This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Developing novel workflows for improved bottom-up, top-down and middle-down mass spectrometry analysis of proteins.

Lavrentis Dimitrios Galanopoulos

PhD in Chemistry
The University of Edinburgh
2021
Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

Lavrentis Dimitrios Galanopoulos
Abstract

Mass spectrometry (MS) has been extensively used to analyse biological samples and has evolved into an essential tool for proteomics research. Recent tremendous technological advancements, including improving instrument accuracy, resolution and sensitivity, and more freely accessible protein databases, have led to the growing importance of this technique. This study contributes to this growing area of research by developing novel orthogonal workflows that push the boundary of mass spectrometry capabilities, and allowed greater insight into the protein environment. Specifically, this research describes new methodology for de novo characterisation of disulfide bonding connectivity and de novo protein sequencing.

Investigating protein 3D structure can provide essential information when addressing the issues in protein folding and function. One of the important structural features in proteins is the disulfide links between cysteine residue pairs, which play crucial roles in sustaining protein 3D structure. Although some conventional experimental techniques can provide information of the disulfide patterns within a specific proteoform, there has been little quantitative analysis due to cost and time limitations. Thus, there is a necessity for new approaches that increase confidence in disulfide mapping and maximise the protein sequence coverage obtained. Here, this study draws attention to the ability of a combination of pepsin and trypsin proteolysis and the usage of dual fragmentation of electron capture dissociation (ECD) and collision induced dissociation (CID) to assigning disulfide connectivity of proteins and maximise sequence coverage. This protease approach is based on the accurate mass measurement of proteins and high-resolution top-down fragmentation MS studies. Here we present our findings, which confirm that the developed method has significant advantages.

Another main challenge in mass spectrometry-based proteomics is de novo protein sequencing, especially for novel proteins such as monoclonal antibodies for which genome information is often limited or not available. However, due to limitations in peptide fragmentation coupled with coverage and ambiguities in spectra interpretation, complete de novo assembly of unknown protein sequences remains challenging. Thus, there is a drive for new strategies that increase fragment ion assignment efficiency in top-down mass spectra and maximise the protein sequence coverage obtained. Here, we use a strategy for selective chemical labelling of the protein N-terminus using reductive alkylation and utilise this chemistry to introduce a halogen-based mass defect tag to the N-terminus of Insulin B chain, Ubiquitin, Myoglobin and Rnase A. We outline the potential advantages of using this simple chemical derivatisation in top-down de novo protein sequencing.
Lay Summary

Proteins are unique entities of natural organisms. One crucial step to understanding proteoforms is to determine their structure. Because by understanding the higher-level structure of proteins enables one to better understand the biophysical interactions within the cell. Thus, the aim of this work was to develop new analytical tools for the investigation of protein structure.

Since each protein's sequence and disulfide pattern are unique, to at least some degree, it helps distinguish proteins. Furthermore, since multiple proteins carry out most fundamental cellular processes through molecular interactions, it helps study protein-protein interactions and molecular biology of proteins. By employing electron capture dissociation (ECD) and collision induced dissociation (CID) fragmentation techniques, we have been able to understand better the disulfide pattern and the sequence of the proteins. A novel approach to interpreting data based on features associated with mass has been adopted for this work.
Acknowledgements

First and foremost, I am incredibly grateful to my supervisor Dr Dave Clarke for his invaluable advice, continuous support, and patience during my PhD study. Dr Clarke’s office was always there whenever I had a question or ran into a trouble spot about my research and writing. His immense knowledge and ample experience have encouraged me in my academic study and daily life. A massive thank you to Thomas Lambert for being the best possible person I could have gone through this PhD journey with. Thank you for the laughs and for listening to my many rants.

To the other members of the Clarke Group, Lulu, Sam, Kelly, Jenn, Maria, and Kiani, this experience would not have been the same without you all! Similarly, I imagine the mass spectrometry suite would have fallen to ruin a long time ago without Faye and Logan.

I need to thank Michalis Papadourakis, Harris Ioannidis, Loukas Manitsas, and Maria Routsi for their unwavering support, constant encouragement, and unshakeable belief in me. I could not have done this without you. None of the successes would have been possible either, so they deserve many thanks for all their help and endless optimism.

Last but not least, I would like to thank my parents and my brother for their support during the whole year.
Table of Contents

Developing novel workflows for improved bottom-up, top-down and middle-down mass spectrometry analysis of proteins................................................................. 1
Abstract.................................................................................................................................................................................. 3
Lay Summary ......................................................................................................................................................................... 4
Acknowledgements............................................................................................................................................................... 5
Table of Contents ................................................................................................................................................................. 6
List of Figures.......................................................................................................................................................................... 10
List of Tables .......................................................................................................................................................................... 19
Glossary .................................................................................................................................................................................... 21
1 Biomolecular Mass Spectrometry ........................................................................................................................................... 22
1.1 Mass Spectrometry ............................................................................................................................................................. 22
1.1.1 Instrumentation................................................................................................................................................................. 22
1.1.2 Nano-electrospray............................................................................................................................................................ 26
1.1.3 Charge states of proteins ions in ESI mass spectra and their relationship to protein structure in solution ........................................................................ 27
1.1.4 Liquid Chromatography .................................................................................................................................................. 29
1.1.5 Isotope Distributions ......................................................................................................................................................... 30
1.1.6 The isotopic fine structure of the protein ...................................................................................................................... 32
1.2 Mass Resolution and Resolving Power .................................................................................................................................... 36
1.2.1 Mass analysers ................................................................................................................................................................. 38
1.2.1.1 Quadrupoles ................................................................................................................................................................. 39
1.2.1.2 Time-of-flight Mass Spectrometry ............................................................................................................................ 40
1.2.1.3 Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS). 43
1.2.1.4 Orbitrap Mass Spectrometry ....................................................................................................................................... 47
1.2.1.5 Hybrid Instruments ......................................................................................................................................................... 48
1.2.1.6 Ion Mobility Mass Spectrometry ................................................................................................................................. 49
1.3 Protein Mass Spectrometry .................................................................................................................................................... 50
1.3.1 Bottom-Up Mass Spectrometry ........................................................................................................................................ 50
1.3.2 Top-Down Mass Spectrometry ....................................................................................................................................... 52
1.3.3 Middle-Down Mass Spectrometry ..................................................................................................................................... 55
1.4 Fragmentation ........................................................................................................................................................................... 56
1.4.1 Collision Induced Dissociation ......................................................................................................................................... 57
1.4.2 Electron Capture Dissociation and ETD ................................................................................................................................... 58
2 Complete mapping of disulfide connectivity in disulfide-rich proteins by high-resolution FT-ICR MS combined with pepsin and trypsin digestion .............................................. 60
3.3.4 Time-course experiments at varying temperatures .................................................. 126
3.3.5 LCMS to Confirm Conjugation and Assess Yield .................................................. 127
3.3.6 Sample Preparation for Top-Down Mass Spectrometry .......................................... 127
3.3.7 Top-Down Mass Spectrometry ............................................................................. 127
3.3.8 Mass Defect Analysis and De Novo Sequencing ..................................................... 128
3.4 Results and Discussion ......................................................................................... 129
3.4.1 Chemical Labelling of Insulin using reductive alkylation incorporating an N-terminal selective iodine-based mass defect tag ............................................................. 129
3.4.2 Time-course experiments of reductive alkylation incorporating N-terminal selective iodine-based mass defect tag at varying temperatures of Insulin ..................... 132
3.4.3 Evaluating Selectivity of the Chemical Labelling of Insulin using a reductive alkylation incorporating N-terminal selective iodine-based mass defect tag ............... 134
3.4.4 De Novo Sequencing of Insulin based on a reductive alkylation incorporating N-terminal selective iodine-based mass defect tag .................................................. 137
3.4.4.1 De Novo Sequencing of Insulin based on CID fragmentation and a reductive alkylation incorporating N-terminal selective iodine-based mass defect tag ............. 138
3.4.4.2 De Novo Sequencing of Insulin based on ECD fragmentation, and a reductive alkylation incorporating N-terminal selective iodine-based mass defect tag ............ 141
3.4.4.3 De Novo Sequencing of Insulin based on CID, ECD fragmentation, and a reductive alkylation incorporating N-terminal selective iodine-based mass defect tag ...... 144
3.4.5 Chemical Labelling of Myoglobin using reductive alkylation incorporating N-terminal selective iodine-based mass defect tag .................................................. 145
3.4.6 Time-course experiments of reductive alkylation incorporating N-terminal selective iodine-based mass defect tag at varying temperatures of Myoglobin .............. 147
3.4.7 Evaluating Selectivity of the Chemical Labelling of Myoglobin using a reductive alkylation incorporating N-terminal selective iodine-based mass defect tag ............. 148
3.4.8 Incorporation of a Bromo-substituted 2-pyridinecarboxaldehyde Mass Defect Tag via Imidazolidinone Formation on Ubiquitin ......................................................... 152
3.4.9 Comparison of top-down ECD fragmentation of Unmodified and BrMDT-modified Ubiquitin .............................................................................................. 155
3.4.10 Evaluating Selectivity of the Chemical Labelling of Ubiquitin incorporating a Bromo-substituted 2-pyridinecarboxaldehyde Mass Defect Tag ..................................... 160
3.4.11 De Novo Sequencing of Ubiquitin based Imidazolidinone Formation incorporating N-terminal selective bromine-based mass defect tag ....................................... 163
3.4.12 Incorporation of a Bromo-substituted 2-pyridinecarboxaldehyde Mass Defect Tag via Imidazolidinone Formation on Myoglobin ...................................................... 165
3.4.13 Evaluating Selectivity of the Chemical Labelling of Myoglobin incorporating a Bromo-substituted 2-pyridinecarboxaldehyde Mass Defect Tag ...................................... 166
3.4.14 De Novo Sequencing of Myoglobin based Imidazolidinone Formation incorporating N-terminal selective bromine-based mass defect tag ..................................... 168
3.4.15 Incorporation of a Bromo-substituted 2-pyridinecarboxaldehyde Mass Defect Tag via Reductive Alkylation on Rnase A ................................................................. 171

