Automated pipeline for isotopic charge states identification in Mass Spectrometry data

.

Konstantinos Pantelidis

Supervisor: Konstantinos Thalassinos
MSc Bioinformatics with Systems Biology
Date: 25/08/2021

ABSTRACT

Understanding the interactions, dynamics and structure of the proteins provides a greater level of comprehension of how a protein works, allowing us to create hypotheses about how to affect it, control it, or modify it.
Mass Spectrometry has been a great analytical technique to study the proteome in the past years and provides a lot of information and data via Mass Spectrum results.
This project focuses on the charge states identification in Mass Spectrum fingerprints and the distinction between true peaks and background noise, that Electrospray Ionisation leaves behind, through an automatic pipeline development instead of a visual inspection.
An algorithm implemented in Python programming language is created, calculates the mass to charge distances and identifies isotopic windows in the whole dataset. It currently searches for +2, +3 and +4 window charge states and uses an intensity threshold at 1.5%, which can be modified by the user according to their needs or dataset complexity and volume.
The results provide the user with the number of isotopic windows for each charge state in the selected dataset as well as creates .txt files that could be used as labelled training datasets for the creation of a deep learning artificial network.

Table of Contents

INTRODUCTION

The central dogma of molecular biology describes the flow of information starting at the genomic level (DNA) to the intermediate mRNA to the final functional product, the protein. Proteins are the essential molecules that regulate and participate in all biological processes[1].

Studying the genome and monitoring changes in DNA, albeit important for understanding the causes of disease in some cases, does not always reflect changes on protein levels and even though all cells of a multicellular organism have the same set of genes, the set of proteins, or **proteome,** differs significantly[2]. Moreover, since proteins orchestrate cellular structure and functions such as metabolism, gene regulation, protein synthesis etc., understanding the interactions, dynamics and structure of proteins has always been a high priority in biomedical research[3].

As mentioned, the proteome is a dynamic entity. The set of proteins present in different cells and tissues varies according to the gene expression[4]. The 3D structure of a protein actively regulates its proper function (or lack thereof) and as such changes in protein spatial conformation can be the primary cause of diseases.
Therefore, studying protein structure is pivotal not only for understanding cellular functions but also for the development of therapies e.g., custom drug design.
The study of all proteins expressed by a given genome and their different properties e.g., sequence, post-translational modifications, protein-protein interactions, at any given time is defined as **proteomics**.

The development of the field of proteomics highlighted the need for tools which could complement genomics and help scientists test their hypotheses that were initially based on gene analysis.
Peptide Mass Fingerprint (PMF), an analytical technique for protein identification which was developed in the 90's, introduced us to the idea that only the masses of the peptides have to be known and that instead of the unnecessary laborious de novo peptide sequencing, the unknown peptide masses are compared to a *database* containing known protein sequences[5].

Amongst the toolkit of techniques for analysing proteins on a large scale, **Mass Spectrometry** (**MS**) has gained popularity because of its ability to handle the complexities associated with the proteome[6].

What is Mass Spectrometry?

Mass spectrometry is a technique that measures and analyses the ratio between the mass and the charge of ionised molecules (*m/z*, mass to charge ratio). In fact, there cannot be mass spectrometry without ionisation[7]. In order to produce the requested ionised molecules, we rely on a bottom-up proteomics workflow.

All proteins from a sample of interest are separated from the cell or tissue according to their size and then are subjected to enzymatic digestion into peptides, typically using trypsin[8].

Subsequently the generated peptides are separated using liquid chromatography and then ionised by being eluted into an electrospray ion source (ionizer) which is used to electrically charge the atoms or molecules converting them into ions.
The ions then travel through a mass analyzer and arrive at different parts of the detector according to their mass/charge (m/z) ratio[9].

Electrospray Ionisation (ESI)

The most sensitive, robust, and reliable ionisation tool over the past years is **Electrospray Ionisation (ESI-MS)** [9, 20]. ESI was first reported by Masamichi Yamashita and John Fenn in 1984, as a need to overcome the propensity of the analyte fragmentation that previous ionization methods (e.g., electron ionization, chemical ionization) could not be able to. This practically means that, generally, no fragmentation occurs upon molecule ionization and thus, among others (e.g., MALDI), it is called a *soft ionisation* technique. Soon, it became indispensable to precisely measure the molecular mass of the biologically important supramolecules like proteins.

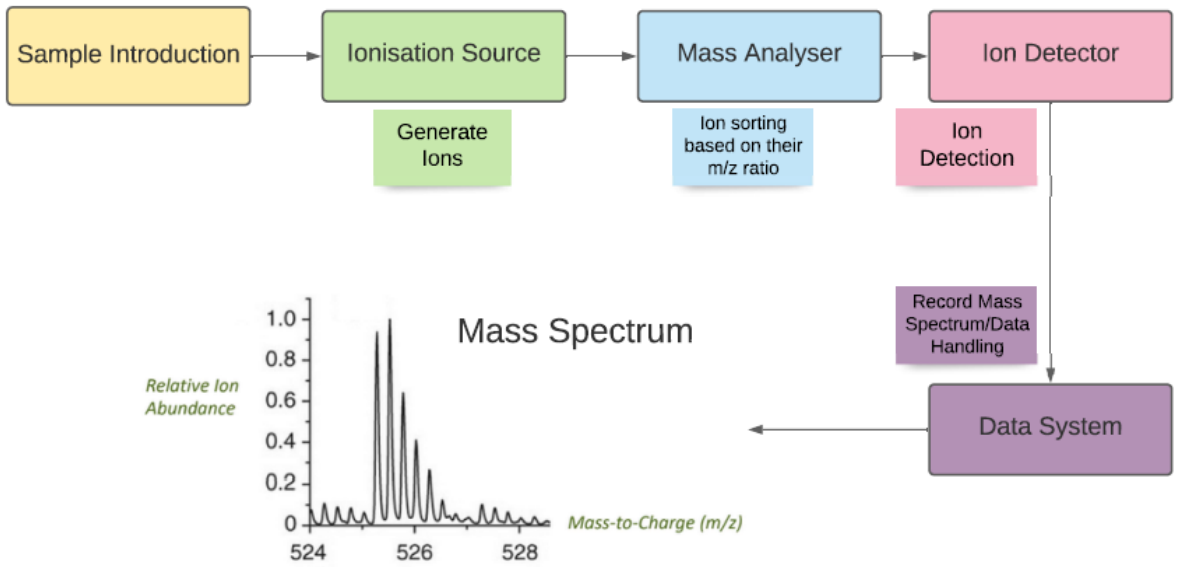


Figure 1 | Components of a Mass Spectrometer. A Mass Spectrometer consists of three components: an ion source, a mass analyzer, and a detector. It shall always perform the showing processes to represent the generated signals into a graph.

After the ionisation, it's time for the ions to travel through the mass analyser in order to get the signals. There are several variations of mass-spectrometry analyses such as quadrupole (Q), ion trap (quadrupole (QIT) or linear (LIT) ion trap), time-of-flight (TOF) etc.

Mass spectrometry in its essence aims to give information on the structure of the protein by measuring its mass and charge. However very little structural information can be gained from the simple mass spectrum obtained.
This can be overcome by coupling ESI with **Tandem Mass Spectrometry (ESI-MS/MS)**, an analytical technique in which two or more analysers are coupled together using an additional

reaction step[21], causing ions in the first mass spectrum (graph representation) to fragment, so that the secondary mass spectrum can reveal new information and increase the chance to analyse chemical samples [9,10].

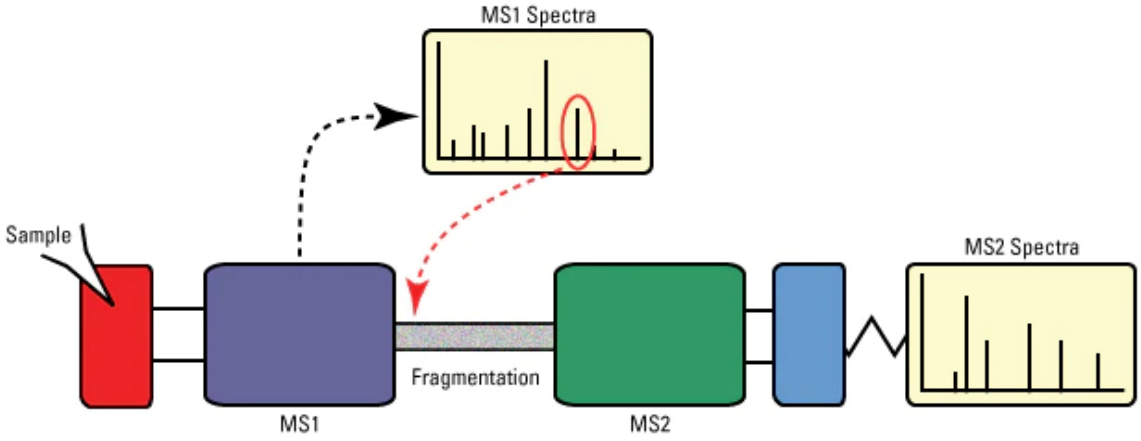


Figure 2 | Tandem Mass Spectrometry Diagram (MS/MS). A sample is injected into the mass spectrometer, ionized, accelerated and analyzed by mass spectrometry (MS1). Ions from the MS1 spectra are then selectively fragmented and analyzed by a second stage of mass spectrometry (MS2) to generate the spectra for the ion fragments. While the diagram indicates separate mass analyzers (MS1 and MS2), some instruments utilize a single mass analyzer for both rounds of MS. (www.thermofisher.com)

Mass Spectrum

The generated signals are collected and separated based on their *m/z* ratio and are represented in a two-dimensional axis, the Mass Spectrum, the most common representation of the mass spec data.
In a Mass Spectrum plot *(Fig.3)*, the horizontal axis represents the *m/z* ratio of ions that have been introduced to the mass analyser while the vertical axis defines the relative ion abundance (intensity). The profile that results from a mass-spec analysis of a specific protein is unique and serves as a fingerprint for this particular molecule [7,11]. Changes in this fingerprint e.g., upon treatment or in disease, represent structural changes that may be important for the functionality of the protein and being able to monitor or measure them can improve our understanding of biological processes or facilitate the development of therapeutic compounds e.g., in case of a disease.

