# Machine Learning in the Analysis of Mass Spectrometry Data

Matthew Mulvey

University of Auckland Summer Research Scholarship 2019/20

## Contents

**1 Introduction** **1**

**2 Methodology** **2**

**3 Results** **2**
3.1 binding_sites.py . . . . . . . . . . . . . . . . . . . . . . . . . . . 2
3.2 peak_labelling.py . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

**4 Discussion** **4**
4.1 binding_sites.py . . . . . . . . . . . . . . . . . . . . . . . . . . . 4
4.2 peak_labelling.py . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

**5 Conclusion** **4**
5.1 binding_sites.py . . . . . . . . . . . . . . . . . . . . . . . . . . . 4
5.2 peak_labelling.py . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

**6 Sources** **5**

## 1 Introduction

The primary motivation for this research project was based on increasing the inter-department collaboration between the School of Computer Science and the School of Chemical Sciences at The University of Auckland. I believe that the School of Computer Science can increase the efficiency and ease of research of its collaborative partners through the use of machine learning and automation techniques which may currently be underutilised in research not directly related to computer science. The particular field of research that this scholarship focused on was the analysis of mass spectrometry data in the study of platinum-based chemotherapeutic drugs.

## 2 Methodology

This project began with the examination of existing research papers which were closely related to the problem I was aiming to address. The papers *'Versatile Tool for the Analysis of Metal-Protein Interactions Reveals the Promiscuity of Metallodrug-Protein Interactions'* and it's associated online tool, and *'Characterization of Platinum Anticancer Drug Protein-Binding Sites Using a Top-Down Mass Spectrometric Approach'* were particularly useful in gaining an understanding of the problem at hand. The primary software package selected to facilitate the development of the Python tools I aimed to create was the pyOpenMS bindings for the OpenMS library. Once I was able to access the Bruker Data Analysis software used by the School of Chemical Sciences, I was able to convert the data they had provided into a more generally accessible file format (the xy file type). These .xy files are attached. The first part of the file names (demo, ub (Ubiquitin), c (Cisplatin), o (Oxaliplatin), or t (Transplatin)) refers to the folder in which the original file was located when received from Christian Hartinger, our primary point of contact in Chemistry. The first number in the file names, refers to the position of each file within its folder (when sorted alphabetically). The second number (where present) refers to the position of that spectrum in Data Analysis where files have more than one spectrum associated with them. For example, the file 'o_1_2.xy' would refer to the first file listed in the 'Ub+O' folder, and specifically the second spectrum listed in Data Analysis from that file.

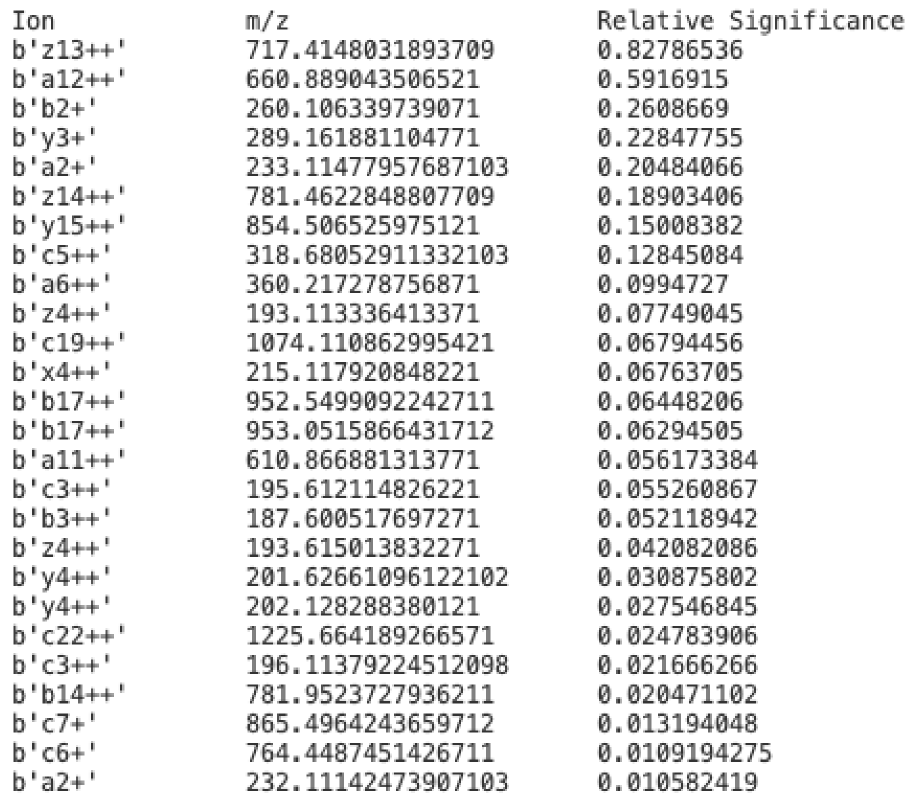
From there I was able to progress with development of the 'binding_sites.py' tool. I used the Python library Matplotlib for visualisation. To match experimental spectra to theoretical, I looked into dtw-python, its companion paper *'Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package'*, and fastdtw. Although I was unable to use dtw-python, its companion paper provided valuable insight into Dynamic Time Warping and how it can be applied to data.