3.4.16 Evaluating Selectivity of the Chemical Labelling of RNase A incorporating a Bromo-substituted 2-pyridinecarboxaldehyde Mass Defect Tag .................................................. 176

3.4.17 De Novo Sequencing of RNase A based on ECD fragmentation, and alkylation described by MacDonald incorporating N-terminal selective bromine-based mass defect tag 178

3.5 Conclusions .................................................................................................................. 180

4. Conclusions .................................................................................................................. 182

Bibliography ..................................................................................................................... 184

Appendix .......................................................................................................................... 196
Figure 1.1: Schematic representation of a mass spectrometer. In a typical MS procedure, the injection of the analyte into the ion source occurs by the inlet. The ionised analyte can obtain either positive or negative charge and be manipulated by applied electric or magnetic fields. Then, the formed ions proceed to the mass analyser, which separates them according to their mass to charge ratio (m/z) by controlling their movement using electric or magnetic fields. The ions are detected by a mechanism capable of determining charged species, such as an electron multiplier. Ultimately, the signals are annotated and interpreted to get a mass spectrum.

Figure 1.2: Schematic layout of electrospray ionization process (in positive ionization mode). The analyte is injected through a hollow needle. A high voltage, typically 2–5 kV, is applied to the needle. Applying an electric field gradient between the sample capillary and the electrode for ion entry distorts the analyte solution's meniscus, forming a Taylor cone that ejects charged droplets that reduce in size until the ions are released into the analyser. Charged droplets are released from the tip of the Taylor cone and evaporate as they move towards the MS inlet to generate analyte ions.

Figure 1.3: Analyte transfer into the gaseous phase. Outline of the different electrospray (ESI) mechanisms. Ion evaporation model (IEM) for small molecules (A), charged residue model (CRM) for folded protein (B) and the chain ejection model (CEM) for unfolded protein (C). This figure was adapted from Konermann et al.

Figure 1.4: Example of a proteoform's theoretical mass spectra in a A) denatured conditions and B) native conditions. The CSD observed in the native folded conformation is more compact at higher m/z than in denatured protein conformation.

Figure 1.5: Simplified schematic representation of an HPLC system. The schematic diagram shows that the basic HPLC system consists of a HPLC pump, injector, column, detector, and integrator or acquisition and display system.

Figure 1.6: Impact of increasing molecular weight on the isotopic distributions. The number on the black is the relative abundance of the monoisotopic peak, and the number on red is the mass difference between the monoisotopic peak and the last one. As the polypeptide mass rises, the biomolecule's increased atomic complexity raises the merging of heavy isotopologues. This emerges as an increased number of isotopic peaks within the configuration spread over a higher mass scale. Consequently, the percentage abundance of the most abounding peak in the distribution becomes lower as the population's overall intensity is dispersed amongst each of the isotopic peaks in the distribution.

Figure 1.7: Mass defects of stable isotopes. The relative isotopic mass defect for common stable isotopes was calculated as the difference in mass between the monoisotopic mass (MIM) of the most abundant isotope and its nominal mass (NM).

Figure 1.8: Isotopic fine structure of a small peptide. The theoretical mathematical model of a small molecular weight polypeptide chain [C$_{69}$H$_{114}$N$_{20}$O$_{26}$S$_{1}$]$^{3+}$ (A) and a closer shot of both the M+1 (B) and the M+2 peak (C) to demonstrate the isotopic fine structure of the peptide.

Figure 1.9: Varying resolution of a protein. At varying resolutions, the information deduced from a spectrum varies. Lower resolution (5000, blue line) generates broader peaks from which an average mass can be determined. As the resolution improves (12,000, green line; 20,000, dashed line), the isotopic peaks become evident. Further increase in resolution (100,000, black line) enables the isotopic distribution to be fully resolved.

Figure 1.10: A schematic of quadrupole electrodes. The poles are arranged as two pairs of electrodes, as illustrated in the electric circuit. The potential differences established between the electrodes leads to a hyperbolic field. Ions are attracted towards the oppositely
charged electrodes, but the applied field results in potential switching of the electrodes and thus, ions oscillate in the x and y directions while travelling in the z-direction. .......................... 39

**Figure 1.11:** A schematic diagram of linear time-of-flight mass analyser demonstrating ions with varying m/z ratios. The ion with a smaller m/z ratio (orange) arrives at the detector before the ion with the higher m/z ratio (green) due to its higher velocity. .............. 41

**Figure 1.12:** Schematic graph of a reflectron ToF mass analyser. Ions with a higher primary kinetic energy (yellow) enter further into the reflectron, whilst ions with less kinetic energy (green) are reflected faster. This raises the mass resolution achieved......................... 42

**Figure 1.13:** The ICR detection cell’s orientation and direction of the electromagnetic field B is shown on the left side. In this magnetic field, ions will orbit the z-axis along the xy axis as displayed to the right; ions of counter-charges will move in a circular motion in reverse directions. This Lorentz motion is associated with the m/z and sets the basis of FT-ICR MS. .................................................................................................................................. 44

**Figure 1.14:** Detection of ions within an ICR detection cell. Ions are excited with a radiofrequency pulse utilising a potential difference between the two detection plates (E) to a greater radius. At this point, the ions will induce a current across the detection plates (D) to generate a complicated overlap of several overlapping frequencies. These radio frequencies are separated by the Fourier transform function and converted into the m/z of the ion.......... 46

**Figure 1.15:** Oscillation of ions within the orbitrap detector. Ions are introduced to the orbitrap parallel to the z-axis and held between the voltages applied to the central spindle and outer barrel. The axial frequency (ω) of the ion along the z-axis is dependent on the m/z of the ion and detected as a free induction decay (FID) and the axial frequency identified using a Fourier transform. From this the m/z ratio of the present ions can be calculated and shown within a mass spectrum. ........................................................................................................... 47

**Figure 1.16:** Simplified overview of the Bruker Daltonics Solarix-Qe FT-ICR mass spectrometer. Regions highlighted in orange indicate where MS/MS may take place. Ions of a particular m/z can be selected and isolated in the quadrupole prior to CID in the collision cell hexapole, or ECD in the ECD cathode........................................................................................................... 49

**Figure 1.17:** Bottom-up approach. In the Bottom-up method, enzymatic digestion is employed to cleave intact proteins into peptides. Peptides are determined through MS, and protein determination can occur by identifying peptides matched to a database. Peptides can be fragmented further to generate extensive sequence information and locate a post-translational modification to one distinct amino acid residue. ......................................................... 51

**Figure 1.18:** Top-down approach. In the Top-Down approach, intact proteins are introduced straight to the MS. The precise mass of the intact macromolecule is determined, followed by direct fragmentation. The generated precursor and fragments are then matched to potential sequences from a database. This provides the potential for the complete assignment of PTMs................................................................. 52

**Figure 1.19:** Top-Down MS/MS spectrum of a 13 kDa protein recorded on an FT ICR MS with isolated precursor ion highlighted (●). The FT ICR MS high-resolution allows for distinguishing the isotope distributions of the generated fragment ions. The inset zoom illustrates the overlapping distributions of the z213+, z304+ and c456+ ions and the separate ion distributions are highlighted by the modelled overlaid theoretical distributions...................... 54

**Figure 1.20:** Middle down approach. The middle-down workflow to PTMs assignment by MS. The intact protein is proteolysed into large peptides utilising enzymes or a chemical cleaving reagent. The large polypeptide chains may contain several PTMs sites yet be small enough to enable extensive fragmentation, thus allowing univocal localisation of PTMs. Middle-down MS is especially useful for providing extensive sequence coverage of a high molecular weight protein................................................................. 55
**Figure 1.21**: A schematic of the fragments that may arise upon top-down fragmentation study of a polypeptide backbone. a, b and c ions are N-terminus fragments and x, y and z ions are C-terminus fragments. The subscript indicates the number of amino acid residues in the generated fragment, counting from the N-terminus for a, b and c ions or the C-terminus for x, y and z ions. ................................................................. 57

