for both axes and computes the distance between each peak in the MS/MS. After it calculates the distances, it recognizes which distances (in Da) align with the weight of an amino acid. Cells with a circle are those where an amino acid could fit between the peaks. The program then computes every single pathway through the distance matrix and ranks the outcomes with a confidence score. It uses a recursive depth-first search to find every single pathway. A single possible pathway is mapped in the figure.



```
,19.0,128.107193,129.102585,136.075821,138.066132,143.118088,147.112747,156.076904,158.092529,173.128571,175.119186,183.113251
250.163086,,,,N,,,C,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
255.146164,VH,,,,,,,V,P,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
267.108826,TF,,,M,E,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
277.155304,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
288.204041,LR,,,,,,,,,D,L,,C,T,,,S,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
304.16095,VW,,,AP,,,,,,M,E,,,,,,C,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
322.187408,FR,GH,,SV,AL,,,,,,F,,H,,K,,,T,V,P,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
356.221436,,VE,VQ,,AF,VN,,SL,VV,,,,GN,,,,,,,,E,,,,,,,T,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
362.182098,,CM,,PE,SH,,SQ,,GF,,,,,,AP,,,,,H,,,,,L,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
380.192444,,DH,NH,PF,LE,,,SH,,,,,,,SV,,,,,,,,,,,M,,,L,C,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
394.219269,,EH,QH,,QK,NH,,TH,VH,,,PN,,,SL,,,,GN,,,,,,,,GS,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
416.191345,,,TW,,MF,SW,,PY,EE,NE,,,CQ,ND,,CN,SQ,,,,,,,GD,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
422.213989,,MY,HR,,QR,KU,QF,EH,TY,,,,,,VE,,,SN,AQ,,,,,,,,GD,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
430.241272,,,DW,GO,,TW,PW,HH,,AW,VR,,NM,GW,VH,CQ,TQ,,,,,SN,,W,,,,,,,AA,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
434.204742,,,,,,QY,TW,MF,EF,NF,QM,NH,,,CH,,,VN,PN,,,,,SC,GK,GQ,SP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
451.228271,,HW,,EW,,,,,HR,MF,EF,MH,CY,TY,KE,DH,SY,VM,NN,PE,SH,,PD,,TT,AM,SN,PV,AL,SS,Y,F,E,,,A,G,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
454.241638,,YY,,,,,,,,,,DR,,,LF,,PR,,CQ,TQ,,PQ,SK,,,,GF,AQ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
463.244141,,,,,,,,,,,,,DY,EF,LR,LU,VY,NK,CH,TH,,PH,SH,,TL,VD,VN,AH,PV,W,,,,,T,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
472.251099,,,,VO,PO,,,,KW,LW,,CW,TW,ER,,SW,DR,NH,,,NM,SR,,NN,,,,,,,SP,AP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
493.277924,OO,QO,,,,,,,,,,FY,,,LW,MY,EY,,NR,,EH,TY,,,,PF,DK,TH,LL,CL,,,GN,H,,L,V,,A,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
524.272583,,,,,,,,MO,EO,NO,YW,,,,,,HW,FR,DW,LW,,,VW,,KF,QF,HH,,TR,,,GY,AM,AP,,GS,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
530.613037,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
533.282043,,,,,,,,,,,,,,,,,,RR,FY,AO,,,GO,CW,KR,HF,PW,MF,,QQ,NM,TQ,PN,,GN,,,,,C,V,,,,,,,,,,,,,,,,,,,,,,,,,,
556.309875,,,,,,,,YO,,,,,,,,,,,FW,,,,,,,,,,,,MH,,PH,AE,PP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
559.295715,,,,WO,,,,,,,FO,,,WW,QO,,,TO,,PO,,,,EW,,FY,,,,,DR,VR,O,,,,,,H,E,,,,,S,,,,,,,,,,,,,,,,,,,,,,,,
564.313843,,,,,,,,,,,,,,,,,,,,,,,,,,EW,QW,,,MR,EF,LF,LE,AH,TT,AL,GL,,AA,,,L,,T,,A,,,,,,,,,,,,,,,,,,,,,,,
580.275269,,,,,,,,,,,,,,,,,,,,,,NO,,,,,,,,FR,EY,EF,EE,SH,AF,AE,GE,,GT,,,E,,,,S,,,,,,,,,,,,,,,,,,,,,,
602.360596,,OO,,,,,,,,,,,,,,,,,,,,,,,,,,,DO,,,,KW,,,VF,,,AH,,,AT,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
```

Figure 2.5: Real example of a distance matrix file. This file represents part of the distance matrix for the peptide HELIEFR, which was fully identified. It is a comma-separated file where the empty spaces represent distances that could not able to be matched to the weight of an amino acid. The shape of the identified amino acids represents an oblong diagonal.

introduced in Tabb et al., 2008. To calculate intensity rank sum (IRS) each peak is assigned a rank based on its intensity so that the highest peak has a rank of 1 and the second highest of 2, as can be seen in Figure 2.6 . When a potential peptide sequence is found, the ranks for each peak belonging to the sequence are summed. The lower the rank sum, the higher the chance that the peptide sequence is real.

$$IRS = \sum_{i=1}^{n} Rank(peak_i)$$

We convert this into a probability of how likely this rank or a smaller one appears in a random distribution. To do this, we calculate a random distribution of rank sums with the same number of peaks as the original and observe how many times a smaller rank sum than our original appears; with this, we obtain p_{rank}.

M/z fidelity represents how well the calculated peptide sequence fits into the real MS/MS. We assume that, when observing two real peaks, the distance between them amounts to exactly the weight of an amino acid. However, there will be occurrences where the distance between two peaks is almost exactly the weight of an amino acid by accident. To avoid these fake hits in our final peptide, we take the lowest m/z value peak of our sequence and attempt to estimate its position. We do this by using the next highest peak and subtracting the amino acid between it and our smallest peak. Repeating this for every peak until we have $n-1$ estimations of our smallest peak, where n is the number of peaks.

$$\begin{aligned}
e_0 &= peak_1 \\
e_1 &= peak_2 - |AA_1| \\
e_2 &= peak_3 - |AA_1 + AA_2| \\
e_3 &= peak_4 - |AA_1 + AA_2 + AA_3| \\
e_{n-1} &= peak_n - |AA_1 + AA_2 + ... + AA_{n-1}|
\end{aligned}$$

Calculating the mean estimation and subtracting it from each estimation gives us an estimated error, which we can square and sum together to receive a sum squared error SSE. Again, the lower this value is, the higher the probability that the sequence is real.

$$e_{mean} = \frac{\sum_{i=1}^{n} e_i}{n}$$

$$SSE = \sum_{i=1}^{n} (e_i - e_{mean})^2$$

Then we can determine the probability that this SSEor a smaller one appears in a random distribution. For this, we calculate a uniform distribution of possible errors. Since we know each error can only fit into the 0.05 threshold we defined, we generate our error distribution within that margin. Then we randomly select n different errors and determine the mean with which we can calculate a simulated SSE score. Using this method, we create a random distribution of SSE scores and observe how many times a smaller SSE value than our real value appears to obtain a p_{mz_fid} .