The fragment ions will always have lower mass than the total molecular ion, which usually lies on the far-right end of the graph as the ion with the highest *m/z* ratio.
The point whereby the signal intensity of a specific *m/z* is higher than the background noise is defined as a peak. The tallest peak, known as base peak, defines the most intense and therefore dominant fragment ion to be formed which is set to 100%, serving as a reference point, and all the other ions have abundances normalised respectfully to this peak[11].
In Mass spectrometry, noise can be defined as the fluctuation in the background signal and can be classified as chemical or electronic. Chemical noise is caused by the presence of chemicals other than the analyte in the sample, while electronic noise is more complex and

may be caused by different factors such as temperature or the electronic circuitry that is associated with the equipment [11,12].

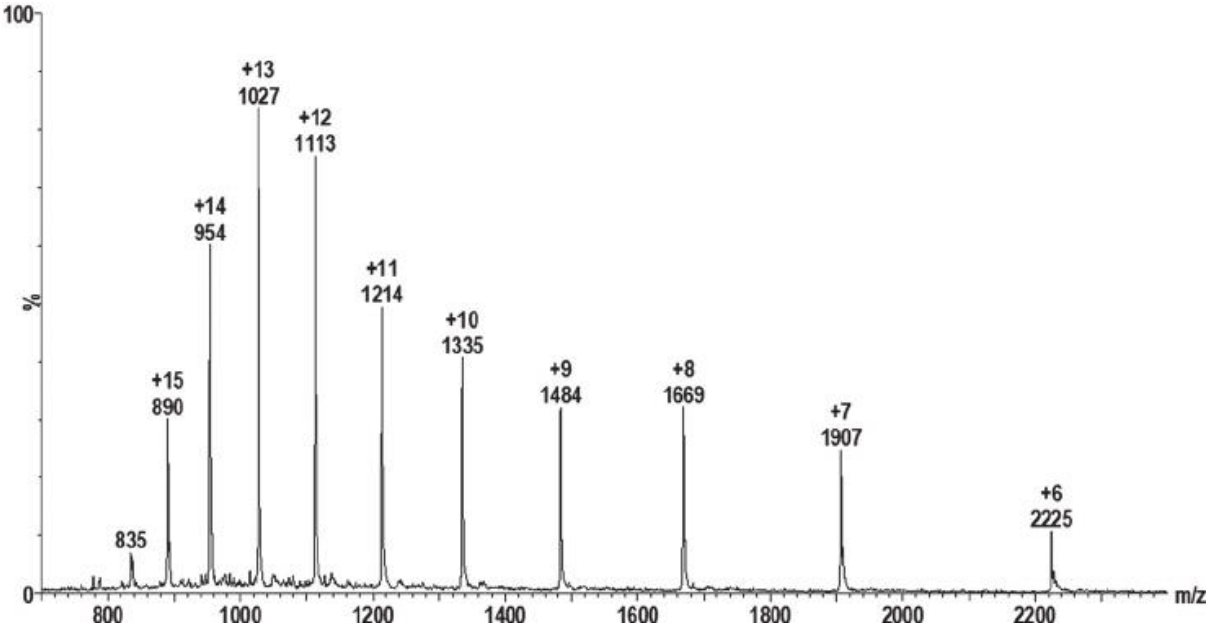


Figure 3 | Mass Spectrum example. Electrospray Ionisation for a protein to show the multiply charged species. The m/z is presented on the x-axis while the y-axis shows the normalised intensity. The different peaks show different charge states[19].

Isotopes

The analyte (ionised protein sample) contains ions of slightly different masses, known as isotopes. In fact, isotopes are atoms that have the same number of protons but different number of neutrons. The term isotope is derived from Ancient Greek root *iso-* ("same") and -*tope* from the word *topos* (means "place") meaning the *"same place"*, thus the meaning of the word defines the "same place" that the different isotopes of the same element occupy in the periodic table[13].

Though, because isotopes contain different numbers of neutrons, each isotope has a unique atomic mass. For example, although there are three isotopes of carbon incorporating six, seven, or eight neutrons yielding ^{12}C, ^{13}C, and ^{14}C, respectively, they all have six protons[11].

Since many elements display unique isotope patterns, the mass analyte can contain isotopic forms of a given element which are represented in the Mass Spectrum as different isotopic peaks. We can think of the Mass Spectra as a distribution of peaks for each ion, corresponding to its different isotopes. Being able to accurately identify the isotopes and their associated patterns in a sample is essential for the correct determination of the sample composition.

Isotopes and abundances

Some isotopes are more naturally abundant on Earth than others, therefore, different isotopes have different relative abundances [7,11]. Relative abundances of all the different isotopes of an

element add up to 100%. In addition, these abundances/intensities can be determined using mass spectrometry and displayed in a Mass Spectrum sample.
This is useful from a physics/chemistry point of view but not from a biological point of view where isotopes are not found alone but in high complexity molecular structures[14].

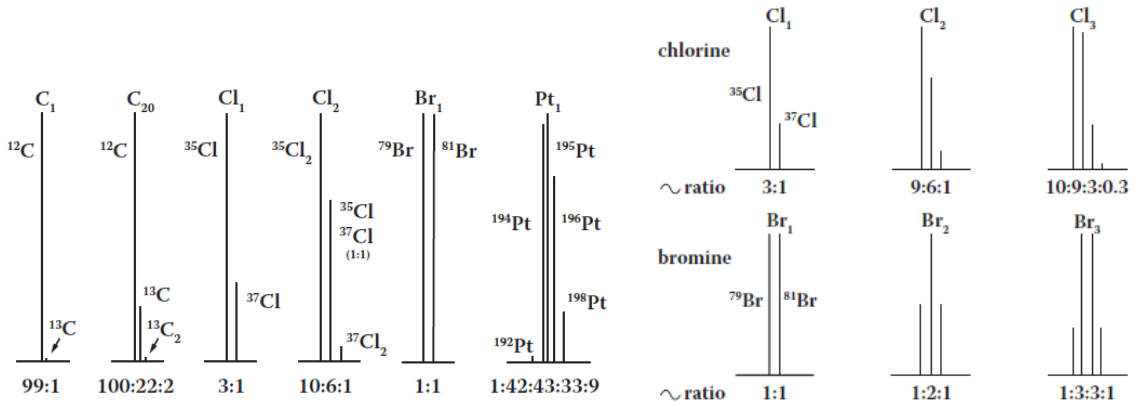


Figure 4 | Isotope patterns examples. The natural abundances of the isotopes of some elements lead to the appearance of unique, distinctive isotope patterns in the spectrum. (John Greaves, John Roboz, Mass Spectrometry for the Novice, 2014, ISBN: 97814200941

Charge states

During the Electrospray Ionisation (ESI) process, the structures of large molecules have multiple sites where charges can be added, and this results in ions with different charge states.
The isotopic peaks in a Mass Spectrum can provide useful information about the different charge states which can be observed at their respective m/z values on the x-axis.
In fact, two adjacent ions shall give two simultaneous equations for the calculation of the mass as well as the charge of a molecule.

As an example, let an ion with a m/z ratio = 980 *(eq. 1)* has an adjacent peak of 891 m/z.
It is also known that the adjacent peak must have an additional charge *(z+1)* on the initial ion *(eq. 2)*.

Example:

$$\frac{m}{z} = 980 \Rightarrow m = 980 \times z \qquad (1)$$

$$\frac{m}{z+1} = 891 \qquad (2)$$

Therefore, the two equations solve first the charge *(eq. 3)* and then the mass of the ion *(eq. 1)*

$$(1), (2) \Rightarrow 980 \times z = 891 \times (z+1) \Rightarrow z \approx 10.011 \Rightarrow z \approx 10 \qquad (3)$$

$$(1), (3) \Rightarrow m = 9800$$

Since the principle behind measuring the charge states is that as isotopes are separated from one another by 1 "mass" (m), for each ionic species we can identify the main isotope *(m)* and its species as *m+1, m+2* and so on and therefore in order to measure the m/z ratio for each species we need to divide the distance on the x axis (m/z ratio) by 1, 2, 3 etc[15].
Following this process, we can predict that isotopes separated by 0.5 m/z from the main isotope will have a charge state of 2, separation of approximately 0.33 translates to charge state of 3 and separation of 0.25 will be a charge state of 4 (double, triple and quadruple charged ions respectfully)[11,16].

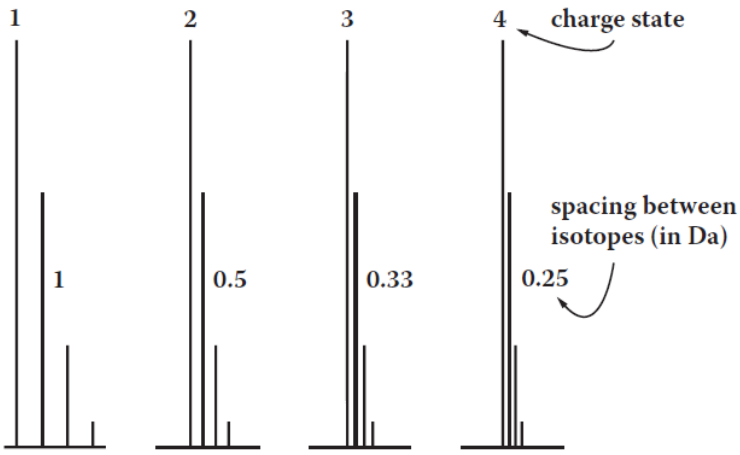


Figure 5 | Components of a Mass Spectrometer. A Mass Spectrometer consists of three components: an ion source, a mass analyzer, and a detector. It shall always perform the showing processes in order to represent the generated signals into a graph. (John Greaves, John Roboz, Mass Spectrometry for the Novice, 2014, ISBN: 9781420094183)

Aim

Knowing the structure of a protein finds applications in areas beyond basic biology such as medicine and chemistry e.g., allowing customised drug design for the treatment of diseases. In addition, following up changes over time or under different conditions e.g., drug treatment, temperature changes etc are important to understand the 3D conformation of a protein and by extent its function in cellular systems.