**Figure 2.1**: Oxidative thiol modifications. Sulfur occupies a unique position in biology because of its ability to adopt a broad range of oxidation states (-2 to +6) and distinct chemical forms (chemotypes). This oxidative diversity, along with the inherent reactivity of the thiol and thiolate modifications, drive to the amino acid cysteine being susceptible to a diverse range of redox reactions which can lead to an array of protein covalent modifications. ........................................................................................................................................................................... 61

**Figure 2.2**: Cellular bond disulfide formation. Schematic representation of electron transfer oxidation and reduction of disulfide bonds. ........................................................................................................................................................................... 62

**Figure 2.3**: Bottom-up proteomics workflow for disulfide connectivity mapping. In the Bottom-up method, enzymatic digestion is employed to cleave intact proteins into peptides. Peptides are then chemically reduced and run by MS, and peptides are identified by running them against a database. Afterwards, oxidised peptides are run by LC-MS/MS to generate extensive sequence information and locate the disulfide bonds. ........................................................................................................................................................................... 64

**Figure 2.4**: The key challenges related to routine disulfide connectivity mapping are enzymatic digestion, closely spaced cysteine residues, and S-S scrambling................. 66

**Figure 2.5**: Fragmentation patterns for persulfide and dehydroalanine fragment ions formed by C-S bond cleavage reactions, and for cysteine and cysteine thioaldehyde fragment ions formed by S-S bond cleavage reactions ........................................................................................................................................................................... 68

**Figure 2.6**: Selective homolytic cleavage of the disulfide linkage with 266 nm wavelength. ........................................................................................................................................................................... 70

**Figure 2.7**: Developed proteomics workflow for disulfide connectivity mapping. In this method, enzymatic digestion is employed to cleave intact proteins into peptides. Peptides are chemically reduced and run by FT-ICR-MS, and peptides are identified by running them against a database. Afterwards, oxidised peptides are run by FT-ICR-MS and fragmented by CID and ECD to generate extensive sequence information and locate the disulfide bonds. 71

**Figure 2.8**: Bovine Serum Albumin primary (a) and secondary (b) structures, the latter containing its native SS cross-links. Based on Majorek et al.44 (Protein Data Bank code 3V03) and Huang et al. (Uniprot Code:PO2769). ........................................................................................................................................................................... 73

**Figure 2.9**: Total number of hits of reduced BSA species (right sequence) using an error tolerance of 2, 5, 10, 20 and 50 ppm error. From the graph, we can conclude that there is a strong correlation between the set error and the total number of hits. .......... 77

**Figure 2.10**: Total number of hits of reduced BSA species (shuffled sequence) using an error tolerance of 2, 5, 10, 20, and 50 ppm error. We can see a strong correlation between the set error and the total number of hits from the graph. ........................................... 77

**Figure 2.11**: Total number of hits of oxidised BSA species (correct sequence) using an error tolerance of 2, 5, and 10 ppm error. The correct and false positives were illustrated with blue and orange, respectively. The graph above shows that as the error of tolerance is set higher, false hits rise significantly. We thus chose 2 ppm error as the optimum since the ratio of correct positives to total hits was the highest. ......... 78

**Figure 2.12**: Total number of hits of reduced BSA species (correct sequence) using an error tolerance of 2, 5, 10, 20, and 50 ppm error. From the graph, we can note a strong correlation between the set error and the total number of hits. Specifically, setting the error of tolerance to higher values leads to a higher number of total hits. ........................................... 79

**Figure 2.13**: Total number of hits of reduced BSA species (shuffled sequence) using an error tolerance of 2, 5, 10, 20, and 50 ppm error. From the graph, we can see a strong
correlation between the set error and the total number of hits. Specifically, setting the error of tolerance to higher values leads to a higher number of total hits.  

**Figure 2.14:** Total number of hits of reduced BSA species (shuffled sequence) using an error tolerance of 2 and 5 ppm error. The correct and false positives were illustrated with blue and orange, respectively. The graph above shows that as the error of tolerance is set higher, false hits rise significantly. We thus chose 2 ppm error as the optimum since the ratio of correct positives to total hits was the highest.  

**Figure 2.15:** Illustration of the various potential disulfide bridged peptides that could be generated.  

**Figure 2.16:** Comparison of the ion [M+3H]$$^3+$$ with mass of 1745.7713 with the simulated isotope distribution of AKTC$_{53}$VADESHAGC$_{62}$EKSL with one disulfide bond. The simulated isotope distribution is plotted over the obtained spectrum as a scatterplot. To add an even higher level of confidence, the third peak of the isotopic distribution was zoomed in and compared to the mathematical simulation. The zooming shows that FT ICR MS resolution is so high that the isotopic composition is similar to the theoretical one. This comparison confirms that the oxidised peptide is AKTC$_{53}$VADESHAGC$_{62}$EKSL with one disulfide bond.  

**Figure 2.17:** CID spectra for the [M+3H]$$^3+$$ charge state of AKTC$_{53}$VADESHAGC$_{62}$EKSL with collisional voltage set to 12 V.  

**Figure 2.18:** The fragment map displays fragment b and y ions resulting from CID fragmentation of AKTC$_{53}$VADESHAGC$_{62}$EKSL. Several peaks have been generated, allowing 100 % sequence coverage.  

**Figure 2.19:** ECD spectra for the [M+3H]$$^3+$$ charge state of AKTC$_{53}$VADESHAGC$_{62}$EKSL with pulse length set to 100 ms. The fragments with black colour are the ones with the disulfide bond intact, while the red ones result from the reduction of disulfide bonds by ECD.  

**Figure 2.20:** Analysis of c6 fragment ion of oxidised AKTC$_{53}$VADESHAGC$_{62}$EKSL. The c6 fragment ion was compared with a mathematical simulation assuming a c6 fragment from the given sequence with a loss of a proton. To add an even higher level of confidence, the third peak of the isotopic distribution was zoomed in and compared to the mathematical simulation. The zooming of the third peak shows that the data matches the theoretical isotopic distribution even at this high level of resolution. All these findings confirm that the c6 ion in the oxidised AKTC$_{53}$VADESHAGC$_{62}$EKSL ECD spectra has lost a hydrogen suggesting a cleavage of the SS–bond between Cys53 and Cys62.  

**Figure 2.21:** Analysis of c7 fragment ion of oxidised AKTC$_{53}$VADESHAGC$_{62}$EKSL. The c7 fragment ion was compared with a mathematical simulation assuming a c6 fragment from the given sequence with a proton loss. This comparison hints that it matches. The zooming of the third peak shows that the data matches the theoretical isotopic distribution even at this high level of resolution. All these findings confirm that the c7 ion in the oxidised AKTC$_{53}$VADESHAGC$_{62}$EKSL ECD spectra has lost a hydrogen suggesting a cleavage of the SS–bond between Cys53 and Cys62.  

**Figure 2.22:** Analysis of c14 fragment ion of oxidised AKTC$_{53}$VADESHAGC$_{62}$EKSL. The c14 fragment ion was compared with a mathematical simulation assuming a c14 fragment from the given sequence with the disulfide bond intact, and this comparison implies that it matches. The zooming of the third peak shows that the data matches the theoretical isotopic distribution even at this high level of resolution.  

**Figure 2.23:** Analysis of c14 fragment ion of oxidised AKTC$_{53}$VADESHAGC$_{62}$EKSL. The c12 fragment ion was compared with a mathematical simulation, and this comparison suggests that it matches. The c12 ion in the oxidised AKTC$_{53}$VADESHAGC$_{62}$EKSL ECD
spectra has gained a sulphur suggesting a cleavage of the SS–bond between Cys53 and Cys52.

**Figure 2.24**: The fragment map displays fragment ions resulting from ECD fragmentation of AKTC53VADESHAGC52EKSL. Several peaks have been generated, allowing 82% sequence coverage.

**Figure 2.25**: Comparison of the ion \([M+5H]^6+\) with a mass of 3584.7837 with the mathematical simulation assuming LKEC460VLHEKPVSEKVTKC474C475TESLVRNPC486F with two disulfide bonds. This comparison suggests that the oxidised peptide is LKEC460VLHEKPVSEKVTKC474C475TESLVRNPC486F with two disulfide bonds.

**Figure 2.26**: ECD spectra for the \([M+6H]^6+\) charge state of NRLC460VLHEKPVSEKVTKC474C475TESLVRNPC486F with pulse length set to 100 ms.

**Figure 2.27**: The fragment map displays fragment ions resulting from ECD fragmentation of NRLC460VLHEKPVSEKVTKC474C475TESLVRNPC486F. Several peaks have been generated, allowing 77% sequence coverage.

**Figure 2.28**: Comparison of the ion \([M+4H]^4+\) with a mass of 2588.1240 with the mathematical simulation assuming VHKEC244C245HGDLLEC251ADDR–C199ASIQK with two disulfide bonds. This comparison suggests that the oxidised peptide is VHKEC244C245HGDLLEC251ADDR–C199ASIQK with two disulfide bonds.