We then normalized these two probabilities and combined them using Fisher's method to get a final confidence score. The lower this combined confidence score, the higher the probability that the sequence is real.

$$X_{2k}^2 = -2 \sum_{i=1}^{k} \ln p_i$$

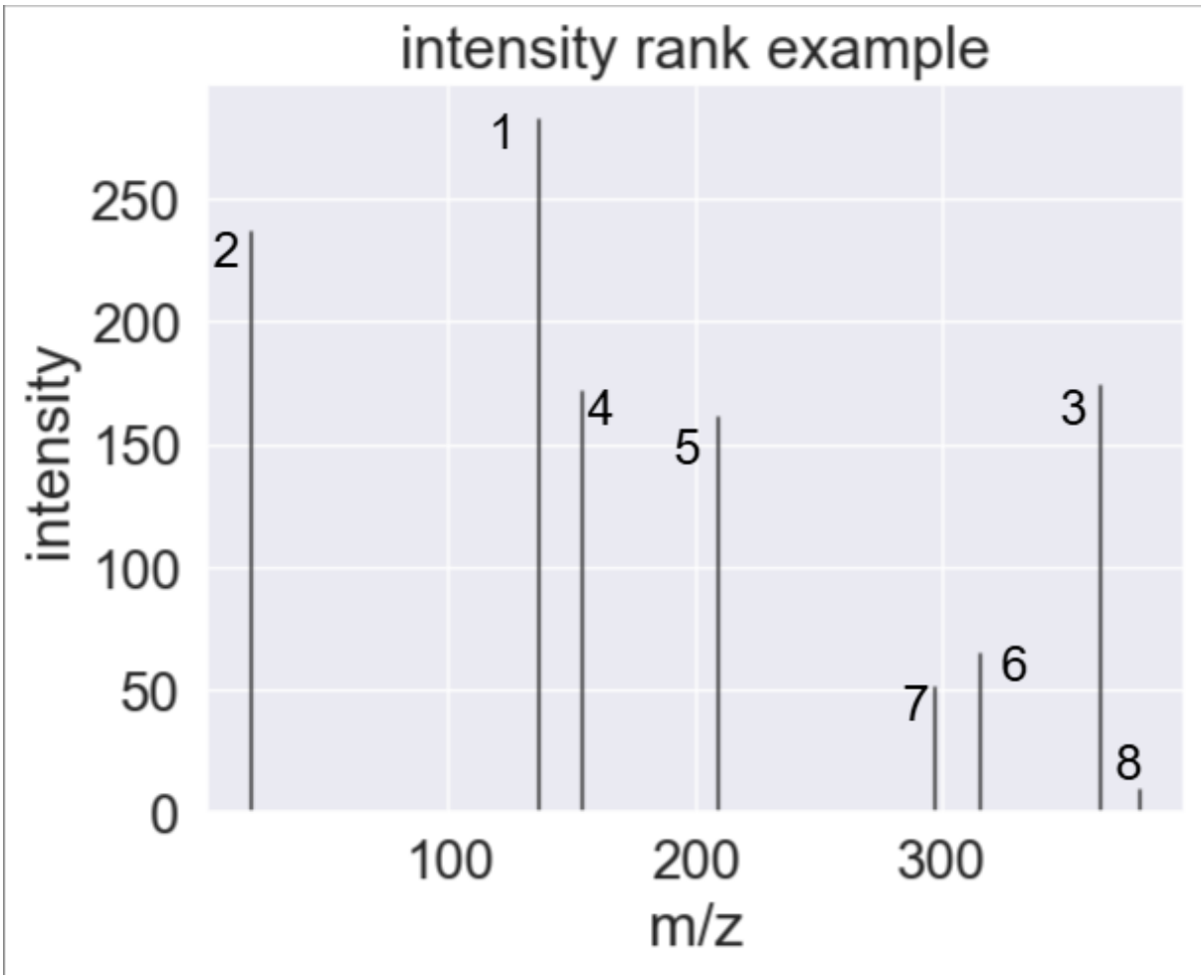


Figure 2.6: Example of intensity ranking. Each peak is given a rank depending on its relative height. The highest peak is given a 1, the second highest a 2, and so on. The ranks for the peaks of an identified sequence are then summed together to receive an intensity rank. This real intensity rank is then compared to a random distribution of intensity ranks with the same number of peaks. We then obtain the likelihood of our real intensity rank randomly occurring by scoring how many ranks in the random distribution are lower or the same value. This is the p_{rank} metric.

Chapter 3

Results and Analysis

This section will present the results obtained from *de novo* sequencing a subset of 487 MS/MS from the MRC-5 cell line. Each MS/MS generated multiple possible peptide sequences using a distance matrix approach, and these were ranked with a confidence score combined from intensity score and m/z fidelity. The accuracy of predictions was then gauged against the ground truth generated through MSFragger, a database search algorithm for peptide spectrum matching (Kong et al., 2017). A score is given based on how well the predicted best peptide overlaps with the ground truth. If a predicted peptide has three or more amino acids in the same order as the ground truth, it counts as an overlap. The overlap function could also identify 2 or 3 separate tags of three or more amino acids that overlapped. The percentage of overlap is determined by how many amino acids are overlapping relative to the full length of the ground truth peptide. 100% overlap means both peptides are the same, and 50% overlap means half of the amino acids from the ground truth are also present in the predicted best peptide in the same order. The mean confidence score represents the accuracy of the scoring metrics for each percentile group. The lower this score, the better. To evaluate the accuracy of this model and its potential, we will first look at the highest rated peptide suggested by the program, then at the Top 10, the Top 50, and the Top 100. Comparing these can indicate how well the confidence score predicts the most accurate peptide sequence.

Percentage identified	Frequency	Mean confidence score	Mean entropy
100%	2.25%	1.1^{-8}	8.01
75%-99%	2.05%	-	8.97
50%-74%	9.24%	1.4^{-7}	8.71
25%-49%	27.72%	0.00054	9.28
1%-24%	11.9%	0.0011	7.47
0%	46.81%	0.051023503	5.72

Table 3.1: Performance metrics for the highest ranking peptide sequence for each of 487 MS/MS.

3.1 Evaluation Metrics

3.1.1 Top 1

For the **Top 1** evaluation, 487 different MS/MS were analysed and inferred through the program. The peptide sequence with the highest confidence score was measured and compared with the ground truth. A 100% overlap occurred 2.25% of the time. This means the program generates a true hit with an accuracy of 2.25% when looking at the highest-ranking peptide. This function for checking overlap was also applied to the rest of the dataset. As seen in Table 3.1: In 2.05% of cases, the highest rated peptide overlapped with more than 75% of the correct sequence. In 9.24% of cases, the highest rated peptide overlapped with more than 50% of the correct sequence. In 27.72% of cases, the highest rated peptide overlapped with more than 25% of the correct sequence. In 11.9% of cases, the highest rated peptide overlapped with more than 1% of the correct sequence. In 46.81% of cases highest rated peptide did not overlap with any part of the peptide. Confidence scores on average improved as identification percentage rose. However, outliers were observed like *AFMTADLPNELLELLEK* whose highest ranking peptide had 0% overlap but a confidence score of 1^{-9} while *LLEETLALK* had a score of 1.9^{-8} and had 100% overlap with the correct peptide.

The absolute numbers for overlap percentages can be seen in Figure 3.1 . Additionally, the figure shows the average peptide length and average entropy in each column. It can be seen that longer peptides are more likely to end up in the lower end of identification, while shorter peptides are more likely to be fully identified. There is also a small trend towards lower entropy being on the lower end of identification and high entropy being on the higher end. The 46.81% of cases had an average of 9592 possible peptide combinations, but the highest rated one of these combinations did not have a single sequence tag of three amino acids or more in common with the correct sequence.

Figure 3.1: Plot displaying how high a percentage of the peptide was identified by the program within the Top 1. Divided into quarters as well as 100% and 0% identification. Using the highest confidence score as a metric. A total of 487 runs.

Percentage identified	Frequency	Mean confidence score	mean entropy
100%	5.95%	4^{-9}	8.3
75%-99%	5.54%	6^{-9}	9.58
50%-74%	17.84%	1.9^{-8}	9.02
25%-49%	26.48%	0.00059	9.15
1%-24%	12.93%	0.0028	7.2
0%	31.21%	-	4.22

Table 3.2: Performance metrics for the best of the 10 highest ranking peptide sequences for each of 487 MS/MS.