Creating a protein "mass fingerprints" database, allows scientists to match proteins and peptides in databases with the unknown peptide targets[17], rather than time-consuming peptide sequencing.
During the peptide mass fingerprint, the protein sequence must be present in the database of interest which leads us to the fact that updating this "database" constantly, by adding new peptide fingerprints, will eventually give us the advantage of having more peptides and protein sequences to compare with the new unknown ones.

Because of ESI, we get numerous charge states which is the raw signal that is needed to process in order to identify the precursor mass of a peptide.
In addition, in Mass Spectrometry, not all the peaks can be identified. Due to the large scale

of raw data, virtual identification of charge states and distances can sometimes be paradoxical and as a result, the charge state recognition would trouble the mass identification of a peptide. As a result, there is a lot of background signal, also known as *noise* which makes the true signal identification complicated.

"A major unexplored opportunity for the application of deep learning with proteomics is for automating the interpretation of large omics datasets" [18].

The aim of this work is to develop a pipeline for the automatic extraction of the various charge states that are found from Mass Spectrum profiles so that we can get and annotate as much information as possible from the available data and differentiate the true signal from the background noise.
As deep learning methods exist for nearly every aspect of the modern proteomics workflow, this will eventually lead to a trained labelled dataset which will be useful for an automatic identification with machine-learning strategies through a deep neural network[18].

METHODS AND MATERIALS

The data files used were provided by Dr. Thalassinos, in .txt file format. The data are a subset of an unpublished project and have been uploaded in PRIDE and held privately at the moment.
The files contain Mass Spectra results in a Tabular form, composed of thousands of data points, corresponding to different m/z and their intensities. The Mass Spectra results are peptides or parts of peptides of different proteins, and they were used for code testing and development.
The data extraction, manipulation, analysis, and visualisation as well as result analysis were all made using Python programming language, Python version 3.7.4.
A Python Environment management system which includes Pandas, NumPy, Matplotlib, Seaborn, Sci-kit Learn packages downloaded as a requirement.
Then, a sharing notebook was created using Jupyter Notebook for documentation, commenting, and editing the code, which can be found online here:
https://github.com/KPantelidis/MSc-Bioinformatics-Project

RESULTS

EXPLORATORY DATA ANALYSIS

What does my data look like?

An exploratory plot representation of the data is a necessary preliminary analysis in order to have an overview of the data structure e.g., peak formation. Using a simple matplotlib scatter plot we not only notice some peaks, but we can also make conclusions on the amount of noise the data has. *(Fig.6|B)* The m/z is presented on the x-axis and their signal intensities on the y-axis.

As mentioned in the introduction section, a mass spectrum is the plot representation of the acquired *m/z* ratios and intensities. However, tabular form is another type of data, and it is the one being used for the purposes of this pipeline development. The data consists of two columns, one depicting the *m/z* ratio for all data points in ascending order and another containing the signal intensity for each specific *m/z* ratio. The data is sorted by increased m/z from the first column. *(Fig.6|A).*

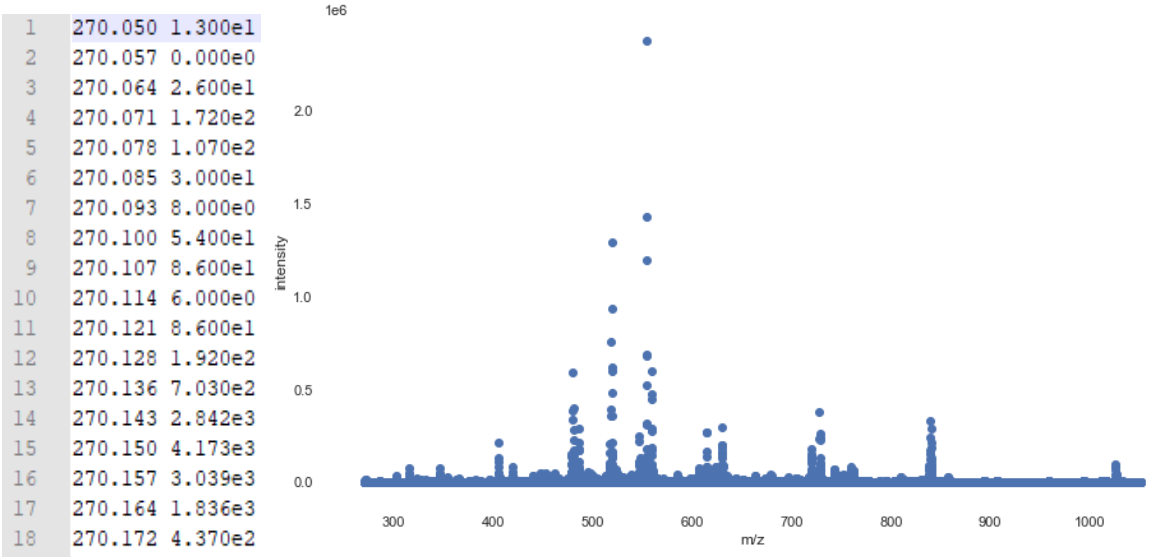


Figure 6 | Exploratory Data Analysis of raw data. A. The first eighteen lines of the raw data .txt files; The m/z of each peak is present on the first column with their intensities on the second one in exponential form. **B.** The raw data presented in a scatter plot with the m/z on the x-axis and their signal intensities in exponential form on the y-axis.

Converting the data into a uniform format suitable for downstream analysis

Since in a Mass Spectrum the most intense fragment ion is set to 100% and all the other peaks are normalised respectfully to it, we work similarly for our data *(Fig.7)*. In order to normalise the data, the method *minmax_scale* is imported from *SciKit-Learn.* In addition, notice how the data is now sorted by the intensity., which will be useful and will be explained in the next steps.

A

	m/z	intensity
28617	553.800	1.000000
28618	553.811	0.601602
25206	519.303	0.543845
28666	554.304	0.503373
25207	519.313	0.391948
25154	518.786	0.318592
28616	553.790	0.287816
28665	554.294	0.287099
25257	519.811	0.259275

B

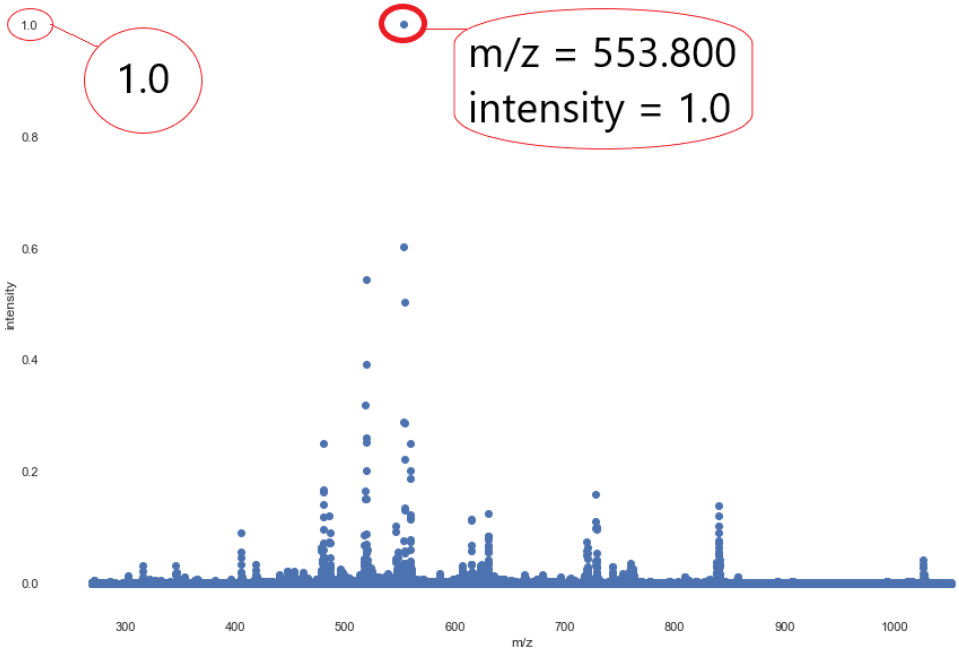


Figure 7 | Graph representation of normalised data by intensity. A. The normalised data is now sorted by intensity in the second column. **B**. The scatter plot of the normalised data; The peak with the highest intensity (1.0) is highlighted.

Minimum Intensity - Noise threshold

Biological systems carry by definition a high level of background (noise) therefore in order to make the subsequent analysis more straightforward it is useful to set a threshold for minimum intensity. To help this decision, several tests that will be discussed more extensively in a later section, took place with results showing that the data points are starting to get clear by setting a minimum intensity at 0.015 or 1.5%.

In the next plot *(Fig.8)* one may notice how the data had been cleaned on the bottom of the plot, very close to 0. A comparison with *(Fig.7|A)* is also suggested.

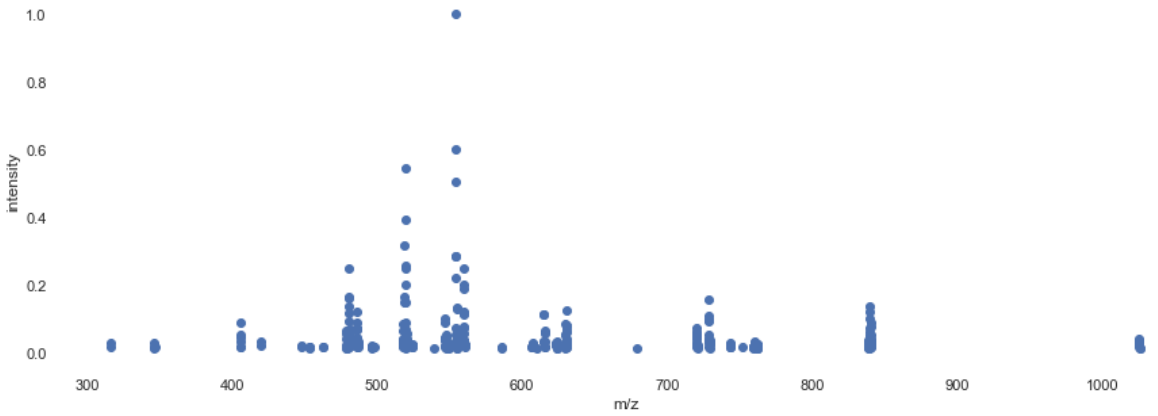


Figure 8 | Plot representation of normalised data with a threshold applied in the intensity. A scatter plot data graph without the background noise. The minimum intensity is set at 1.5%.