**Figure 2.29**: CID spectra for the \([M+4H]^4+\) charge state of VHKEC244C245HGDLLEC251ADDR–C199ASIQK with the collisional voltage set to 12V.

**Figure 2.30**: The fragment map displays fragment b and y ions resulting from CID fragmentation of VHKEC244C245HGDLLEC251ADDR–C199ASIQK. Several peaks have been produced, allowing 56% sequence coverage.

**Figure 2.31**: ECD spectra for the \([M+4H]^4+\) charge state of VHKEC244C245HGDLLEC251ADDR–C199ASIQK with pulse length set to 120 ms.

**Figure 2.32**: The fragment map displays fragment ions resulting from ECD fragmentation of VHKEC244C245HGDLLEC251ADDR–C199ASIQK assuming the intramolecular disulfide bond intact.

**Figure 2.33**: The fragment map displays fragment ions resulting from ECD fragmentation of VHKEC244C245HGDLLEC251ADDR–C199ASIQK assuming the intramolecular disulfide bond reduced.

**Figure 2.34**: The fragment map displays fragment ions resulting from ECD fragmentation of VHKEC244C245HGDLLEC251ADDR–C199ASIQK. Several peaks have been produced, enabling 87% sequence coverage.

**Figure 2.35**: Comparison of the ion \([M+6H]^6+\) with a molecular weight of 3815.8090 with the mathematical simulation assuming LKEC277C278DKPLLEK–SHC288IAVEK–YIC264DNQDTISISK with two disulfide bonds. This comparison suggests that the oxidised peptide is LKEC277C278DKPLLEK–SHC288IAVEK–YIC264DNQDTISISK with two intermolecular disulfide bonds.

**Figure 2.36**: ECD spectra for the \([M+6H]^6+\) charge state of LKEC277C278DKPLLEK–SHC288IAVEK–YIC264DNQDTISISK with pulse length set to 65 ms.

**Figure 2.37**: ECD spectra for the \([M+6H]^6+\) charge state of LKEC277C278DKPLLEK–SHC288IAVEK–YIC264DNQDTISISK with pulse length set to 65 ms.

**Figure 2.38**: ECD spectra for the \([M+6H]^6+\) charge state of LKEC277C278DKPLLEK–SHC288IAVEK–YIC264DNQDTISISK with pulse length set to 65 ms.

**Figure 2.39**: The fragment map displays fragment ions resulting from ECD fragmentation of LKEC277C278DKPLLEK–SHC288IAVEK–YIC264DNQDTISISK, assuming both intermolecular disulfide bonds reduced.
Figure 2.40: The fragment map displays fragment ions resulting from ECD fragmentation of LKEC\textsubscript{277}C\textsubscript{278}DKP\textsubscript{LLEK-SHC\textsubscript{288}IAEVEK-YIC\textsubscript{294}DNQDTISSK}. Several peaks have been generated, allowing 86% sequence coverage. ........................101

Figure 3.1: Diagram outlining structural aspects of a human antibody and how it can be broken down for middle down MS. The first step in the workflow involves separating by proteolytic cleavage of the IgG CDR-containing Fab from the constant Fc utilizing the IdeS/GingisKHAN enzyme, which cleaves just underneath the disulfides bonds (yellow) among the heavy chains. Afterwards, the generated peptides are chemically reduced. ..........................104

Figure 3.2: A) Mass spectra of bovine Ubiquitin (Ubiq). The charge state distribution of Ubiq. B) The MS spectra of the [M+8H]\textsuperscript{8+} charge state of Ubiq. The Ubiq was compared with the mathematical simulation assuming a molecular formula of C\textsubscript{378}H\textsubscript{629}N\textsubscript{105}O\textsubscript{118}S\textsubscript{1} (8559.616712 Da). ..........................105

Figure 3.3: ECD fragmentation spectrum of standard Ubiq. The bottom spectrum shows a magnified region of the fragmentation spectrum with all assigned signals labelled.............106

Figure 3.4: The fragment map displays fragment ions resulting from ECD fragmentation of Ubiquitin. Red hooks indicate identified fragments. Specifically, ECD fragmentation led to 91% sequence coverage, of which 34% of the identified ions were assigned within a 3 ppm error. ..........................................................106

Figure 3.5: The potential protein sequences of Ubiq determined de novo employing conventional ECD fragmentation. Ambiguous sequences are highlighted in Red and unknown amino acids with green. ........................................................................107

Figure 3.6: MD against nominal mass along the length of three polymers along with the hypothetical poly(averagine). ..........................................................................110

Figure 3.7: MD against nominal mass along the length of three polymers along with the hypothetical poly(averagine). ..........................................................................111

Figure 3.8: Mass spectra of protein treated without (top) and with (bottom) mass defect tag. The c and z ions are closely nested on the top, leading to potentially overlapping species that would be hard to distinguish. The signals have been spread out across the m/z range on the bottom, allowing more straightforward and more confident detection. ........112

Figure 3.9: Chemical Structure of dibromo-(2'-iodo) acetanilide. .............................113

Figure 3.10: Mass defect analysis illustrated using the peptide GYSKEASAL. The MS software provides the list of the monoisotopic masses (MIM). Afterwards, this mass is corrected by the averagine factor to standardize it for every protein (MIM\textsubscript{avg}). The nominal mass (NM\textsubscript{avg}) is calculated by rounding the previously estimated number, and finally, the MD is calculated by subtracting from the MIM\textsubscript{avg} the NM\textsubscript{avg}. The MD plot was constructed by plotting MIM against MD. As illustrated, c and z ions overlap with each other, thus making it impossible to distinguish between the two classes without any prior knowledge..............116

Figure 3.11: Mass defect analysis illustrated using the peptide GYSKEASAL labelled on the N-termini with an MD tag (C\textsubscript{4}H\textsubscript{9}I, monoisotopic mass: 215.943598). As before, the MS software provides the list of the monoisotopic masses (MIM). Afterwards, this mass is corrected by the average factor to standardize it for every protein (MIM\textsubscript{avg}). The nominal mass (NM\textsubscript{avg}) is calculated by rounding the previously estimated number, and finally, the MD is calculated by subtracting from the MIM\textsubscript{avg} the NM\textsubscript{avg}. The MD plot was constructed by plotting MIM against MD. A closer look indicates two trends corresponding to c (red) and z ions (blue ions). Therefore, labelling selectively either the N- or C-terminus allows for easy resolution of tagged from untagged fragments and, in the case of a terminus-selective label, among N- and C-terminal fragments. ............................226,227 ..........................117

Figure 3.12: Standard strategies for modifying N-terminal amino acid residues. (a) Thioesters react with N-terminal Cys by native chemical ligation. (b) Thiazolidines are formed by the reaction between aldehydes with N-terminal Cys. (c) Sodium periodate is
employed to form aldehydes from N-terminal serine and threonine residues. (d) Pictet-Spengler reactions are undergone between N-terminal tryptophans and aldehydes. (e) The N-terminal amino group was converted to a carbonyl group by transamination reagents. (f) N-terminal prolines react with aminophenols by an oxidation reaction (g) 2-pyridinecarboxaldehyde (2PCA) was employed to label the N-termini selectively, the nitrogen of the adjacent amino acid cyclises on the imine to form an imidazolidinone product at the N termini.                                                                                                                119

Figure 3.13: The Top-down workflow in combination with a mass defect labelling. Protein is selectively labelled on the N-terminus with a covalent MD label before top-down fragmentation. The large MD allows the classification of fragment ions into N-terminal (c-ions) and C-terminal (z-ions) fragment series without prior sequence information.                                                                                                    121

Figure 3.14: Structural features of a) insulin, b) ubiquitin, myoglobin and Ribonuclease A. a) Primary structure of the structure of porcine insulin (5.7 kDa). The protein contains 6 cysteine residues, which form 3 disulfide bonds. The disulfide connective is highlighted. b) Primary structure of the structure of Ubiquitin (8.5 kDa). c) Primary structure of the structure of Myoglobin (16.9 kDa). d) Primary structure of the structure of bovine Ribonuclease A (13.7 kDa). The protein contains 8 cysteine residues, which form 4 disulfide bonds. The disulfide connective is highlighted.                                                                                     123

Figure 3.15: Reaction scheme for the N-terminal selective incorporation of 4-iodobenzaldehyde into Ins and Myo proteins.                                                                                                    125

Figure 3.16: Reaction scheme for the N-terminal selective incorporation of 5-Bromo-substituted 2-pyridinecarboxaldehyde into proteins.                                                                                     126

Figure 3.17: Reaction scheme for the N-terminal selective incorporation of a) benzaldehyde and b) 4-iodobenzaldehyde into proteins.                                                                             129