In the later stages of the project a meeting with Christian provided invaluable feedback on the 'binding_sites.py' tool and significant insight into the type of software tool that would be most helpful to his work. This led to the development of the 'peak_labelling.py' tool. This tool was intended to allow mass spectrometry peaks to be automatically labelled on a visual graph, rather than the 'binding_sites.py' tool which was designed to find the specific location of platinum binding sites in ubiquitin.

## 3 Results

### 3.1 binding_sites.py

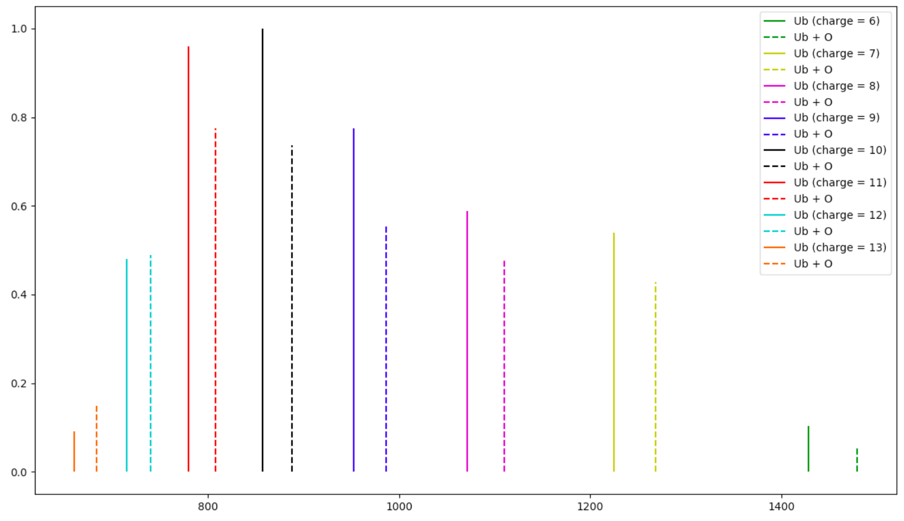
When run, this program prints out a list of fragments (in the Ion column) of Ubiquitin as shown in the image below. These are suspected to contain platinum binding sites. It also displays the mass to charge ratio of each fragment, and the relative significance of each fragment as a binding site predictor. This program requires two files of type xy in the same folder as itself. One should be a sample of ubiquitin with no binding (this filename should be specified on line 86). The other should be a sample of ubiquitin bound with either cisplatin, oxaliplatin, or transplatin (this filename should be specified on line 107).



```
Ion             m/z                     Relative Significance
b'z13++'        717.4148031893709       0.82786536
b'a12++'        660.889043506521        0.5916915
b'b2+'          260.106339739071        0.2608669
b'y3+'          289.161881104771        0.22847755
b'a2+'          233.11477957687103      0.20484066
b'z14++'        781.4622848807709       0.18903406
b'y15++'        854.506525975121        0.15008382
b'c5++'         318.68052911332103      0.12845084
b'a6++'         360.217278756871        0.0994727
b'z4++'         193.113336413371        0.07749045
b'c19++'        1074.110862995421       0.06794456
b'x4++'         215.117920848221        0.06763705
b'b17++'        952.5499092242711       0.06448206
b'b17++'        953.0515866431712       0.06294505
b'a11++'        610.866881313771        0.056173384
b'c3++'         195.612114826221        0.055260867
b'b3++'         187.600517697271        0.052118942
b'z4++'         193.615013832271        0.042082086
b'y4++'         201.62661096122102      0.030875802
b'y4++'         202.128288380121        0.027546845
b'c22++'        1225.664189266571       0.024783906
b'c3++'         196.11379224512098      0.021666266
b'b14++'        781.9523727936211       0.020471102
b'c7+'          865.4964243659712       0.013194048
b'c6+'          764.4487451426711       0.0109194275
b'a2+'          232.11142473907103      0.010582419
```