3.1.2 Top 10

For the **Top 10** evaluation, the top 10 highest-ranking peptides were checked by the overlap function, and the peptide with the highest overlap was chosen. The distribution changes slightly when running the overlap checking function over the top 10. In a practical setting, it would require a manual check to identify the correct peptide out of the top 10, which is why these numbers only show potential for the program to improve, not actual effectiveness. All the frequencies of higher percentage identification have gone up. The frequency of identifying full peptides has doubled. The mean confidence score for full identification has worsened. The mean confidence score for 75 - 99% of identification has improved, and the frequency of observation has doubled. Also, a visible increase in frequency of observation for 50-74% peptides and the mean confidence score has improved slightly.

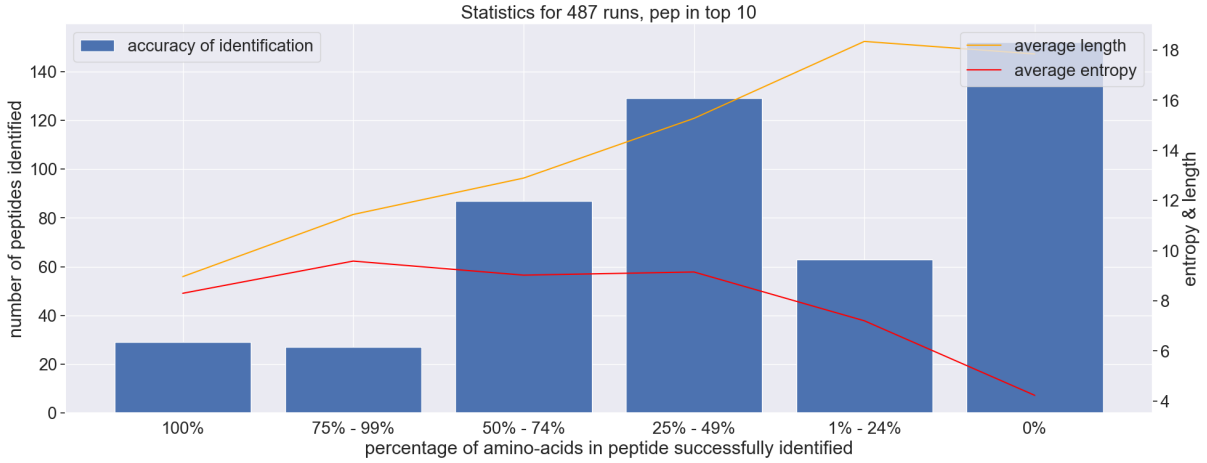


Figure 3.2: Plot displaying how high a percentage of the peptide was identified by the program within the Top 10. Divided into quarters as well as 100% and 0% identification. Using the highest confidence score as a metric. A total of 487 runs.

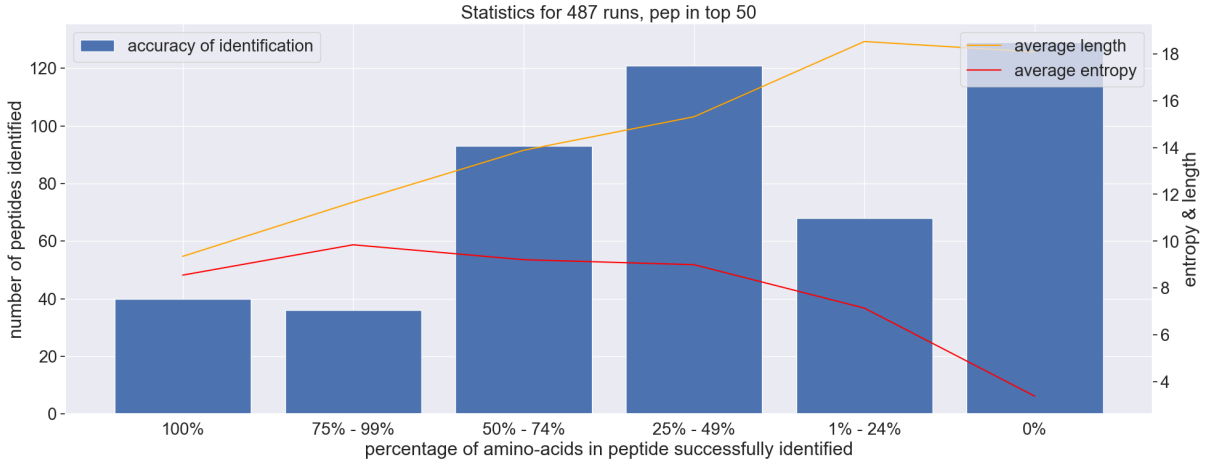


Figure 3.3: Plot displaying how high a percentage of the peptide was identified by the program within the Top 50. Divided into quarters as well as 100% and 0% identification. Using the highest confidence score as a metric. A total of 487 runs.

3.1.3 Top 50

For the **Top 50** we chose the best overlapping peptide out of the top 50 ranked peptides per MS/MS. This time we saw improvements again, but diminished compared to those we had when we went from top 1 to top 10, even though this time we increased our search scope by 5 times the amount. Showing that we receive diminishing returns from incorporating more peptides from the ranking. However, an 8.21% frequency of observing full peptides is seen. Whereas 50 - 99% improved by 2% , and 25 - 45% identification decreased. 0% identification decreased while 1-24% increased.

Percentage of peptide identified	Frequency	Mean confidence score	Mean entropy
100%	8.21%	3^{-9}	8.54
75%-99%	7.39%	5^{-9}	9.84
50%-74%	19.09%	9.05^{-6}	9.2
25%-49%	24.84%	0.00066	8.99
1%-24%	13.96%	0.00275	7.13
0%	26.48%	-	3.37

Table 3.3: Performance metrics for the 50 highest ranking peptide sequences for each of 487 MS/MS.

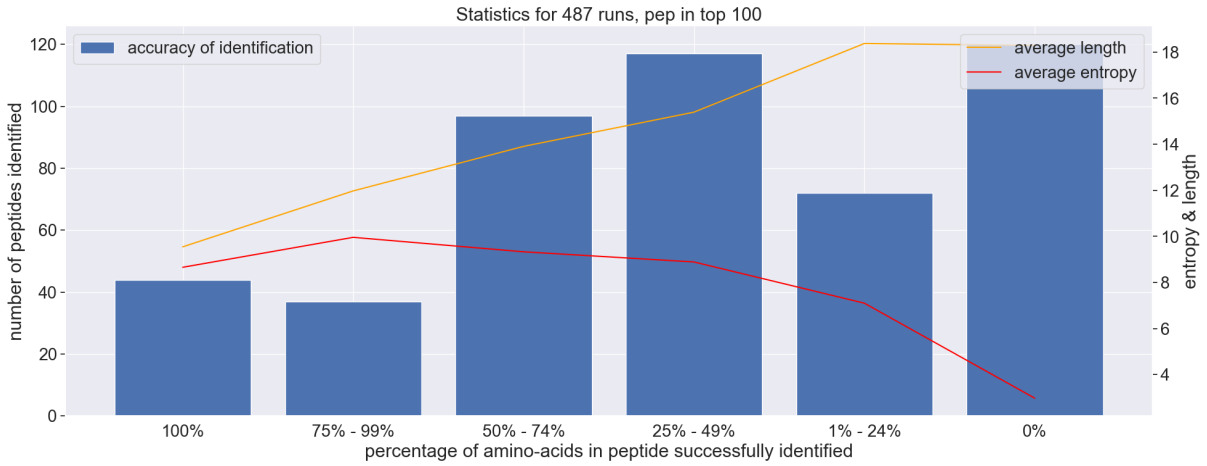


Figure 3.4: Plot displaying how high a percentage of the peptide was identified by the program within the Top 100. Divided into quarters as well as 100% and 0% identification. Using the highest confidence score as a metric. A total of 487 runs.

3.1.4 Top 100

When moving from **Top 50** to **Top 100** we only see increases or decreases within a 1% margin when looking at the observation frequency.