Figure 3.18: A) Mass spectra of insulin (Ins). Top: Controls showing the charge state distributions of the unmodified protein. Bottom: Spectra of the crude reaction mixture with a combination of unmodified protein and species with higher mass. B) Observed isotopic distributions of Ins overlaid with simulated isotopic distributions for the a) unmodified (C_{254}H_{477}N_{56}O_{75}S_{6}, red), b) singly modified (C_{261}H_{482}N_{56}O_{75}S_{6}, blue) and di-modified (C_{268}H_{487}N_{56}O_{76}S_{6}, yellow) protein as represented by coloured rings. 131

Figure 3.19: Relative yields from Ins within a time course experiment at varying temperatures. As illustrated, the unmodified proteoform disappeared throughout the time as it was used up in the chemical reactions to generate the modified species. Furthermore, the relative yield of the di-modified conjugate of Ins had a steady increase over the whole time range without reaching a plateau, and the same can be observed for the tri- and tetra-modified conjugates. The most striking observation from the data comparison was that rise was already visible at the 1 hr time point. The highest yield of the di-modified species was 31 %, and it was achieved after 23 hours of reaction time at 47 °C.                                                                                               133

Figure 3.20: ECD fragmentation spectrum of Ins modified with iodobenzaldehyde. The charge state distribution of Ins, 5+ charge state isolated (middle) and with ECD (bottom). The bottom spectrum shows a magnified region of the fragmentation spectrum with all assigned signals labelled.                                                                                                           134

Figure 3.21: Sequence of the Ins B-chain with an MD tag on the N-terminus (top) or Lys-29 (bottom) matched against the ECD fragmentation data for Ins modified under initial conditions. Red hooks indicate identified fragments.                                                                                                          135

Figure 3.22: Plot showing the intensities of all fragments assigned from ECD fragmentation data for Ins modified under initial conditions. The displayed ratios highlight the relatively low intensity of the fragments suggesting tagging of Lys-29 compared to their counterparts, suggesting an N-terminal tag.                                                                                136
Figure 3.23: A) CID fragmentation spectrum of Ins B chain modified with iodobenzaldehyde. B) Mass defect plot derived employing averagine scaling correcting factor for the assigned fragment ions from Ins CID fragmentation dataset.

Figure 3.24: On the top, a mass defect plot derived employing averagine scaling correcting factor for the assigned fragment ions from Ins CID fragmentation dataset and on the bottom a zoomed region (600-1400 Da) of this plot.

Figure 3.25: The protein sequence of Ins B chain determined de novo combining an N-termini MD tag and CID cleavage. Although the performance was not ideal, we nevertheless managed to determine 90% of the primary sequence was assigned. Ambiguous sequences are highlighted in Red and unknown amino acid acids with green.

Figure 3.26: A) ECD fragmentation spectrum of Ins B chain modified with iodobenzaldehyde. B) Mass defect plot derived employing averagine scaling correcting factor for the assigned fragment ions from Ins ECD fragmentation dataset.

Figure 3.27: On the bottom, a mass defect plot derived employing averagine scaling correcting factor for the assigned fragment ions from Ins ECD fragmentation dataset and on the right, a zoomed region (300-800 Da) of this plot.

Figure 3.28: The protein sequence of Ins B chain determined de novo combining an N-termini MD tag and ECD fragmentation. Ambiguous sequences are highlighted in Red.

Figure 3.29: The protein sequence of Ins B chain determined de novo combining an N-termini MD tag and CID & ECD fragmentation. Ambiguous sequences are highlighted in Red.

Figure 3.30: A) Mass spectra of Myo. Top: Controls showing the charge state distributions of the unmodified protein. Bottom: Spectra of the crude reaction mixture with a combination of unmodified protein and species with higher mass. B) Observed isotopic distributions of Myo overlaid with simulated isotopic distributions for the unmodified (red) and singly-modified (IMDT) (blue) protein as represented by coloured rings.

Figure 3.31: Relative yields from Myo within a time course experiment at varying temperatures.

Figure 3.32: ECD fragmentation spectrum of Myo modified with iodobenzaldehyde. The charge state distribution of Myo, 20+ charge state isolated (middle) and with ECD (bottom). The bottom spectrum shows a magnified region of the fragmentation spectrum with all assigned signals labelled.

Figure 3.33: Sequence of the Myo with an MD tag on the N-terminus (top) or the C-terminus (bottom) matched against the ECD fragmentation data for Myo modified under initial conditions. Red hooks indicate identified fragments.

Figure 3.34: Plot showing the intensities of all fragments assigned from ECD fragmentation data for Myo modified under initial conditions. The displayed ratios highlight the relatively low intensity of the fragments suggesting tagging of an internal Lysine compared to their counterparts which suggest an N-terminal tag.

Figure 3.35: Reaction scheme for the N-terminal selective incorporation of a) of 2PCA-biotin reagent and b) 5-bromo-substituted 2-pyridinecarboxaldehyde into proteins.

Figure 3.36: a) The mass spectra of Ubiquitin (Ubiq) treated with 5-Bromo-substituted 2-pyridinecarboxaldehyde and denaturing agent. Observed isotopic distributions of Ubiq overlaid with simulated isotopic distribution for the the b) unmodified (red) and c) singly-modified (BrMDT) (blue) protein as represented by coloured rings.
**Figure 3.37:** ECD fragmentation spectra for Ubiq. Top-down ECD fragmentation of standard (top) and BrMDT-modified (bottom) protein. From the graph, it can be observed that higher fragmentation coverage is achieved for standard Ubiq.

**Figure 3.38:** The fragment map displays fragment ions resulting from ECD fragmentation of standard (top) and modified (bottom) Ubiquitin. Top-down FT ICR MS suggested that the N-termini modification leads to lower sequence amino acid coverage of the protein. Specifically, ECD fragmentation of the standard and modified led to 35 and 87% sequence coverage respectively within a 5 ppm error.

**Figure 3.39:** The number of c and z ions produced by standard and Modified Ubiq under the same conditions.

**Figure 3.40:** The sequence coverage obtained by employing 3 V, 5 V, and 10 V ECD bias of 1.5, 2.5, and 1.8 and gas control flow of 30 % and 10 %. This study indicates that the optimum collisional voltage, ECD bias, and gas control flow are 5 V, 1.5 V, and 10 %, respectively. Thus, the highest number of c*(58) and z ions (191) are generated at the optimum conditions.

**Figure 3.41:** ECD fragmentation spectrum of Ubiq modified with 5-Bromo-substituted 2-pyridinecarboxaldehyde. The charge state distribution of Ubiq, 11+ charge state isolated (middle) and with ECD (bottom). The bottom spectrum shows a magnified region of the fragmentation spectrum with all assigned signals labelled.

**Figure 3.42:** Sequence of the Ubiq with an MD tag on the N-terminus (top) or C-terminus (bottom) matched against the ECD fragmentation data for Ubiq modified. Red hooks indicate identified fragments.

**Figure 3.43:** A) ECD fragmentation spectrum of Ubiq modified 5-Bromo-substituted 2-pyridinecarboxaldehyde. B) Mass defect plot derived employing averagine scaling correcting factor for the assigned fragment ions from Ubiq ECD fragmentation dataset.

**Figure 3.44:** The protein sequence of Ubiq determined de novo combining an N-termini MD tag and ECD fragmentation. Ambiguous sequences are highlighted in Red and unknown amino acid acids with green. The simplified de novo sequence process described here led to 91 % of the primary sequence being determined.

**Figure 3.45:** A) Mass spectra of Myo. Top: Controls showing the charge state distributions of the unmodified protein. Bottom: Spectra of the crude reaction mixture with a combination of unmodified protein and species with higher mass. Observed isotopic distributions of Myo overlaid with simulated isotopic distributions for the B) unmodified (red) and C) singly-(blue) protein as represented by coloured rings.

**Figure 3.46:** ECD fragmentation spectrum of Myo modified with BrMDT. The bottom spectrum shows a magnified region of the fragmentation spectrum with all assigned signals labelled.

**Figure 3.47:** Sequence of the Myo with an MD tag on the N-terminus (top) or C-terminus (bottom) matched against the ECD fragmentation data for Myo modified under initial conditions. Identified fragments are indicated by red hooks.

**Figure 3.48:** A) ECD fragmentation spectrum of Myo modified with 5-Bromo-substituted 2-pyridinecarboxaldehyde. B) Mass defect plot derived employing averagine scaling correcting factor for the assigned fragment ions from Myo ECD fragmentation dataset.

**Figure 3.49:** The protein sequence of Myo determined de novo combining an N-termini MD tag and ECD fragmentation. Ambiguous sequences are highlighted in Red and unknown amino acid acids with an “X”. The proposed simplified de novo sequence process described here led to 70 % of the primary sequence.
**Figure 3.50:** A) Mass spectra of oxidised RNase A. Top: Controls showing the charge state distributions of the unmodified protein. Bottom: Spectra of the crude reaction mixture with a combination of unmodified protein and species with higher mass. B) Observed isotopic distributions of oxidised Rnase A overlaid with simulated isotopic distributions for the a) unmodified (red) and b) singly-(blue) protein as represented by coloured rings.