Focusing on a window as an example *(Fig.9)*, having eliminated the noise, we can identify the peak with the highest intensity as well as its isotopes. Noticing the distances between the peaks we can make a preliminary conclusion on the charge state of this isotopic window too. We are now able to let the algorithm identify the charge state of each isotopic window by checking the distances automatically instead of visual inspection.

In the next plot *(Fig.9),* there is an example of an isotopic window where we notice 17 clear data points, after setting the threshold. The data points look like they are grouped in four sets, which are actually the four different peaks on this window. One can tell by visual inspection that there is a rough 0.5m/z distance between the grouped data points sets/peaks, and as a result, it would be a sensible conclusion that this is a charge state +2 window.

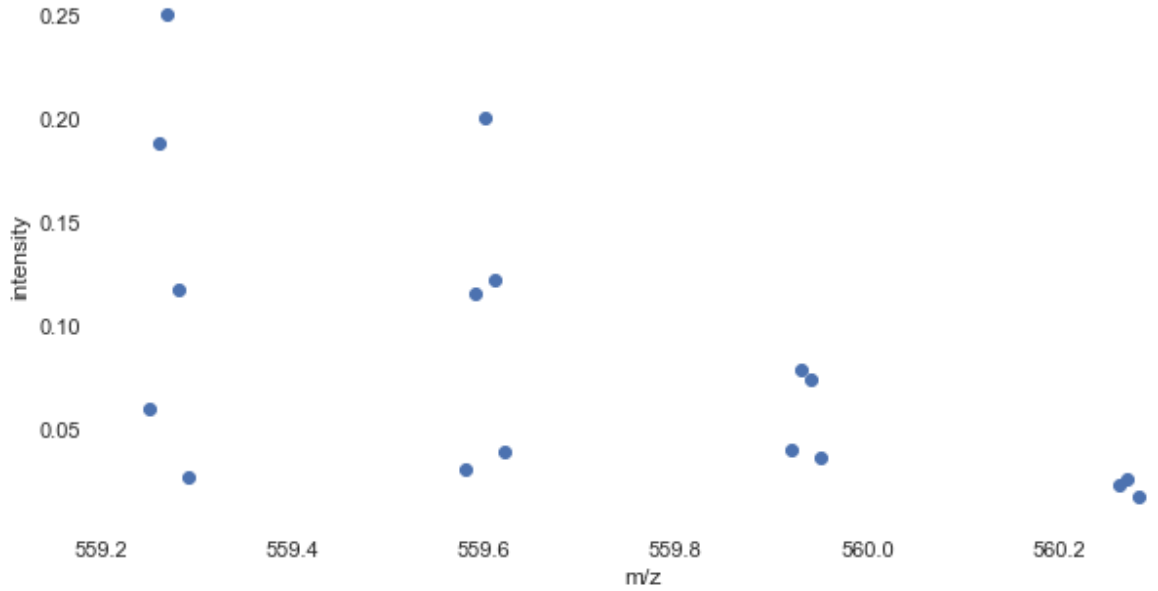


Figure 9 | Example of an isotopic window. The grouped peaks have ~0.5m/z distance which means that this is a charge state +2 isotopic window.

ALGORITHM

Main Idea

The main idea for the algorithm development is to work in one specific isotopic window per time, extracting results and extracting information on their charge state before moving to the next one.
In each window, the algorithm searches for the highest peak (highest intensity) and sets it as the main one (point of reference) and then moves in both directions (right and left) looking for peaks at specified distances. The distances used for the purposes of this thesis are of 0.5 or 0.33 or 0.25 stating double, triple and quadruple charged ions respectively. However, the algorithm offers the flexibility to be modified and search for different charge states (+5, +6 etc.) depending on the need of the analysis.
Once a peak is found, the algorithm keeps a record of it in a list and sets this new peak as the main one. The same process is followed again until no more peaks within the specified distances are found for this window.
This procedure is followed for both directions and the end result is a .txt file having two columns for m/z and their intensities, exactly as described earlier for the initial data files, but this time only including the data points of this specific window which is also labelled with its charge state.
Since the noise is filtered out in the pre-processing step, the data depict only the main peak and its isotopes. *(Fig10|C)*. It is important to note that the background noise that has been filtered in the beginning may affect the peak determination in the downstream analysis and it would therefore be useful to examine the peaks whilst maintaining the noise in the same plot *(Fig.10|D)*. Such a plot gives an indication whether the baseline (noise) has been correctly determined or further adaptations are necessary e.g., not too high so as to obscure peak selection or charge state identification.
As a result, both file and plot include the initial noise that had been cut as well as the important peaks.

In the next figure, the algorithmic steps for the window charge stage identification are explained thoroughly.

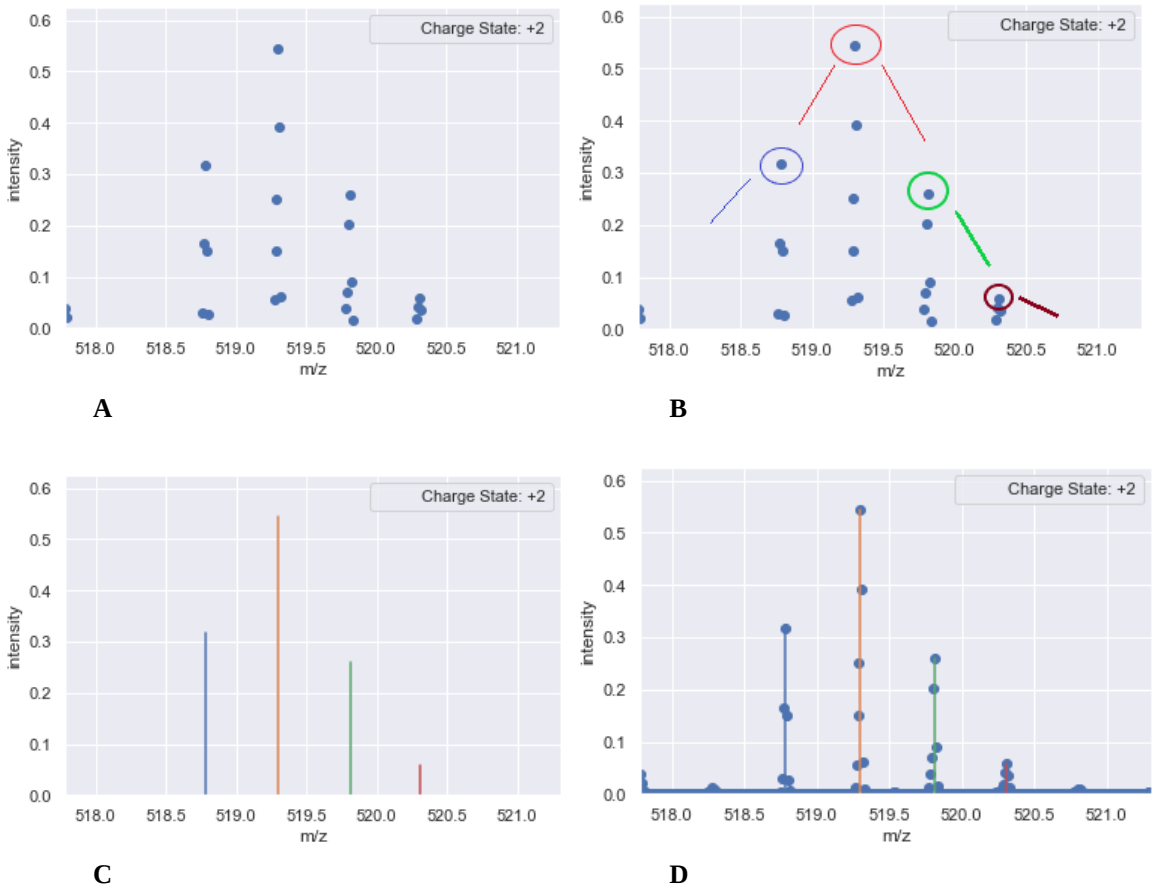


Figure 10 | Steps of Window Charge State Identification. A. Four grouped peaks are found after the noise cut. **B.** The algorithm finds the highest and starts searching left and right for the next highest in the requested distances. Since it finds it, the algorithm does the same process now starting from the new highest peak/point of reference, until there are no other high peaks left or right. **C.** The peaks that are found are all saved and then shown without the rest unwanted peaks. **D.** The plot shows the requested peak as well as all the others in the specific window, even the ones that were cut in the very first beginning.

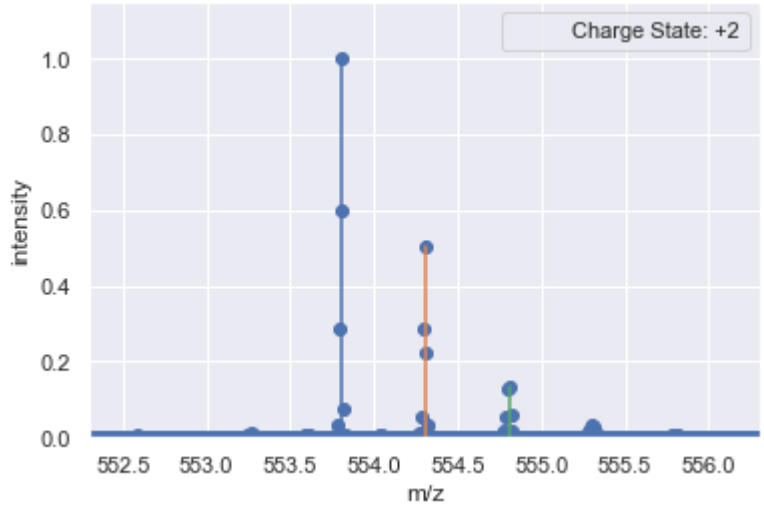
It's important to highlight that although the exact distances from the main peak are 0.5, 0.33 and 0.25, the actual experimental data show slight deviations from these numbers due to technical experimental reasons and therefore it is important to take this into account when determining the distances to look at from the main selected peak. To this end, the distance has been set from 0.47 to 0.53 for charge 2+, 0.30-0.36 for charge 3+ and 0.22-0.28 for charge +4, giving the algorithm the chance to identify peaks with greater accuracy. A decision which will also be discussed extensively in a later section.