Percentage of peptide identified	Frequency	Mean confidence score	mean entropy
100%	9.03%	3e-09	8.65
75%-99%	7.59%	5e-09	9.95
50%-74%	19.91%	8.67e-06	9.32
25%-49%	24.02%	0.00068	8.88
1%-24%	14.78%	0.00261	7.09
0%	26.64%	-	2.97

Table 3.4: Performance metrics for the 100 highest ranking peptide sequences for each of 487 MS/MS.

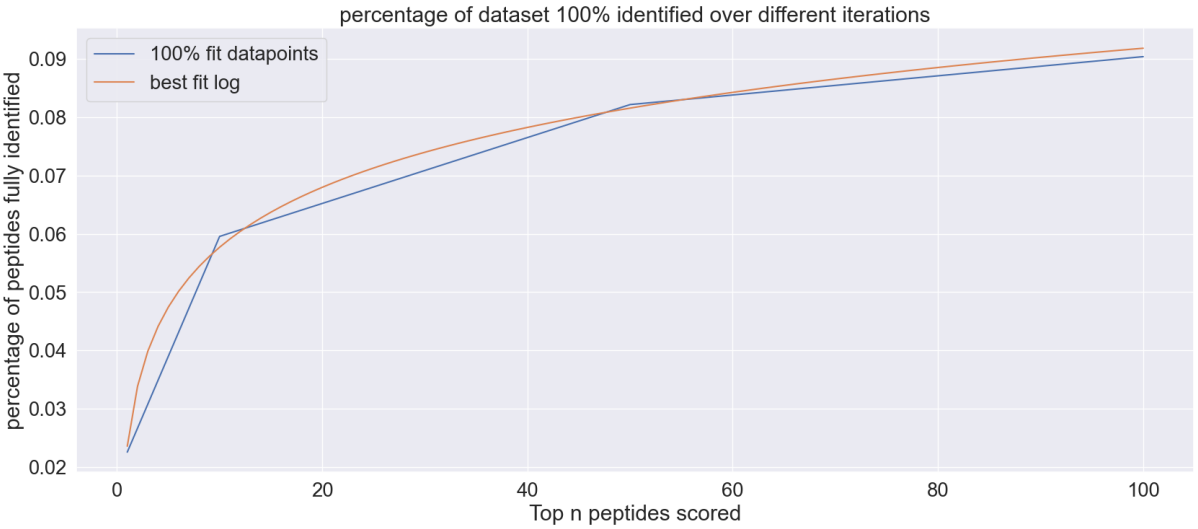


Figure 3.5: This graph displays how much of the dataset was 100% revealed, depending on how many of the top-rated peptides were being checked for overlap. A log function was discovered to be the best fit to describe this relationship $(0.0148 * \log Top\_n + 0.0236)$. The strategy of checking more of the top-rated peptides seems to give diminishing returns. Suggesting that while the confidence score does not perfectly evaluate each peptide, it is not an arbitrary measure.

3.2 Data over Top1 - Top100

The percentage of full identification of peptides has gone up each time when observing more of the Top n peptides, but is yielding diminishing returns. The curve can be best described by a logarithmic function $(0.0148 * \log Top\_n + 0.0236)$. Figure 3.6 shows the confidence score distribution for all 487 runs collectively. The shape of the graph shows a diagonal cutoff declining as the index number grows.

3.3 Individual Spectra

Figure 3.7 shows the confidence score distribution for *LLEETLALK*. Its highest ranking possible peptide sequence for the MS/MS was identified 100% correctly. Figure 3.8 shows the confidence score distribution for *AFMTADLPNELLELLEK*. For which the highest ranking possible peptide sequence was identified 0% correctly. The graph of LLEETLALK shows that the confidence score is slowly worsening as it ascends over the index in a slower curve. For the graph AFMTADLPNELLELLEK, we see the confidence score worsening a lot quicker before it flattens.

Looking at the individual confidence score distributions, we can observe a few more trends. As the combined confidence score improves, we see rank intensity slowly approaching better

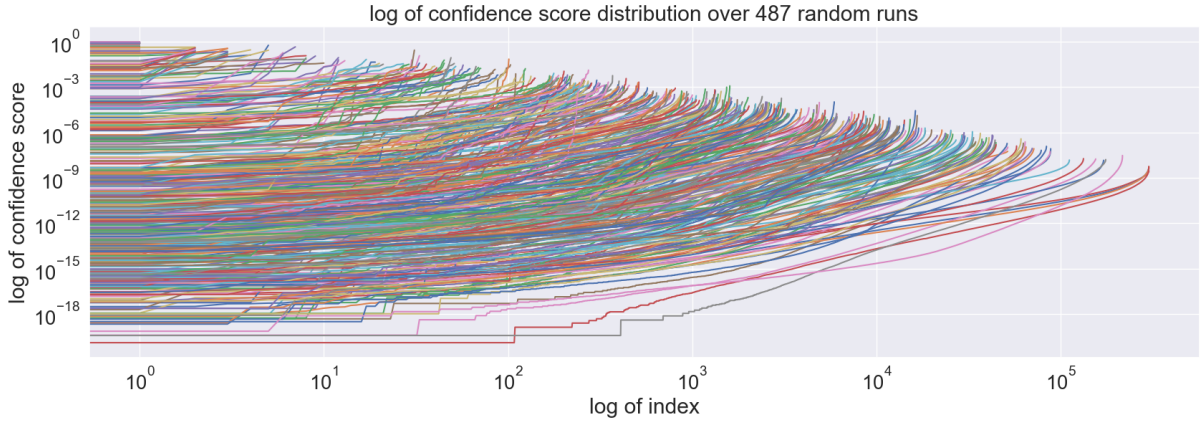


Figure 3.6: The distribution of confidence scores over all 487 runs. The lower the confidence score, the better. An interesting trend can be observed. A diagonal cutoff is visible, suggesting that peptides with a bigger index might have better confidence scores.

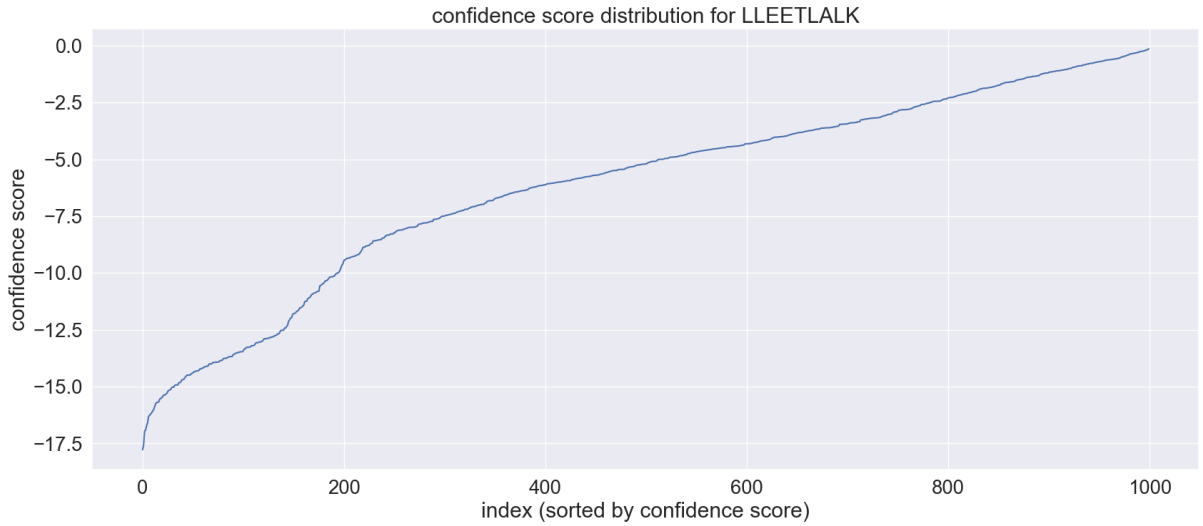


Figure 3.7: The confidence score distribution for *LLEETLALK*. The graph shows a slow curve over the confidence score improving.