**Figure 3.51:** The two protocols studied to modify the N-terminus and reduce and alkylate the cysteines selectively. In the A protocol, on the first step, cysteines are reduced, and then alkylated and on the final step the N-terminus is modified. On the B protocol, the N-terminus is first modified, and then cysteines are reduced and modified...

**Figure 3.52:** Top: Mass spectra of Modified RNase when protocol A is followed. In the A protocol, on the first step, cysteines are reduced, alkylated, and the N-terminus is modified on the final step. Bottom: Mass spectra of Modified RNase when protocol B is followed. On the B protocol, the N-terminus is first modified, and then cysteines are reduced and modified.

**Figure 3.53:** A) Mass spectra of Modified RNase A. B) Observed isotopic distributions of NEM and N-terminally modified Rnase A overlaid with simulated isotopic distributions for the singly-(blue) protein as represented by coloured rings.

**Figure 3.54:** ECD fragmentation spectrum of RNase A modified with 5-Bromo-substituted 2-pyridinecarboxaldehyde. The bottom spectrum shows a magnified region of the fragmentation spectrum with all assigned signals labelled.

**Figure 3.55:** Sequence of the RNase A with an MD tag on the N-terminus (top) or C-terminus (bottom) matched against the ECD fragmentation data for Myo modified under initial conditions. Identified fragments are indicated by red hooks.

**Figure 3.56:** A) ECD fragmentation spectrum of RNase A modified with 5-Bromo-substituted 2-pyridinecarboxaldehyde. B) Mass defect plot derived employing averagine scaling correcting factor for the assigned fragment ions from RNase A ECD fragmentation dataset.

**Figure 3.57:** The protein sequence of RNase A determined de novo combining an N-termini MD tag and ECD fragmentation. Ambiguous sequences are highlighted in Red and unknown amino acid acids with green. The proposed simplified de novo sequence process described here led to 79% of the primary sequence.

---

**List of Tables**

<table>
<thead>
<tr>
<th>Table 1.1</th>
<th>Natural isotopic distribution of common elements. This table lists the mass and per cent natural isotopic abundance for the stable nuclides observed in proteins.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1.2</td>
<td>Nominal- and monoisotopic mass and mass defect of elements found in proteins and iodine.</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Nominal- and monoisotopic mass and mass defect of elements found in proteins and iodine and bromine.</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Halogen-based MD tags employed in this research to increase fragment ion assignment efficiency in top-down mass spectra and therefore maximise the protein sequence coverage obtained. Such a tag aids classification between N- and C-terminal fragments from top-down MS data, thus simplifying fragment assignment.</td>
</tr>
</tbody>
</table>
Table 3.3: Identified fragments sorted according to ion type and whether they are consistent with a tag on the N-terminus or Lys-29. The ratios indicate selectivity based on the summed intensity of fragments suggesting N-terminal labelling versus fragments suggesting labelling of Lys-29. 

Table 3.4: Comparison of specific identified indicating a tag on the N-terminus or Lys-29. The ratios indicate selectivity based on the summed intensity of fragments suggesting N-terminal labelling versus fragments suggesting labelling of Lys-29.

Table 3.5: Identified fragments sorted according to ion type and whether they are consistent with a tag on the N-terminus or unselective. The ratios indicate selectivity based on the summed intensity of fragments suggesting N-terminal labelling versus fragments suggesting unselective labelling.

Table 3.6: Different tested ECD conditions. The impact of ECD bias (V), collisional voltage (V), and the gas flow control (%). The available evidence seems to suggest that the optimum conditions are the ones described by experiment 6.

Table 3.7: Identified fragments sorted according to ion type and consistent with a tag on the N-terminus (selective) or an internal lysine (unselective). The ratios show selectivity based on the summed intensity of fragments suggesting N-terminal labelling versus fragments suggesting unselective labelling.

Table 3.8: Identified fragments sorted according to ion type and consistent with a tag on the N-terminus or an internal lysine. The ratios indicate selectivity based on the summed intensity of fragments suggesting N-terminal labelling versus fragments suggesting unselective labelling.
# Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACN</td>
<td>Acetonitrile</td>
</tr>
<tr>
<td>BrMDT</td>
<td>bromine-based mass defect tag</td>
</tr>
<tr>
<td>BSA</td>
<td>Bovine Serum Albumin</td>
</tr>
<tr>
<td>CID</td>
<td>collision induced dissociation</td>
</tr>
<tr>
<td>CNBr</td>
<td>Cyanogen Bromine</td>
</tr>
<tr>
<td>CSD</td>
<td>accumulation of multiple charges</td>
</tr>
<tr>
<td>Cys</td>
<td>Cysteine</td>
</tr>
<tr>
<td>Da</td>
<td>Dalton</td>
</tr>
<tr>
<td>ECD</td>
<td>electron capture dissociation</td>
</tr>
<tr>
<td>ESI</td>
<td>electrospray ionisation</td>
</tr>
<tr>
<td>ETD</td>
<td>electron transfer dissociation</td>
</tr>
<tr>
<td>EThcD</td>
<td>electron-transfer/higher-energy collision dissociation</td>
</tr>
<tr>
<td>ExD</td>
<td>electron transfer or capture dissociation</td>
</tr>
<tr>
<td>FA</td>
<td>Formic acid</td>
</tr>
<tr>
<td>FT-ICR</td>
<td>Fourier transform-ion cyclotron resonance</td>
</tr>
<tr>
<td>HCD</td>
<td>higher-energy collisional dissociation</td>
</tr>
<tr>
<td>Hsp33</td>
<td>heat shock protein 33</td>
</tr>
<tr>
<td>IMDT</td>
<td>iodine-based mass defect tag</td>
</tr>
<tr>
<td>Ins</td>
<td>Insulin</td>
</tr>
<tr>
<td>KE</td>
<td>kinetic energy</td>
</tr>
<tr>
<td>KM</td>
<td>Kendrick mass</td>
</tr>
<tr>
<td>KMD</td>
<td>Kendrick mass defect</td>
</tr>
<tr>
<td>LC</td>
<td>liquid chromatography</td>
</tr>
<tr>
<td>mAb</td>
<td>monoclonal antibody</td>
</tr>
<tr>
<td>MALDI</td>
<td>Matrix-assisted laser desorption/ionisation</td>
</tr>
<tr>
<td>MD</td>
<td>mass defect</td>
</tr>
<tr>
<td>MS</td>
<td>mass spectrometry</td>
</tr>
<tr>
<td>nESI</td>
<td>nanoelectrospray ionisation</td>
</tr>
<tr>
<td>NMR</td>
<td>nuclear magnetic resonance</td>
</tr>
<tr>
<td>PDB</td>
<td>protein data bank</td>
</tr>
<tr>
<td>PTM</td>
<td>post-translational modification</td>
</tr>
<tr>
<td>ROS</td>
<td>Reactive oxygen species</td>
</tr>
<tr>
<td>SNAP</td>
<td>sophisticated numerical annotation procedure</td>
</tr>
<tr>
<td>TCEP</td>
<td>Tris (2-carboxyethyl) phosphine</td>
</tr>
<tr>
<td>THRASH</td>
<td>thorough high-resolution analysis of spectra by Horn</td>
</tr>
<tr>
<td>TOF</td>
<td>time of flight</td>
</tr>
<tr>
<td>UVPD</td>
<td>ultraviolet photodissociation</td>
</tr>
</tbody>
</table>
1 Biomolecular Mass Spectrometry

1.1 Mass Spectrometry

1.1.1 Instrumentation

The basis of this thesis is the development of novel mass spectrometry (MS) methods to study protein structure. In the past two decades, MS has become an essential analytical tool, and it can provide both qualitative and quantitative information of analytes.\(^1\) Researchers from all disciplines of life sciences and beyond are working with this technique to identify and characterise biomolecules.\(^2\) Even though this analytical technique has only been widely applied to life sciences recently, the tremendous technological advancements, including instrument accuracy and sensitivity and more freely accessible protein databases, have established this technique as an essential tool for studying many biological systems. This is perhaps best exemplified in the field of proteomics, where MS has become the core analytical technology. This discipline, which involves the identification of polypeptide chains in complex mixtures including cell lysates, tissue extracts or clinical fluids, relies on sensitivity, speed and resolution afforded by modern MS.\(^3,4\)

MS nominally measures the \(m/z\) ratio of gas-phase ions allowing the determination of molecular masses of analytes.\(^5\) Fundamentally, a basic mass spectrometer comprises an ion source, a mass analyser and a detector (Figure 1.1). The injection of the analyte into the ion source occurs by the inlet under a vacuum. The ionised analyte can obtain either positive or negative charge and be manipulated by applied electric or magnetic fields. Then, the formed ions proceed to the mass analyser, which separates them by their mass to charge ratio (\(m/z\)) by influencing their movement by applying electric or magnetic fields.\(^6,7\) An electron multiplier is employed to sense the presence of ion signals emerging from the mass analyzer of an MS. The task of the electron multiplier is to determine every ion of the specified mass passed by the mass filter. The primary physical process that enables an electron multiplier to function is based on the secondary electron emission theory. When a charged particle (ion or electron) hits a surface, it induces secondary electrons to be emitted from atoms in the surface layer.\(^8\) Ultimately, the signals are annotated and interpreted to get a mass spectrum. By examining the mass spectrum, the analyte can be defined as each molecule has a unique mass spectrum.\(^6,7\)
However, given the non-volatile nature of proteins, it is only since the development in the 1980s of the soft ionisation techniques of matrix-assisted laser desorption ionisation (MALDI)\(^9\) and electrospray ionisation\(^{10}\) (ESI) made MS capable of analysing intact protein ions.