Identifying Isotopic Windows complication

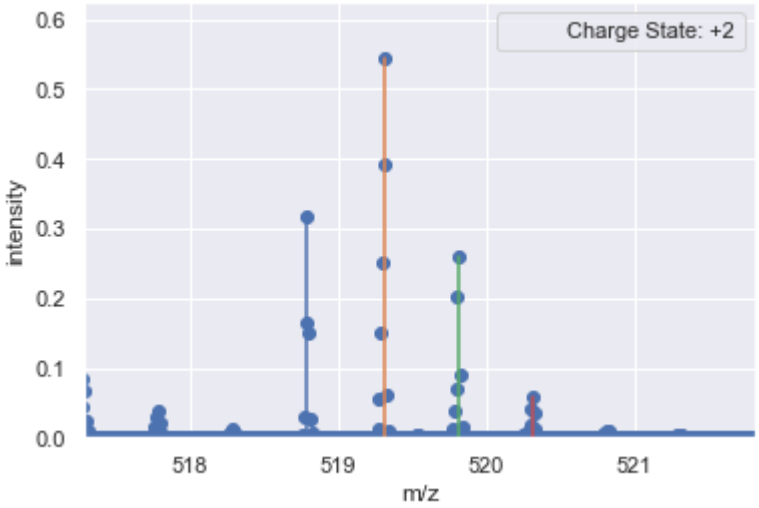
The challenge of this approach is the extraction of charge state information from a dataset with many isotopic windows.
Since the raw data file runs from the smallest m/z to the largest, or left to right, if we consider a plot like the one suggested above, we can see that the start and end of each window is not clearly identified and therefore identifying the peak with the highest intensity is not possible.

The second column that shows the intensity for each m/z is the one that the algorithm uses to identify the requested peaks, therefore, this is the reason that the normalisation and sorting out the data by intensity ascending order is the very first step in the process.
Following the intensity normalisation, it is easier for the algorithm to determine which is the peak with the highest abundance which is then set as the main peak (point of reference) to begin determining the charge states from.
As a next step, the algorithm performs the peak searching as described above, sets and names the window as a charge state +2, +3 or +4 and creates two values, *max_del* and *min_del*, which are the 0.5 m/z and -0.5 m/z distances respectfully from the highest and lowest values in the m/z-sorted list. These values are also used to set the x-axis limits when creating the plot.
Moreover, these *max_del* and *min_del* values are used in order to delete the currently analysed MS-window once the information on charge state and intensities has been extracted, and at the same time to create a new data set excluding the window that we have already examined. This step has a key role in the algorithm flow as it can now identify a new highest peak and set it as the new point of reference and follow the same procedure for this newly identified window and peak.

In the next plot, notice how the intensity reduces from 1.0 *(Fig 11/A)* to approximately 0.55 *(Fig 11/B)*. This is due to the fact that the first identified isotopic window is not present in the dataset when the second search begins, thus, the highest peak/point of reference in the second search is lower than 1.0. More isotopic windows can be found in the Appendix section.



A



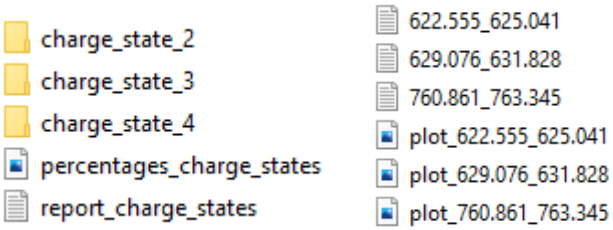
B

Figure 11 | Identified isotopic windows for the same dataset The results after two searches of the algorithm where the intensity cut is set at 1.5%.

ANALYSIS

Output - File naming and sorting into different directories

As mentioned, prior to every new search the data is sorted and saved/exported to a .txt file as described earlier. In addition, the algorithm creates a directory that includes three more directories, one for each charge state *(Fig 12A).* Then, each tabular-data-type file as well as its own plot are sorted in the appropriate directory regarding their charge state *(Fig 12B)*
The name of the files, as decided after discussion, describe the first and the last m/z of this specific isotopic window. In that way, one can know which isotopic window and area of the peptide is referring to.



A **B**

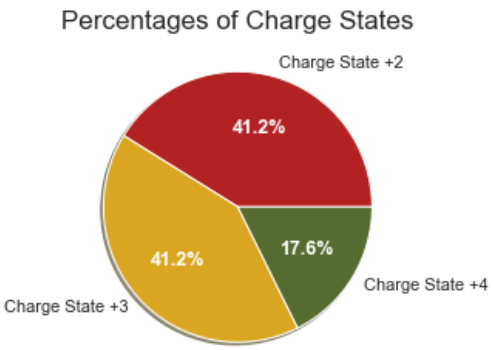
Figure 12 | Output - Directories description - Files sorting - Files naming. A. Three directories are created by the algorithm and each isotopic window result is sorted by its own charge state. **B.** An example of the files in the charge state directory. Here, there are three isotopic windows and for each one there is a .txt file including the information that was described earlier in addition to its own plot.

Output - Creating Reports

After the algorithm runs all the available windows and sorts out the files with their plots, two statistical report files (a .txt file and a pie chart .png file), which show the number and percentage of the different charge states in the whole initial data file, are created *(Fig 13A).*

```
REPORT FOR THE SELECTED FILE ;
Noise cut at 1.5%

There were found 17 windows in total
Specifically :

Charge_State +2: 7
Charge_State +3: 7
Charge_State +4: 3

with respectful percentages:

Charge_State +2: 41.1764705882353
Charge_State +3: 41.1764705882353
Charge_State +4: 17.647058823529413
```

A



B

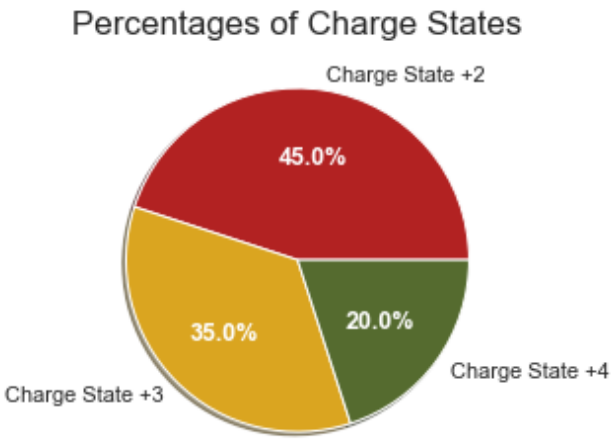
Figure 13 | Output -Creating Reports - Examples. A. An example of a .txt report created after the analysis of all the isotopic windows where the number and the percentages of all different charge state windows are shown. **B.** An example of a pie report of the percentages of the isotopic windows that are found after the algorithm process corresponding to A.

Intensity cut - Noise threshold

Since the minimum intensity cut is one of the first steps in the algorithm process, it is obvious that the less the noise threshold, the more the identifying windows. With even small changes, the number of windows the algorithm finds, and saves can be double, triple or more and as a result, sometimes there are differences in the percentages of the charge states in the whole file.

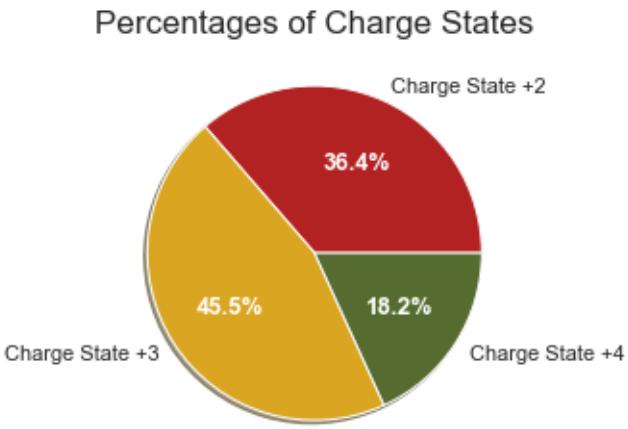
Examples of the same data file where the intensity cut is set at 0.01 and 0.02 in the next figure.
Although the percentages do not change drastically, the number of windows is roughly the half when the minimum intensity increased from 1% to 2%.

```
REPORT FOR THE SELECTED FILE ;
Noise cut at 1%

There were found 20 windows in total
Specifically :

Charge_State +2: 9
Charge_State +3: 7
Charge_State +4: 4

with respectful percentages:

Charge_State +2: 45.0
Charge_State +3: 35.0
Charge_State +4: 20.0
```



A

```
REPORT FOR THE SELECTED FILE ;
Noise cut at 2%

There were found 11 windows in total
Specifically :

Charge_State +2: 4
Charge_State +3: 5
Charge_State +4: 2

with respectful percentages:

Charge_State +2: 36.36363636363637
Charge_State +3: 45.45454545454545
Charge_State +4: 18.181818181818183
```