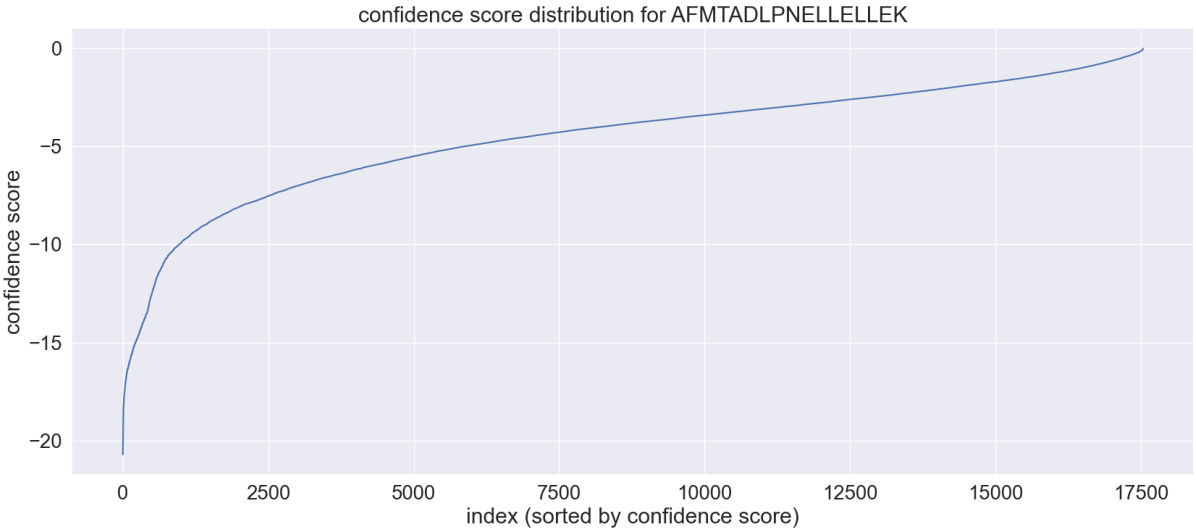


Figure 3.8: The confidence score distribution for *AFMTADLPNELLELLEK*. The graph shows many similar confidence scores across the dataset, with a few strong ones creating a peak in the graph.

and better scores. M/z fidelity scores, however, quickly plateau at 10^{-6}. This can be seen in Figure 3.9 and Figure 3.10.

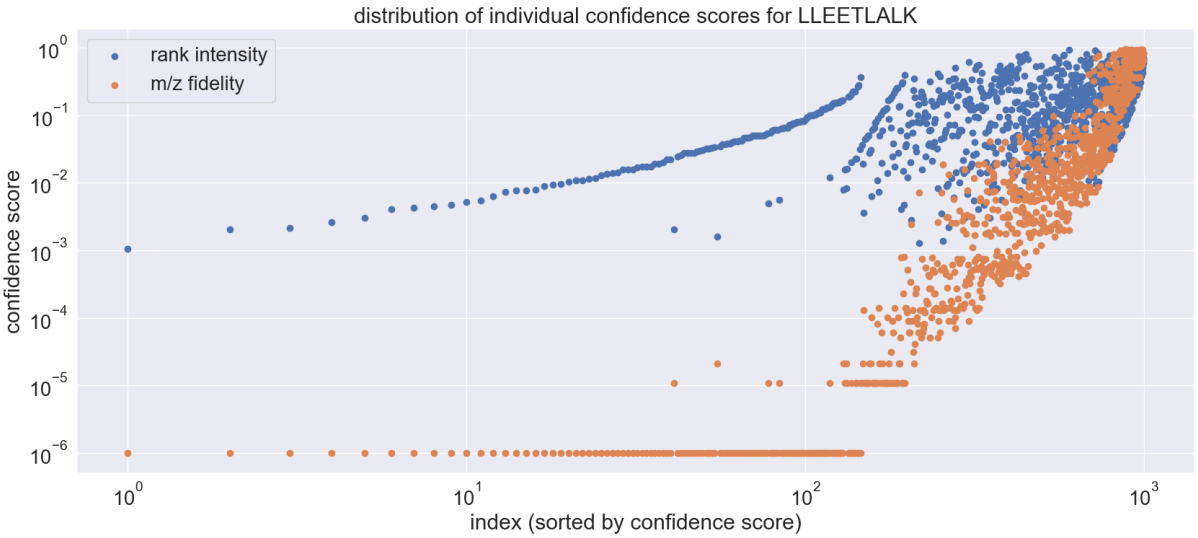


Figure 3.9: Plotting distribution of m/z fidelity score (p_{mz_fid}) and rank intensity (p_{rank}) for *LLEETLALK*. These individual scores are being summed together using Fisher's method to obtain the final confidence score metric. A random distribution can be seen towards the higher indices. As indices increase, the rank intensity slowly increases as well while m/z fidelity plateaus at 10^{-6} and 10^{-5}.

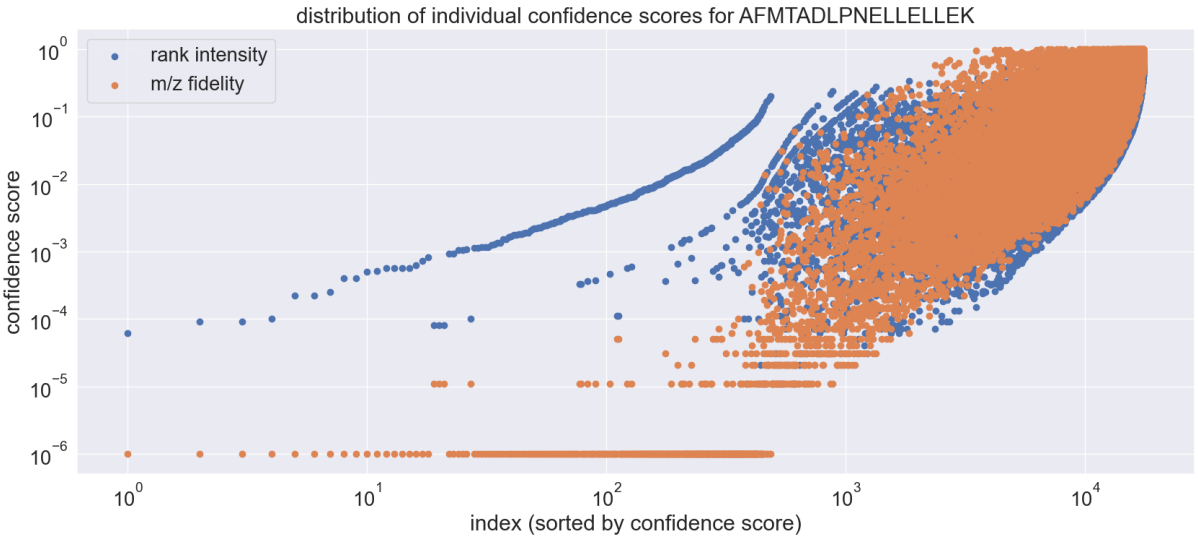


Figure 3.10: Plotting distribution of m/z fidelity score (p_{mz_fid}) and rank intensity (p_{rank}) for *AFMTADLPNELLELLEK*. These individual scores are being summed together using Fisher's method to obtain the final confidence score metric. A random distribution can be seen towards the higher indices. As indices increase, the rank intensity slowly increases as well, while, similar to Figure 3.9, m/z fidelity plateaus at 10^{-6} and 10^{-5}, but also at $10^{-4.7}$ and $10^{-4.5}$.

Chapter 4

Discussion

This section will discuss the results obtained from the model. Identify weaknesses and strengths of the approach and give insight into possible improvements.