![Diagram of a mass spectrometer](image)

**Figure 1.1: Schematic representation of a mass spectrometer.** In a typical MS procedure, the injection of the analyte into the ion source occurs by the inlet. The ionised analyte can obtain either positive or negative charge and be manipulated by applied electric or magnetic fields. Then, the formed ions proceed to the mass analyser, which separates them according to their mass to charge ratio (m/z) by controlling their movement using electric or magnetic fields. The ions are detected by a mechanism capable of determining charged species, such as an electron multiplier. Ultimately, the signals are annotated and interpreted to get a mass spectrum.

A key feature in making MS broadly applicable has been the development of ionisation processes for polar macromolecules such as proteins and peptides. In particular, ESI\(^{11}\) was first used for the ionisation of biomolecules by Yamashita and Fenn\(^{12}\) and is a "soft" low energy
– ionisation technique that enables solution-phase particles to be transferred to the gas-phase, generating protonated, even-electron ions.\textsuperscript{12}

\textbf{Figure 1.2: Schematic layout of electrospray ionization process (in positive ionization mode).} The analyte is injected through a hollow needle. A high voltage, typically 2–5 kV, is applied to the needle. Applying an electric field gradient between the sample capillary and the electrode for ion entry distorts the analyte solution’s meniscus, forming a Taylor cone that ejects charged droplets that reduce in size until the ions are released into the analyser. Charged droplets are released from the tip of the Taylor cone and evaporate as they move towards the MS inlet to generate analyte ions.\textsuperscript{13}

In ESI, the analyte is sampled directly from the solution through a capillary tube (Figure 1.2). A voltage usually of up to 4 kV is applied to the capillary, with an auxiliary electrode typically about 2-3 cm away.\textsuperscript{14,15} The auxiliary electrode is employed to generate an electric field that induces separation of the charges in solution in the capillary tip.\textsuperscript{16} This forms a Taylor cone due to the high density of charge at the solution meniscus. A fog of charged droplets, typically several hundred μm in diameter, emerged from the Taylor cone, is directed into the mass spectrometer's inlet due to a potential and pressure gradient. The ESI-source temperature and the nitrogen drying gas progressively reduce the droplets' size by evaporating the solvent. Evaporation of the volatile solvents decreases the droplets' size and subsequently increases the density of charge within the droplet. Eventually, the droplet reaches the point where the surface tension cannot sustain the Coulomb force of repulsion, known as the Rayleigh limit.\textsuperscript{11} At this point, Coulomb fission happens, and the primary droplet breaks down into tiny droplets,
typically 10-50 nm in diameter. Eventually, the charged droplets are ejected into the gaseous phase.\textsuperscript{13,17}

\textbf{Figure 1.3: Analyte transfer into the gaseous phase.} Outline of the different electrospray (ESI) mechanisms. Ion evaporation model (IEM) for small molecules (A), charged residue model (CRM) for folded protein (B) and the chain ejection model (CEM) for unfolded protein (C). This figure was adapted from Konermann et al.\textsuperscript{15}

The positive or negative charges will be dragged by a high electric field. The positive mode will produce a positively charged aerosol. Basic molecules, consisting of an amino group, would be analysed in positive ion mode because they are readily protonated to give positive
charge ions. In contrast, a negatively charged aerosol is produced if the mode is a negative ion. Negative ion mode is used for the analysis of acidic compounds.\textsuperscript{18}

Three different release models describe the final step in the ionisation process – the release of analyte ions into the gaseous phase. Lower molecular weight molecules are thought to move to the edge of the charged droplet, where the electric field favours ejection of the analyte with charge once the droplet is adequately small (<10 nm).\textsuperscript{11} This is referred to as the ion evaporation model (IEM), displayed in Figure 1.3. In the IEM, a single droplet will contain numerous small particles which are ejected from the droplet. Two competing mechanisms have been suggested for larger biomolecules, such as native, structured proteins. Ordered proteins in their native state follow the charged residue model (CRM), where they are left inside the charged droplet during the IEM of small molecules and evaporation of the droplet.\textsuperscript{19} Ultimately, this generates droplets carrying one analyte, which undergoes charge transfer when the solvent fully evaporates. (Figure 1.3).

In contrast, denatured, disordered and hydrophobic proteoforms display a different ESI behaviour described by the chain ejection model (CEM).\textsuperscript{20} In this procedure, as the higher-order structure has meddled due to hydrophobic and electrostatic interactions, residues usually buried within the folded structure become solvent accessible in the unfolded structure. As the folded polypeptide’s internal residues tend to be predominantly hydrophobic, these result in the polypeptide series being pushed to the droplet’s edge. Finally, similarly to the IEM, the Rayleigh limit provokes the polypeptide chain’s emission with a wide distribution of charge states.\textsuperscript{21}

1.1.2 Nano-electrospray

When low volumes and amounts of sample are available, nanoelectrospray ionisation (nESI) can be employed. nESI\textsuperscript{22,23} is a modified ESI variation and follows the equivalent ion ejection mechanisms. In nESI, glass capillaries are utilised as spray capillaries pulled out at one end either mechanically or by a laser to give orifices of just 1–10 µm in diameter. For adequate conductivity, the capillaries are sputter-coated with a conductive material, such as gold, or a thin alloy wire is injected into the capillary. The capillaries are loaded from the back with only 1–5 µL of the sample solution. A micromanipulator is needed to regulate the spray needle under stereomicroscopic or video camera control towards the auxiliary electrode orifice. The separating distance between the needle and the opening is only 0.5 to 2 mm. No liquid feed system is employed; the liquid’s dispersion and the produced flow rates of about 20–50 nL/min
are only due to the electrical fields' forces when voltages between 0.5 and 1 kV are applied. Some backpressure to the spray capillary by a gas-tight syringe is regularly implemented and helps launch the flow. Spray capillaries are only employed once, which proffers the further advantage that contamination and long-lasting memory effects are prohibited. 24

From the above, it can be concluded that nESI 's first primary advantage is the tiny sample (1–3 µL) required for analysis. 25 Moreover, the nESI needle can be placed relatively close to the MS's orifice, particularly for ESI sources employing the heated transfer capillary for which only crude and uncritical alignment is required. The gas flow into the transfer capillary opening ejects the nanospray aerosol thoroughly. Hence, the ESI mass spectra recorded under the same mass spectrometric conditions typically display more intense signals despite a lower liquid flow. Furthermore, the reduced sample size is less probable to contain high salt concentrations, enhancing sensitivity. Therefore, if only limited amounts of sample are available (as is usually the case in biochemical research), nESI is often the method of choice.

However, unfortunately, the performance of nESI is susceptible to minor variations during the study, for instance, deviations in the emitter geometry, electric current, flow rate, and needles alterations. Other standard issues include arcing and capillary tip breakage, mainly when highly conductive aqueous solutions are studied. 24

1.1.3 Charge states of proteins ions in ESI mass spectra and their relationship to protein structure in solution

It was shortly afterward the advent of ESI; in the late 1980s, MS was used to investigate intact proteoforms for the first time. 26,27 As well as determining the accurate intact mass of the polypeptides, it was apparent that hypotheses related to the structural form of both polypeptide and non-covalent proteoform complexes could be achieved under controlled experimental conditions. Thus, much was to be reached from investigating proteins ionised from physiologically relevant conditions that do not disturb their formations. This form of protein MS is known as native MS. Native MS allows the high sensitivity and resolution of MS to be applied to more detailed, structural studies of individual proteoforms. 28–30 It also enables studying proteoforms that are especially challenging to analyze by other systems, such as membrane proteins 31 or large viral capsids. 32 Finally, the high throughput of MS means that large-scale studies of multiple complexes can be conducted. 33
Furthermore, preliminary studies were conducted by Loo and co-workers,\textsuperscript{34} which affirmed that the accumulation of multiple charges (CSD) on a single protein ion during ionisation was exceedingly affected by the solvent composition associated with the solution state formation of a protein throughout ionisation. In the literature, CSD usually refers to the collection of charge states demonstrated for a particular molecule under a given set of experimental conditions.\textsuperscript{35} Since then, several other studies\textsuperscript{34,36,37} have provided convincing evidence that the number of protonation sites of a polypeptide is associated with the protein's solvent-accessible surface area.\textsuperscript{36}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{mass_spectra.png}
\caption{Example of a proteoform's theoretical mass spectra in \textbf{A) denatured conditions and B) native conditions}. The CSD observed in the native folded conformation is more compact at higher m/z than in denatured protein conformation.}
\end{figure}

For a higher molecular weight folded protein, the range of surface protonation is narrowed because the amino acidic side-chain groups are buried within the substrate, resulting in a narrow CSDs, low overall charge and higher m/z values (Figure 1.4). In contrast, in an denatured conformation, the same proteoform will have more exposed amino acid side chains and thus a higher chance for surface protonation, leading to higher charge states, wider CSD distribution, and lower m/z. High coulomb repulsion resulting from the increased surface
charge is also a causative factor towards a larger, more unfolded proteoform ion at higher charge states. These hypotheses correspond well with the view that as a protein unfolds, its solvent accessible surface area increases, and hence higher charge states are observed. Though it implements structural insights, considerable care should be taken to study CSD since it can be affected by experimental conditions such as high instrumental voltages and solvent composition.