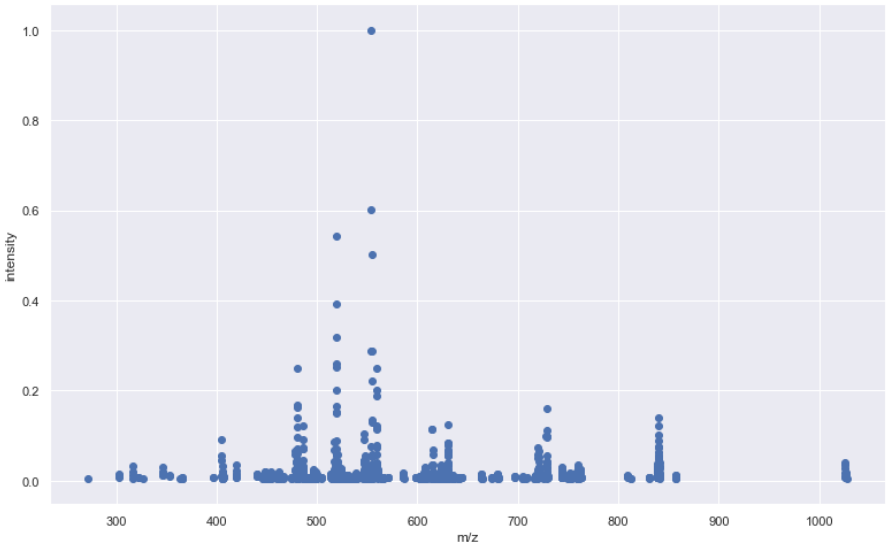


B

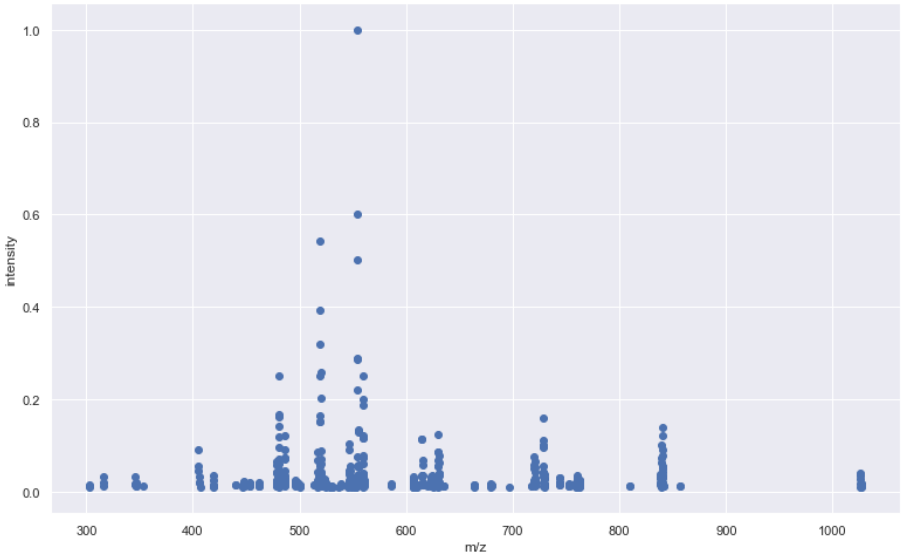
Figure 14 | Noise cut - Examples. A. The .txt report file with its relevant pie chart percentages of different charge states in the selected file. Noise threshold at 1% **B.** The same selected file but the minimum intensity has now been set at 2%

In addition to the reports, differences can be observed in the scatter plot of the data after the noise cut. And although we might not see big differences between 1% or 2% noise cut *(Fig. 15A, 15B)*, a perceptible change can be observed when we compare these to a 0.5% or 5% minimum intensity selection.

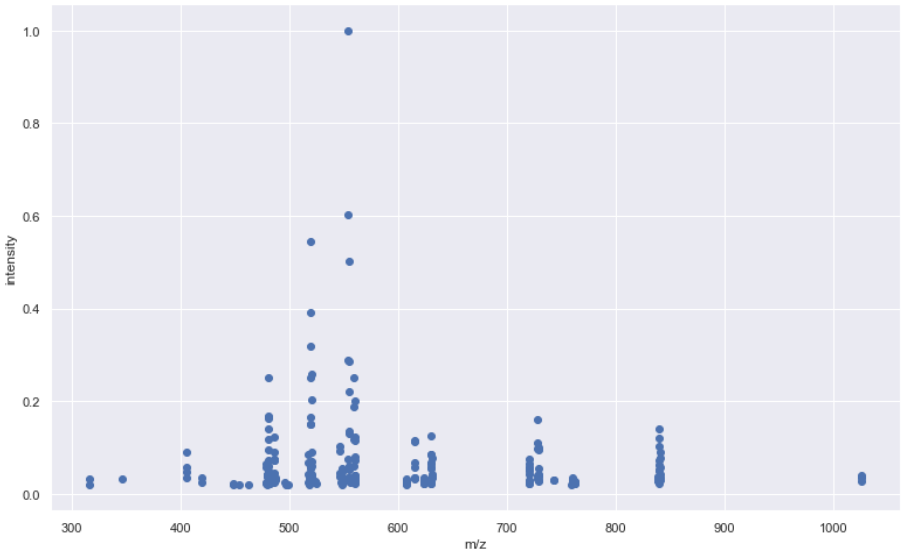
In the next figures, notice how the peaks are starting to “fade” and eventually disappear in the bottom of the plot as we increase the minimum cut of the intensity.


intensity
m/z
1.0
0.8
0.6
0.4
0.2
0.0
300
400
500
600
700
800
900
1000

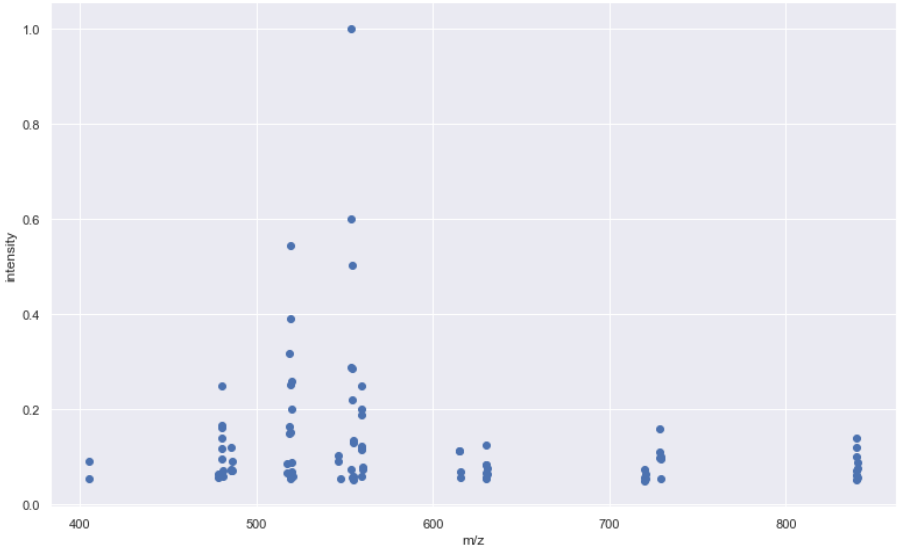
A


intensity
m/z
1.0
0.8
0.6
0.4
0.2
0.0
300
400
500
600
700
800
900
1000

B


intensity
m/z
1.0
0.8
0.6
0.4
0.2
0.0
300
400
500
600
700
800
900
1000

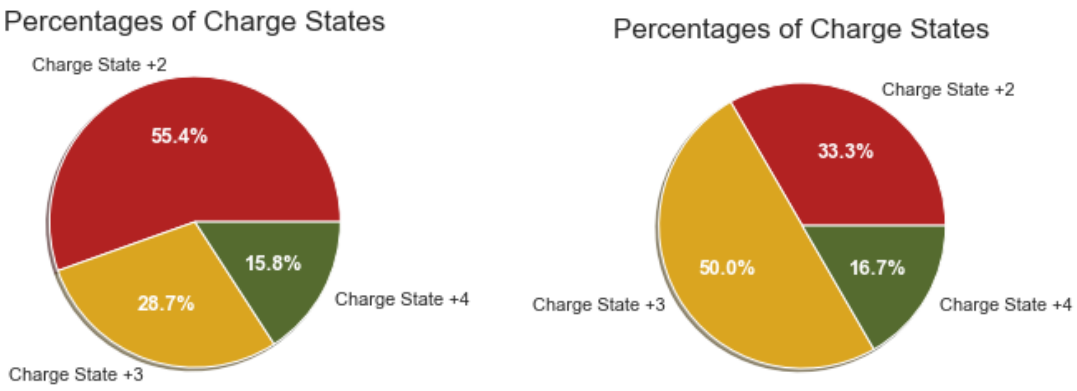
C



D

Figure 15 | Plot representation of the same data set with different minimum intensity cuts. A.B.C.D. The noise cut is set at 0.5%, 1%, 2% and 5% respectfully.

It is also interesting to notice the window percentages when the difference in the minimum intensity is big. The smaller the minimum intensity, the more the background noise in the dataset and the more difficult for the algorithm to identify true peaks.



A **B**

Figure 16 | Pie chart of the percentages of the charge states in the same data set with different minimum intensity cuts. A. The noise cut is set at 0.25% **B.** The noise threshold is at 5%

DISCUSSION

The project is focusing on identifying the charge state of an isotopic window without a visual inspection. One may focus on a specific area of the data set and easily make conclusions on the charge state of a window just by running the code in the selected area which would be time-consuming and unorthodox. A method that searches the whole data set and not specific areas, one by one, should be applied for time saving.

What characterises this algorithm special is the part where each window is deleted from the data set after its identification and a new algorithmic search begins for the next highest peak and as a result a new window. Since the previous highest peaks are not present in the data set anymore, the algorithm is now given a new highest peak - a new point of reference - and makes it capable for a new search and a new isotopic window identification without a manual order from the user.

After running the algorithm in the very first attempts, there have been noticed peaks which although were very close to the exact distances of 0.5, 0.33 and 0.25 for the charges +2, +3, and +4, were not identified by the algorithm. Due to technical experimental reasons, it is common to notice slight deviations and as a result, for greater accuracy, there is an adaptation to the algorithm to investigate a range of numbers. After testing the algorithm with different selected ranges, a +/- 0.03m/z to the initial exact searches was decided and as a result the ranges are 0.47 to 0.53 for charge 2+, 0.30-0.36 for charge 3+ and 0.22-0.28 for charge +4. Furthermore, thanks to the code adaptability, the user can alter the number ranges according to the dataset needs.