4.1 Performance Insights

With only 2.25% of MS/MS identified correctly based on the highest ranking peptide for each MS/MS, the performance of this *de novo* engine is not yet fit for practical application. Correlations between high confidence scores and accurate identification can be observed. However, there are also strong outliers such as the *AFMTADLPNELLELLEK* peptide. Beyond that, scans interpreting the **Top 10** or **Top 50** managed to double and even triple the 100% identification frequency, respectively. This shows that 100% fits do exist within the dataset, but don't perform good enough in confidence scoring to reach first place reliably and therefore cannot confidently be identified. The **Top 100** showed little to no further improvements in full identification percentage. As seen in Figure 3.5 the improvement gained over checking more of the Top n highest scored peptides is logarithmic. Meaning that while the confidence score is not capturing every peptide, it is also not arbitrary. The mean confidence scores in the 100% identification were consistently better than in lower identification ranges. This reflects on the confidence score metrics, suggesting that it could be a good absolute metric. However, outliers like the *AFMTADLPNELLELLEK* peptide weaken that claim. Along with low full identification, this likely means there is significant room for improvement for these scoring metrics. The entropy measurement was originally introduced to assess the likelihood of a confidence score distribution containing a hit. If a strong peak is observed, it would suggest one single peptide

combination strongly outperforming the others. The mean entropy measurements were unexpected. If the distribution of confidence scores showed a strong peak, as would be expected for the correct peptide prediction compared to other peptide predictions, it would indicate a peptide much more likely than the rest. However, as we can see in Figure 3.7 , the distribution slowly curves towards better confidence scores and doesn't show one strong peak. Rather, it shows multiple low fit scores and slowly more and more likely scores. For Figure 3.8 a few very strong confidence scores appear compared with a quick worsening in confidence scores as they flatten over higher indices. This suggests that the entropy score could be used as a measurement in the exact opposite way as originally proposed. A distribution like in Figure 3.7 shows a slow curve because as every single path along the distance matrix gets explored longer and longer correct sequences are being found with better and better confidence scores. Until finally, the full sequence is being explored with likely the best confidence score. Therefore, the expected distribution would be a slow curve, giving a high entropy score. For distributions like in Figure 3.8 , where a good confidence score is wrongfully assigned, will most likely not have the same slow ascension towards that confidence score, but just a few sequences that, by random chance, fit the confidence measures well. Therefore, resulting in a low entropy score. Looking at the distribution of individual confidence scores, it is visible that m/z fidelity quickly plateaus at 10^{-6} and beyond that point can no longer differentiate between different quality peptide predictions. This is likely the case because the m/z fidelity scores have become so small compared to the generated distribution of 100.000 random SSE's for m/z fidelity that there is no single random value above them, which allows the p_{mz_fid} value to reach the minimum. To adjust the metric to this issue, either the number of generated distributions needs to be increased, or the range in which random errors are generated needs to be decreased. Additionally, the shape that emerges in Figure 3.6 indicates that the longer the index list of a distribution, meaning the more possible peptide sequences for a single MS/MS, the better the worst confidence score in the distribution compared to other distributions. This is characterized by the diagonal cutoff visible. Since, logically, a MS/MS with higher m/z values has a bigger distance matrix, more amino acids, and therefore likely a longer list of possible peptide sequences. This could indicate a bias in the confidence scoring towards longer peptide sequences, which were shown to have a lower likelihood of being successfully identified (see Table 3.1).

4.2 Comparing spectra and outliers

As of now, we have seen a correlation between high confidence scores and successful identification. However, the peptide *AFMTADLPNELLELLEK*, though having a high confidence score, did not find a match. To figure out how a wrong peptide sequence got such a high score, it is helpful to go by how each score is measured and see how it could be abused. One thing that needs to be mentioned first is that there were several different MS/MS that had a peptide spectrum match for this peptide. Figure 4.1 and Figure 4.2 are two examples. There is a high likelihood that the program would fail to interpret Figure 4.2 as it has few peaks and does not match the length of the peptide sequence it is supposed to represent. There are quite a few suboptimal MS/MS in this dataset that are virtually impossible to match through *de novo*. This will always be the case in a realistic scenario. A few things that are also seen in 4.1 are the missing peaks at y6, y15 and y16. This peptide has a length of 17 amino acids, meaning y17 would be the full-length peptide. Since it is missing two amino acids right at the beginning, this being a gap, the program can't close; it will likely miss the first two amino acids completely. Literature suggests that this is oftentimes the case since fragmentation at the N-terminal amino acids is difficult (Taylor & Johnson, 2001). But besides that and possibly y9 and y8 being removed through too aggressive peak picking, this MS/MSshould be interpretable. It was assumed that the figure 4.2 is most likely the reason for the bad match. It could have received a good intensity rank sum because there are very few peaks over a large area. However, with such few peaks, it should get a bad m/z fidelity score. After looking at the program logs, however, it was found that Figure 4.1 was used for identification. Meaning that this spectrum became an outlier by randomly fitting the confidence metrics. For a successful interpretation of MS/MS, similar to this, more metrics are likely needed. A possible way to circumvent low information MS/MS like Figure 4.2 could be to look for matching MS/MS within the unidentified dataset, and if they show a significant overlap, merge them into a single MS/MS. This process needs to be monitored so it does not lead to misinterpretation in case of a mismatched merge.

4.3 Ideas for improvement

Possible Ideas for improvement include attempting to more closely replicate the original idea for intensity rank sum, where only tags of length 3 are being evaluated. Splitting the peptide into multiple tags and ranking them individually. This would also be a good idea for the m/z

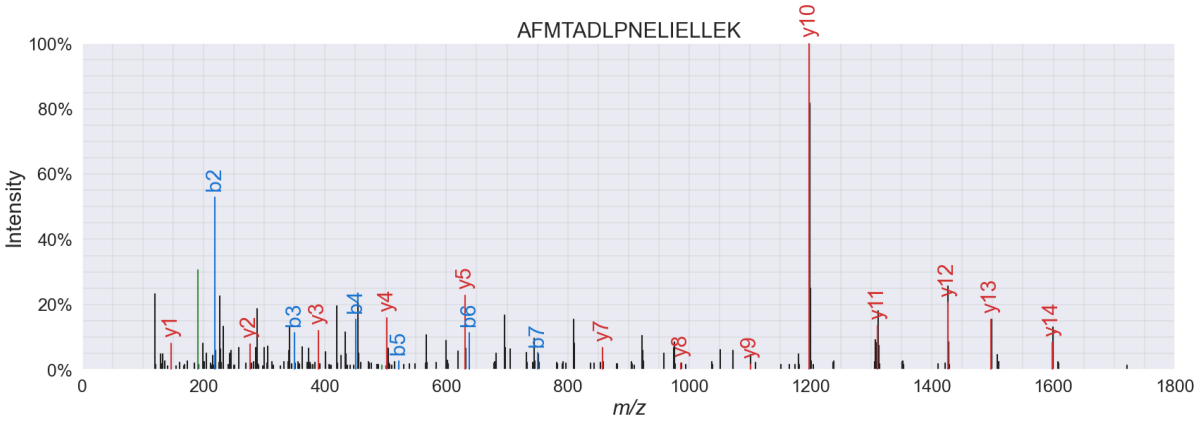


Figure 4.1: Automatically annotated version of *AFMTADLPNELLELLEK* with pyteomics. Automatic annotation based on a previous database search identified a total of 13 out of 16 peaks associated with y-ions