## 3.2 peak_labelling.py

This program requires a single input file in the xy format (specified on line 39). From that sample it is able to identify significant peaks representing ubiquitin which have charges of +1 to +20 (inclusive). For each ubiquitin peak identified, it is also able to identify the first adduct of that peak, which should represent the combination of ubiquitin, platinum, and the oxaliplatin ligand. All the identified peaks are then displayed visually, as shown in the image below. The x-axis represents mass to charge ratio, and the y-axis represents normalised intensity.



## 4 Discussion

### 4.1 binding_sites.py

As yet the results obtained from the binding sites program are not validated against reliable data. I suspect there are several sources of significant error, and the quality of results will likely improve once those are resolved. The most obvious point of potential improvement would be accounting for differences between MS1 and MS2 data. The program currently does not do this, and I believe it would allow the fragments of potential binding sites (some of which are currently quite large) to be identified at a higher resolution, giving a more precise prediction.

### 4.2 peak_labelling.py

The results from the peak labelling program have been quite promising. The peaks identified in the broadband sample of oxaliplatin (file o_1_1.xy) match those identified manually very closely. However, oxaliplatin files seem to be the easiest to produce results from, so further development may be needed to generalise the program to samples of cisplatin and transplatin.

## 5 Conclusion

### 5.1 binding_sites.py

At this stage, it seems that significant work would be needed to improve the binding sites tool to the point where it would produce reliable results. However, machine learning techniques may provide methods of increasing its performance significantly.

### 5.2 peak_labelling.py

The peak labelling tool serves as a useful proof of concept for something that could save chemistry researchers significant amounts of time by removing the need for manually identifying mass spectrometric peaks. The next logical improvement for anyone who wished to extend this project would be to make the tool more generalised. Currently it is only able to detect the two types of peaks mentioned above (Ubiquitin and Ubiquitin + Platinum + Oxaliplatin ligand). If it was modified to identify a wider range of proteins and drugs, or even identify platinated sites in MS2 data, it could be of great use to researchers in this area.

## 6 Sources

**1. Versatile Tool for the Analysis of Metal-Protein Interactions Reveals the Promiscuity of Metallodrug-Protein Interactions**
https://www.ncbi.nlm.nih.gov/pubmed/29053254

**2. Characterization of Platinum Anticancer Drug Protein-Binding Sites Using a Top-Down Mass Spectrometric Approach**
https://www.ncbi.nlm.nih.gov/pubmed/29053254

**3. OpenMS**
https://www.openms.de/

**4. Matplotlib**
https://matplotlib.org/

**5. dtw-python**
https://dynamictimewarping.github.io/

**6. Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package**
https://www.jstatsoft.org/article/view/v031i07

**7. fastdtw**
https://github.com/slaypni/fastdtw