A main mass spectrometry challenge is to manage identifying all the peaks in a mass spectrum if possible. Especially in data sets that are extracted from large peptides or proteins, distinguishing the true peaks from the background noise is vital.
One of the main decisions one needs to take when running the algorithm for a selected data file is the noise threshold. As shown in a previous section, small changes in the minimum intensity selection affect the number of isotopic windows that are extracted.
After testing different noise cuts in several data files, the noise threshold is advised to be set between 1-2%. For code testing and developing during this project, the noise threshold is set at 1.5%. Plots can be found in the Appendix section. However, an exploratory data analysis is advised in the beginning of the search as the data set may vary and as a result the minimum intensity threshold needs an adaptation.

As mentioned, this project is aiming to the identification of true peaks and charge state labelled isotopic windows. On top of that, the creation of a good training labelled dataset could give opportunities for machine learning applications and eventually a deep artificial neural network. The .txt file results can be used to train a model and eventually create a database of known isotopic window fingerprints.

REFERENCES

[1] Cobb M (2017) 60 years ago, Francis Crick changed the logic of biology. PLoS Biol 15(9): e2003243.

[2] Cho WC. Proteomics technologies and challenges. *Genomics Proteomics Bioinformatics*. 2007;5(2):77-85. doi:10.1016/S1672-0229(07)60018-7

[3] Breda A, Valadares NF, Norberto de Souza O, et al. Protein Structure, Modelling and Applications. 2006 May 1 [Updated 2007 Sep 14]

[4] Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition.

[5] Karl R. Clauser, Peter Baker, and Alma L. Burlingame. Analytical Chemistry 1999 *71* (14), 2871-2882. DOI: 10.1021/ac9810516

[6] Han X, Aslanian A, Yates JR 3rd. Mass spectrometry for proteomics. Curr Opin Chem Biol. 2008 Oct;12(5):483-90. doi: 10.1016/j.cbpa.2008.07.024. PMID: 18718552; PMCID: PMC2642903.

[7] Thalassinos, K., Nobeli, I. MSc Bioinformatics with Systems Biology, Systems Biology Module Material, Proteomics I. 2018.

[8] Schmidt, A., Forne, I. & Imhof, A. Bioinformatic analysis of proteomics data. *BMC Syst Biol* 8, S3 (2014).

[9] Ho CS, Lam CW, Chan MH, et al. Electrospray ionisation mass spectrometry: principles and clinical applications. Clin Biochem Rev. 2003;24(1):3-12.

[10] Shibdas Banerjee, Shyamalava Mazumdar, "Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte", International Journal of Analytical Chemistry, vol. 2012, Article ID 282574, 40 pages, 2012.

[11] John Greaves, John Roboz, Mass Spectrometry for the Novice, 2014, ISBN: 9781420094183, Published September 18, 2013 by CRC Press

[12] Wong CC, Cociorva D, Venable JD, Xu T, Yates JR 3rd. Comparison of different signal thresholds on data dependent sampling in Orbitrap and LTQ mass spectrometry for the identification of peptides and proteins in complex mixtures. J Am Soc Mass Spectrom. 2009;20(8):1405-1414. doi:10.1016/j.jasms.2009.04.007

[13] Soddy, Frederick (12 December 1922). "The origins of the conceptions of isotopes". Nobelprize.org. p. 393. Retrieved 9 January 2019.

[14] von Korff, M., Sander, T. Molecular Complexity Calculated by Fractal Dimension. Sci Rep 9, 967 (2019). https://doi.org/10.1038/s41598-018-37253-8

[15] Cardoza JD, Parikh JR, Ficarro SB, Marto JA. Mass spectrometry-based proteomics: qualitative identification to activity-based protein profiling. Wiley Interdiscip Rev Syst Biol Med. 2012;4(2):141-162. doi:10.1002/wsbm.166

[16] Yuan Z, Shi J, Lin W, Chen B, Wu FX. Features-based deisotoping method for tandem mass spectra. Adv Bioinformatics. 2011;2011:210805. doi: 10.1155/2011/210805. Epub 2012 Jan 4. PMID: 22262971; PMCID: PMC3259476.

[17] Cottrell JS. Protein identification by peptide mass fingerprinting. Pept Res. 1994 May-Jun;7(3):115-24. PMID: 8081066.

[18] Jesse G. Meyer, Deep learning neural network tools for proteomics, Cell Reports Methods, Volume 1, Issue 2, 2021, 100003, ISSN 2667-2375

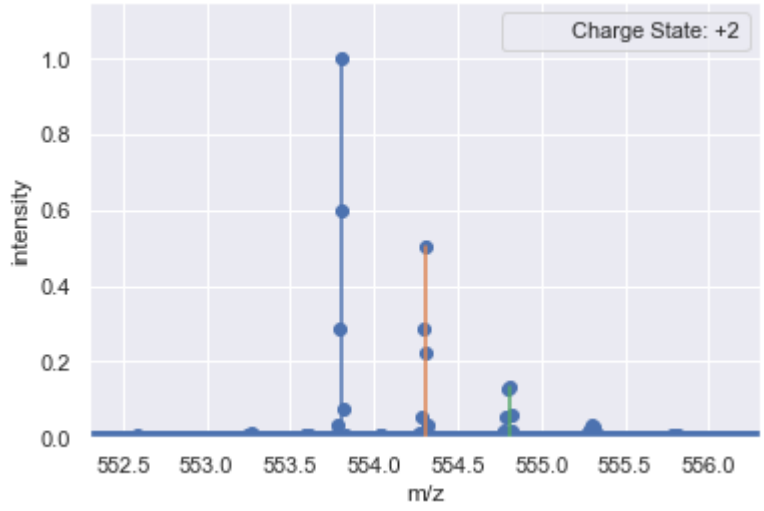
[19] Tyrefors, Niklas & Michelsen, Peter & Grubb, Anders. (2014). Two new types of assays to determine protein concentrations in biological fluids using mass spectrometry of intact proteins. Cystatin C in spinal fluid as an example.. Scandinavian journal of clinical and laboratory investigation. 74. 546-554. 10.3109/00365513.2014.917697.

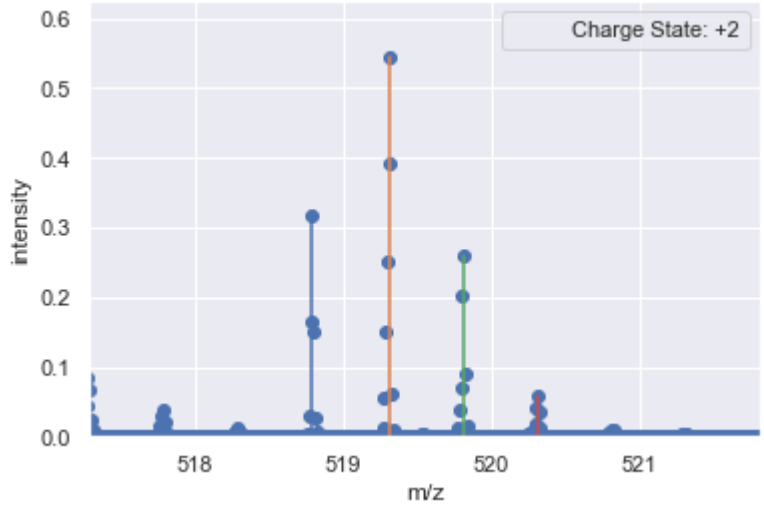
[20] Hui Liu, Jiyang Zhang, Hanchang Sun, Changming Xu, Yunping Zhu, Hongwei Xie, The Prediction of Peptide Charge States for Electrospray Ionization in Mass Spectrometry, Procedia Environmental Sciences, Volume 8, 2011, Pages 483-491, ISSN 1878-0296

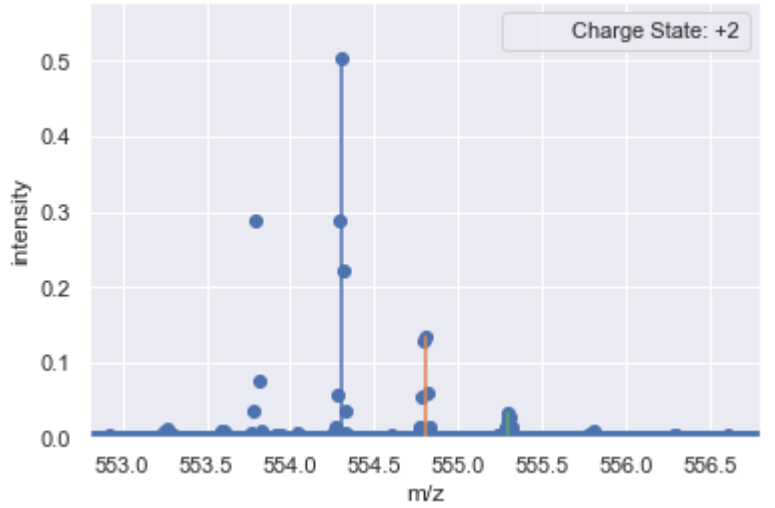
[21] Shi, J., Wu, FX. Peptide charge state determination of tandem mass spectra from low-resolution collision induced dissociation. Proteome Sci 9, S3 (2011).