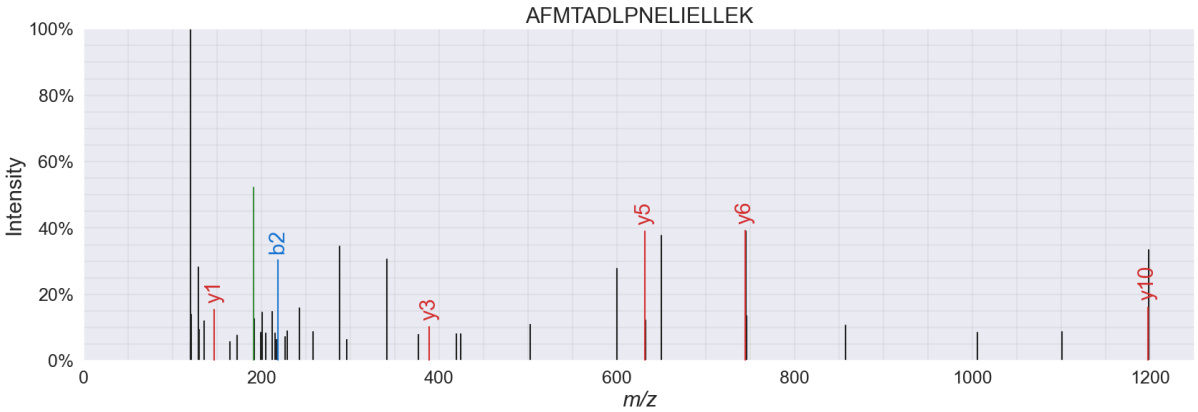


Figure 4.2: Another automatically annotated version of *AFMTADLPNELLELLEK* with pyteomics. A difficult-to-interpret spectrum since even the automatic annotation can only find 5 out of 16 peaks. Many Spectra in the MRC-5a dataset have similarly little information and are therefore more difficult to interpret.

fidelity metric since one major drawback of it includes the fact that if one of the earlier peaks is skewed, it skews every following peak, increasing the chance of a bad m/z fidelity score as peptides get longer. Additionally, Tabb et al., 2008 suggests a third evaluation metric for sequence tags, complementarity. This metric involves looking for the complementary b-ion peak that we expect to find for our y-ion peak. B-ions are less abundant than their counterparts, and the successful identification of a complementary peak would improve confidence in its validity significantly. Identifying isotopes could provide an additional confidence metric. If the y-ion peak we identify has another, smaller, peak exactly 1 Da away, this again suggests validity. A second, even smaller, peak 2 Da away is very rare but can also be a good metric. The same thing applies to fragment ions that lost ammonium or water, as they would weigh around 17 or 18 Da less, respectively. Another obvious trend we can observe in the data is that the longer the peptide sequences get, the harder the identification becomes. This could again be because of how the confidence metrics approach the entire peptide instead of parts. Moving from a direct interpretation approach to one where intermediary tags are being inferred before moving to the full peptide shows promise to improve this program. This also follows the interpretation of Figure 3.6 , which portrays a possible bias for longer peptide sequences in the confidence scores. This would more efficiently circumvent missing peaks as well as possibly identify isolated parts of the distance matrix. Adding the possibility of circumventing missing peaks without needing specific amino-acid masses to fill the space could lead to more accurate tag sequences if applied correctly. Additionally, I believe it to be important to improve the peak picking as the SciPy find peaks algorithm missed some important peaks in manual testing. When trying other algorithms during testing, such as peak-utils and spectrum-utils results were comparable to SciPy. When removing SciPy completely, however, the distance matrix increased in size substantially and extended runtime considerably while also losing accurate peptide sequences that were found in trials before. Peak picking using concepts such as full width at half maximum (FWHM) is a possible approach. This would, however, require restructuring of the entire pipeline as conversion with MSConvert centroids data and removing the width measurement. During the peak picking process, it would also be helpful to remove peaks we can already identify as not being y-ions, such as the doubly-charged parent ion, its isotopes, and again, possible variations of it. Implementing a pre-review and sorting MS/MS by their ability to be identified could provide another valuable heuristic for the program. Identifying the length of the possible peptide, given the largest observed m/z value, is simply done by dividing by the average amino acid weight.

The number of peaks can also serve as a heuristic for identification. Another drawback that was not implemented and needs to be improved is that the overlap function for peptides does not optimally consider pairs of amino acids. Since a pair of amino acids A and B can be paired as BA or AB, it might not perfectly match the original sequence. This has not been accounted for in the overlap function.

4.4 Comparison with Literature

Current state-of-the-art *de novo* software like PepNet, PointNovo, and DeepNovo achieves better results than the software presented here, both in accuracy and runtime. However, this software's approach to *de novo* sequencing differs from theirs. Even a similar approach like Lutefisk's sequence graph has some distinct differences. Compared to Lutefisk, Uni-Novo, and pNovo+ this program provides a visually representative matrix that shows an oblong diagonal region whose visual structure might lead to the development of new features and heuristics. This study also demarcates successful vs unsuccessful identifications as well as providing the scoring functions for doing so.

4.5 Limitations

The most prominent limitation this project encountered is time. With additional time to work, the suggested improvements could have been added to possibly improve the program. Having more time to fine-tune thresholding and peak picking can allow a higher accuracy. Acknowledging that in a real scenario MS/MS will always be suboptimal, it still needs to be mentioned that some MS/MS are not able to be identified because of their lack of information quality, which definitely affected the final identification score. An Additional drawback is the complexity of the topic of mass spectrometry and the limited time given to gather expertise.

Chapter 5

Conclusion

5.1 Summary

The thesis set out to explore a rule-based expert-system approach to *de novo* sequencing, and while the program had a low Top 1 accuracy (2.25%), as a proof of concept, this is a successful implementation. To the extent of my knowledge, there are no distance matrix applications similar to the one presented here mentioned in the literature. Furthermore, even with the program's current drawbacks, this approach has potential. When suggested improvements get implemented, an increase in accuracy is to be expected. Meaning, ultimately, this system could be a baseline for a future AI-driven system for optimized spectra annotation.

5.2 Future Work

The next steps for this project would be to continue to aim for improving the accuracy of the model. Improvement in the scoring metrics is needed, to a point where they reliably show the real peptide as the best fit. Improving upon the current scoring metrics by finetuning them as well as implementing complementarity from the Tabb et al., 2008 paper is the first step. Additionally, an improved peak picking method is needed as the current one omits too much information. Working on these two steps, as well as identifying new scoring metrics, possibly based more on biochemical validity, could yield an increases accuracy score. With this, a rule-based expert system *de novo* sequencing program could be integrated into a bigger context. An AI-driven system using a large language model could be trained to use the logical rule set adapted from the manual *de novo* sequencing methods, whose baseline ideas were presented

here. With additional integration of contextual information, accurate, large-scale, and human-readable annotations of tandem mass spectra can be achieved.

Bibliography

Al-Amrani, S., Al-Jabri, Z., Al-Zaabi, A., Alshekaili, J., & Al-Khabori, M. (2021). Proteomics: Concepts and applications in human medicine. *World Journal of Biological Chemistry*, *12*(5), 57–69. https://doi.org/10.4331/wjbc.v12.i5.57

Armenian, P., Vo, K. T., Barr-Walker, J., & Lynch, K. L. (2018). Fentanyl, fentanyl analogs and novel synthetic opioids: A comprehensive review. *Neuropharmacology*, *134*, 121–132. https://doi.org/10.1016/j.neuropharm.2017.10.016

Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M.-Y., Paulse, C., Creasy, D., ... Mallick, P. (2012). A cross-platform toolkit for mass spectrometry and proteomics [Publisher: Nature Publishing Group]. *Nature Biotechnology*, *30*(10), 918–920. https://doi.org/10.1038/nbt.2377

Chi, H., Chen, H., He, K., Wu, L., Yang, B., Sun, R.-X., Liu, J., Zeng, W.-F., Song, C.-Q., He, S.-M., & Dong, M.-Q. (2013). pNovo+: De novo peptide sequencing using complementary HCD and ETD tandem mass spectra [Publisher: American Chemical Society]. *Journal of Proteome Research*, *12*(2), 615–625. https://doi.org/10.1021/pr3006843