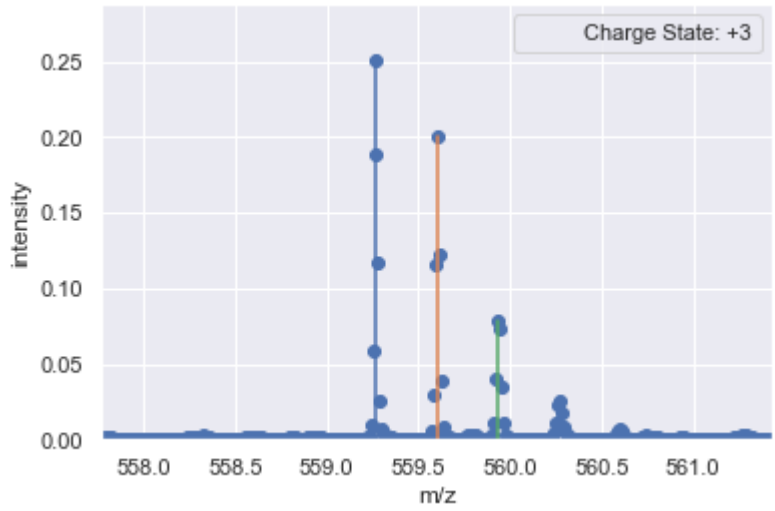
APPENDIX

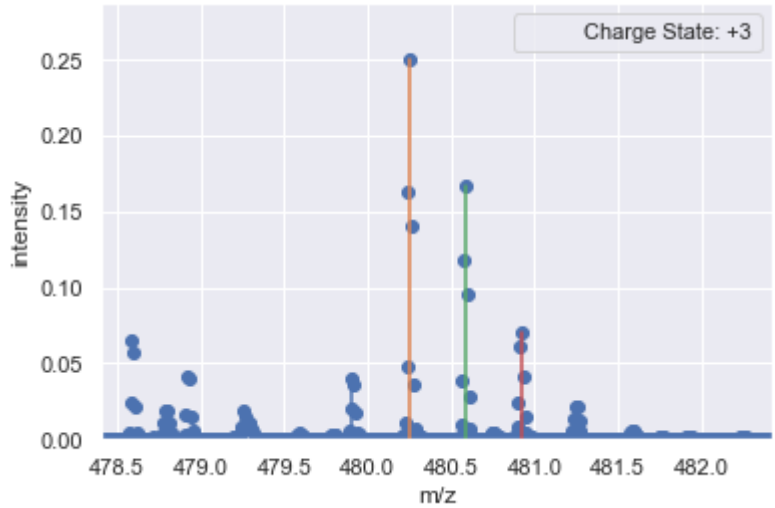
Plots examples of the first ten isotopic windows in the dataset selected for the purpose of this thesis with noise cut at 1.5%.

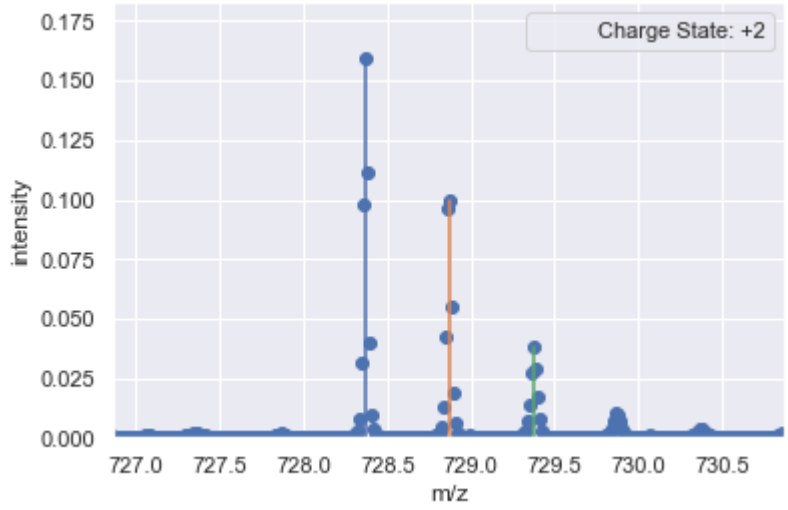


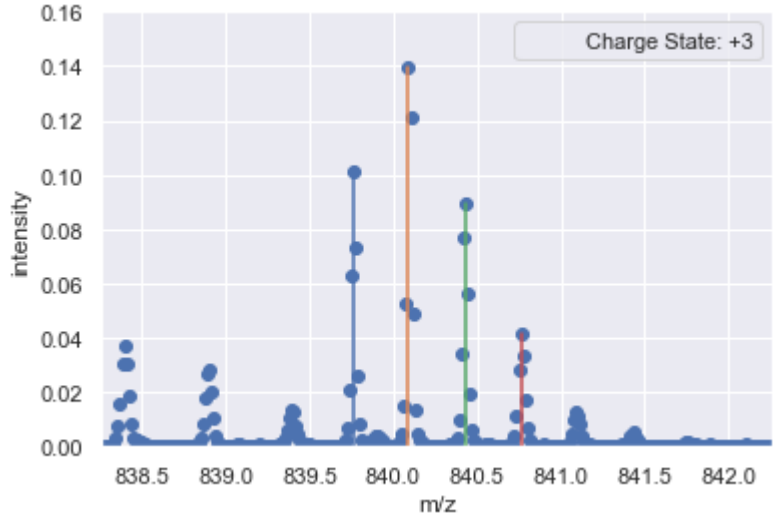


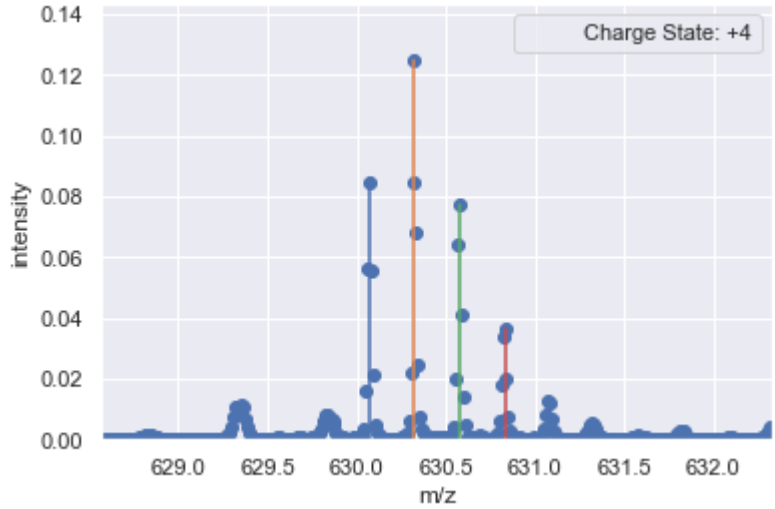

Charge State: +2
intensity
0.0
0.1
0.2
0.3
0.4
0.5
553.0
553.5
554.0
554.5
555.0
555.5
556.0
556.5
m/z

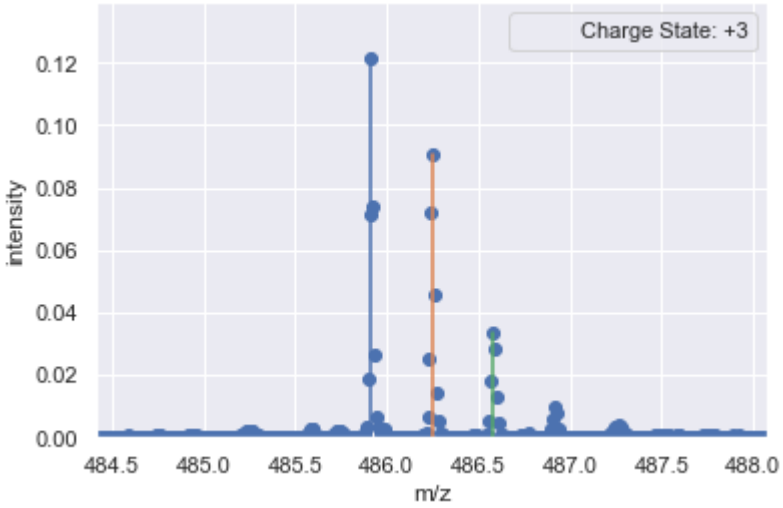

Charge State: +3
intensity
0.00
0.05
0.10
0.15
0.20
0.25
558.0
558.5
559.0
559.5
560.0
560.5
561.0
m/z

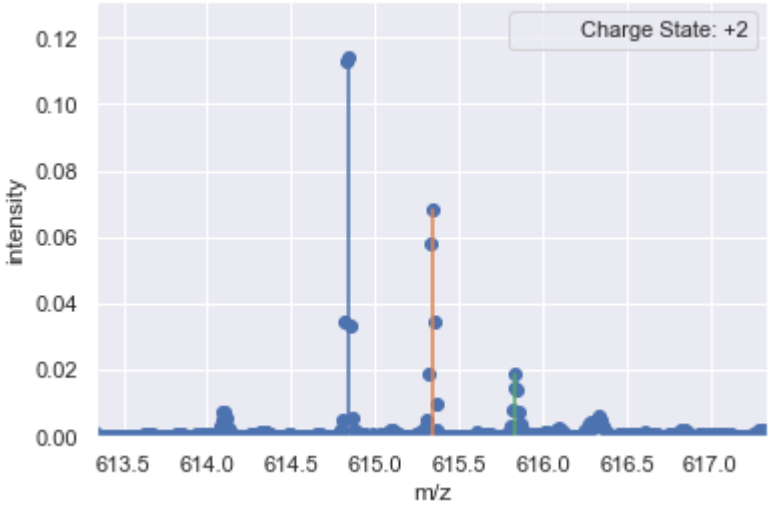

Charge State: +3
intensity
0.00
0.05
0.10
0.15
0.20
0.25
478.5
479.0
479.5
480.0
480.5
481.0
481.5
482.0
m/z


Charge State: +2
intensity
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
727.0
727.5
728.0
728.5
729.0
729.5
730.0
730.5
m/z


Charge State: +3
intensity
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
838.5
839.0
839.5
840.0
840.5
841.0
841.5
842.0
m/z


Charge State: +4
intensity
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
629.0
629.5
630.0
630.5
631.0
631.5
632.0
m/z


Charge State: +3
intensity
0.12
0.10
0.08
0.06
0.04
0.02
0.00
484.5
485.0
485.5
486.0
486.5
487.0
487.5
488.0
m/z


Charge State: +2
intensity
0.12
0.10
0.08
0.06
0.04
0.02
0.00
613.5
614.0
614.5
615.0
615.5
616.0
616.5
617.0
m/z