Frank, A., & Pevzner, P. (2005). PepNovo: de novo peptide sequencing via probabilistic network modeling [Publisher: American Chemical Society]. *Analytical Chemistry*, *77*(4), 964–973. https://doi.org/10.1021/ac048788h

Jannetto, P. J., & Fitzgerald, R. L. (2016). Effective use of mass spectrometry in the clinical laboratory. *Clinical Chemistry*, *62*(1), 92–98. https://doi.org/10.1373/clinchem.2015.248146

Jeong, K., Kim, S., & Pevzner, P. A. (2013). UniNovo: A universal tool for de novo peptide sequencing. *Bioinformatics*, *29*(16), 1953–1962. https://doi.org/10.1093/bioinformatics/btt338

Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., & Nesvizhskii, A. I. (2017). MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometrybased proteomics [Publisher: Nature Publishing Group]. *Nature Methods*, *14*(5), 513–520. https://doi.org/10.1038/nmeth.4256

Liu, K., Ye, Y., Li, S., & Tang, H. (2023). Accurate de novo peptide sequencing using fully convolutional neural networks. *Nature Communications*, *14*(1), 1–11. https://doi.org/10.1038/s41467-023-43010-x

Loponte, H. F., Zheng, J., Ding, Y., Oliveira, I. A., Basse, K., Todeschini, A. R., Horvatovich, P. L., & Lageveen-Kammeijer, G. S. M. (2025). GlycoGenius: The ultimate high-throughput glycan composition identification tool [Preprint]. *bioRxiv*. https://doi.org/10.1101/2025.03.10.642485

Ma, B. (2015). Novor: Real-time peptide de novo sequencing software. *Journal of the American Society for Mass Spectrometry*, *26*(11), 1885–1894. https://doi.org/10.1007/s13361-015-1204-0

Paizs, B., & Suhai, S. (2005). Fragmentation pathways of protonated peptides [_eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/mas.20024]. *Mass Spectrometry Reviews*, *24*(4), 508–548. https://doi.org/10.1002/mas.20024

Qiao, R., Tran, N. H., Xin, L., Chen, X., Li, M., Shan, B., & Ghodsi, A. (2021). Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices [Publisher: Nature Publishing Group]. *Nature Machine Intelligence*, *3*(5), 420–425. https://doi.org/10.1038/s42256-021-00304-3

Rost, H. L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., Andreotti, S., Ehrlich, H.-C., Gutenbrunner, P., Kenar, E., et al. (2016). Openms: A flexible open-source software platform for mass spectrometry data analysis. *Nature methods*, *13*(9), 741–748.

Shao, C., Zhang, Y., & Sun, W. (2014). Statistical characterization of HCD fragmentation patterns of tryptic peptides on an LTQ orbitrap velos mass spectrometer. *Journal of Proteomics*, *109*, 26–37. https://doi.org/10.1016/j.jprot.2014.06.012

Sinha, A., & Mann, M. (2020). A beginners guide to mass spectrometrybased proteomics. *The Biochemist*, *42*(5), 64–69. https://doi.org/10.1042/BIO20200057

Sturm, M., & Kohlbacher, O. (2009). TOPPView: An open-source viewer for mass spectrometry data [Publisher: American Chemical Society]. *Journal of Proteome Research*, *8*(7), 3760–3763. https://doi.org/10.1021/pr900171m

Tabb, D. L. (2015). The SEQUEST family tree. *Journal of the American Society for Mass Spectrometry*, *26*(11), 1814–1819. https://doi.org/10.1007/s13361-015-1201-3

Tabb, D. L., Ma, Z.-Q., Martin, D. B., Ham, A.-J. L., & Chambers, M. C. (2008). DirecTag: Accurate sequence tags from peptide MS/MS through statistical scoring. *Journal of proteome research*, *7*(9), 3838–3846. https://doi.org/10.1021/pr800154p

Tan, C. F., Wahidin, L., Khalil, S. N., Tamaldin, N., Hu, J., & Rauterberg, M. (2016). The application of expert system: A review of research and applications. *11*, 2448–2453.

Taylor, J. A., & Johnson, R. S. (2001). Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Analytical Chemistry*, *73*(11), 2594–2604. https://doi.org/10.1021/ac001196o

Toby, T. K., Fornelli, L., & Kelleher, N. L. (2016). Progress in top-down proteomics and the analysis of proteoforms. *Annual review of analytical chemistry (Palo Alto, Calif.)*, *9*(1), 499–519. https://doi.org/10.1146/annurev-anchem-071015-041550

Tran, N. H., Zhang, X., Xin, L., Shan, B., & Li, M. (2017). De novo peptide sequencing by deep learning [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *114*(31), 8247–8252. https://doi.org/10.1073/pnas.1705691114

Trifonova, O. P., Maslov, D. L., Balashova, E. E., & Lokhov, P. G. (2021). Mass spectrometry-based metabolomics diagnostics – myth or reality? *Expert Review of Proteomics*, *18*(1), 7–12. https://doi.org/10.1080/14789450.2021.1893695

Tully, B., Balleine, R. L., Hains, P. G., Zhong, Q., Reddel, R. R., & Robinson, P. J. (2019). Addressing the challenges of high-throughput cancer tissue proteomics for clinical application: ProCan. *Proteomics*, *19*(21-22), 1900109. https://doi.org/10.1002/pmic.201900109

Urban, J., Jin, C., Thomsson, K. A., Karlsson, N. G., Ives, C. M., Fadda, E., & Bojar, D. (2024). Predicting glycan structure from tandem mass spectrometry via deep learning. *Nature Methods*, *21*(7), 1206–1215. https://doi.org/10.1038/s41592-024-02314-6

Urbiola-Salvador, V., Jaboska, A., Miroszewska, D., Kamysz, W., Duzowska, K., Drek-Chya, K., Baber, R., Thieme, R., Gockel, I., Zdrenka, M., rutek, E., Szylberg, ., Jankowski, M., Baa, D., Zegarski, W., Nowikiewicz, T., Makarewicz, W., Adamczyk, A., Am-

bicka, A., ... Chen, Z. (2024). Mass spectrometry proteomics characterization of plasma biomarkers for colorectal cancer associated with inflammation. *Biomarker Insights*, *19*, 11772719241257739. https://doi.org/10.1177/11772719241257739

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wedd, N. (n.d.). Tandem peptide spectra [[Online; accessed 12 June 2025]]. https://www.weddslist.com/ms/tandem.html

Wehr, T. (2006, September 1). Top-down versus bottom-up approaches in proteomics [Publisher: MJH Life Sciences]. Retrieved July 7, 2025, from https://www.chromatographyonline.com/view/top-down-versus-bottom-approaches-proteomics-0

Appendices

Appendix A

Ethical Declaration

By submitting this thesis/ uploading this document to Brightspace, I certify:

- That I am the author of this document
- That nothing was taken from other sources without proper references
- That this document is the result of my own discussions and preparations
- that I have not used AI tools except when this was part of the research question

Appendix B

Use of AI tools

AI-assisted tools during the writing and coding process were used in limited and defined ways. All well within the restrictions on usage. Language models (such as ChatGPT) were used to help generate and correct BibTeX citation entries, including resolving syntax errors and formatting tables within the BibTeX and LaTeX file. Additionally, AI assistance was used to help with finding initial thesis topics and gathering information early on. It was used for some applications in coding, mainly to help with improving runtime, exploring implementation options, and understanding the use of the Habrok supercomputer. Basic writing tools like Grammarly were used to help with punctuation and comma placement.

Below are a few example prompts that were given to ChatGPT:

- "How can I cite ProteoWizard msConvert. And if you know how can you give it to me in BibTeX format?"
- "I want to represent intensity rank sum as a mathematical formula in LaTeX. How can I do that?"
- "How do I apply a function to a dataframe that requires 2 arguments?"
- "What text do I have to use again to activate my environment, and can you explain how it works?"