1.1.4 Liquid Chromatography

Liquid chromatography (LC) facilitates the separation of molecules based on their relative affinities to a solid stationary phase and a liquid mobile phase and is a valuable tool to couple to MS to implement analytes’ orthogonal separation (Figure 1.5). This is especially valuable for analysing complex samples such as those resulting from proteomic experiments. LC fractions can be collected and further separated by electrophoresis before MALDI ionisation, but regularly LC is coupled to MS (LC-MS) using ESI ionisation, allowing for consecutive detection throughout the LC run. Traditionally, for protein and peptide separation, a reversed-phase column is implemented to separate molecules based on hydrophobicity, so the most soluble samples in water move through the column faster.\(^7\) LC-MS in this project has been used to obtain an accurate mass of polypeptides as it eliminates the necessity of transferring the protein into an ESI-compatible buffer.\(^36\) The non-volatile inorganic salts present in most physiological buffers are unsuited to ESI as they form salt adducts such as \([\text{M}+(n-1)\text{H}+\text{Na}]^+\) ions, with high ionisation efficiency or surface activity that compete with analytes during ion evaporation. Thus, this leads to a vastly decrease in the signal-to-noise ratio and obscures the determination of the accurate protein mass while also generating salt formations on the ESI capillary, likely leading to blockages. These issues are efficiently subdued during reversed-phase LC-MS, as inorganic salts pass through the apolar column, have a short interaction with the stationary phase, and are eluted first. This leaves the protein, now unfolded, to be eluted from the column by increasing non-polar solvent concentration.\(^38-40\)
1.1.5 Isotope Distributions

The isotopic distribution shown by biomolecules is a set of elemental and naturally occurring isotopic abundances. A typical case within proteins is the element of carbon. The most naturally occurring isotope of carbon is $^{12}\text{C}$, yet a smaller amount of the heavier isotope $^{13}\text{C}$ (~1.1 %) is also present (Table 1.1). The addition of one $^{13}\text{C}$ atom would raise the mass of a protein by roughly one Dalton; similarly, this shift in mass would rise with each addition of $^{13}\text{C}$. Heavy isotope merging is distinctive for each protein. As biomolecules carry numerous carbons, the resulting molecular mass follows a mathematical distribution of units. The units are chemically analogous while varying in isotopic distribution, referred to as 'isotopologues', whereas the composition is known as the 'isotope distribution' or 'isotopologue distribution'.

The possibility of heavy isotope incorporation is identical at every carbon site along the protein chain. Hence for polypeptides over 10 kDa, the molecular weight is sufficiently large that the isotope peak referred to as the monoisotopic peak has a low amplitude (Figure 1.6), which is inadequate for distinguishing it from the noise. This is why for biomolecules over 5 kDa, the average mass is most commonly used, or if the monoisotopic mass is required, the average mass is used from the observed isotope distribution; however, this process can introduce errors. The reality of the isotopologue distribution is even more complex as all the common biomolecule elements (C, N, H, O and S) have several stable isotopes that contribute to the total mass of the protein (Table 1.1). Interestingly, the configuration of the isotope distribution is primarily regulated by the size of the polypeptide chain. The available evidence suggests that as the biomolecule's molecular weight rises, so does the scope of inconsistency for merging $^{13}\text{C}$ and other less abundant isotopes.
Table 1.1: Natural isotopic distribution of common elements. This table lists the mass and per cent natural isotopic abundance for the stable nuclides observed in proteins.  

<table>
<thead>
<tr>
<th>Z</th>
<th>Name</th>
<th>Symbol</th>
<th>Mole Fraction</th>
<th>Relative Atomic Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hydrogen</td>
<td>$^1\text{H}$</td>
<td>[0.999 72 , 0.999 99]</td>
<td>1.007 825 0322</td>
</tr>
<tr>
<td></td>
<td>Deuterium</td>
<td>$^2\text{H}$</td>
<td>[0.000 01, 0.000 28]</td>
<td>2.014 101 7781</td>
</tr>
<tr>
<td>6</td>
<td>Carbon</td>
<td>$^{12}\text{C}$</td>
<td>[0.9884, 0.9904]</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$^{13}\text{C}$</td>
<td>[0.0096, 0.00116]</td>
<td>13,003 354 835</td>
</tr>
<tr>
<td>7</td>
<td>Nitrogen</td>
<td>$^{14}\text{N}$</td>
<td>[0.995 78 , 0.996 63]</td>
<td>14.003 074 004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$^{15}\text{N}$</td>
<td>[0.003 37, 0.004 22]</td>
<td>15.000 108 899</td>
</tr>
<tr>
<td>8</td>
<td>Oxygen</td>
<td>$^{16}\text{O}$</td>
<td>[0.997 38, 0.997 76]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$^{17}\text{O}$</td>
<td>[0.000 367, 0.000 400]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$^{18}\text{O}$</td>
<td>[0.001 87, 0.002 22]</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Sulphur</td>
<td>$^{32}\text{S}$</td>
<td>[0.9441 , 0.9529]</td>
<td>31.972 071 174</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$^{33}\text{S}$</td>
<td>[0.007 29 , 0.007 97]</td>
<td>32.971 458 910</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$^{34}\text{S}$</td>
<td>[0.0396 , 0.0477]</td>
<td>33.967 8670</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$^{36}\text{S}$</td>
<td>[0.000 129 , 0.000 187]</td>
<td>35.967 081</td>
</tr>
</tbody>
</table>
Figure 1.6: Impact of increasing molecular weight on the isotopic distributions. The number on the black is the relative abundance of the monoisotopic peak, and the number on red is the mass difference between the monoisotopic peak and the last one. As the polypeptide mass rises, the biomolecule's increased atomic complexity raises the merging of heavy isotopologues. This emerges as an increased number of isotopic peaks within the configuration spread over a higher mass scale. Consequently, the percentage abundance of the most abounding peak in the distribution becomes lower as the population's overall intensity is dispersed amongst each of the isotopic peaks in the distribution.

1.1.6 The isotopic fine structure of the protein

A definite amount of energy, known as the binding energy, is endowed to a nucleus through its formation. This is the force released if an atomic nucleus was to be formed from its constituent distinct subatomic particles. This release of energy translates into a "loss of mass" as a portion of a nucleus's mass is converted into that binding energy. Thus, the mass of any given atom or molecule is regularly slightly less than the sum of its protons (p), neutrons (n), and electrons (e) of which the atom is made off. This discrepancy is referred to as the
mass defect (MD), and it is exceptionally illustrated in Einstein’s equation $E = mc^2$, in which mass is related to energy. The following equation gives the mass defect (MD) of an atom:

$$\Delta m = Z(m_p + m_e) + (A - Z)m_n - m_{\text{atom}}$$

Where $\Delta m$ is the MD (amu), $m_p$ is the mass of a $p$ (1.007277 amu), $m_n$ is the mass of a $n$ (1.008665 amu), $m_e$ is the mass of an $e^-$ (0.000548597 amu), $m_{\text{atom}}$ is the mass of the nuclide (amu), $Z$ is the atomic number (number of $p$), and $A$ is the mass number (number of nucleons).

For example, iodine, with $Z = 53$, a nominal mass $A = 127$ –defined as the integer mass of an atom, and an atom mass of 126.90004 amu has an MD $\Delta m = 1.1574$.

![Mass Defect Graph](image)

**Figure 1.7: Mass defects of stable isotopes.** The relative isotopic mass defect for common stable isotopes was calculated as the difference in mass between the monoisotopic mass (MIM) of the most abundant isotope and its nominal mass (NM).\(^{47}\)

The definition of MD is in the field of MS slightly differs from that above and just refers to the difference between the nominal mass (NM) of an atom (i.e., 127 amu for iodine) and its monoisotopic mass (MIM) (126.904473 amu), which gives an MD of -0.09553 for iodine. By convention, carbon was assigned an MD of zero, while any other existing isotope has either a positive or negative MD related to its binding energy to $^{12}$C, Figure 1.7.\(^{48}\)
The effect of the mass defect on the isotopic distribution of a proteoform is illustrated in figure 1.4 and is defined as the isotopic fine structure of the isotopic distribution. A modelled peptide was chosen with a molecular formula $\text{C}_{69}\text{H}_{114}\text{N}_{20}\text{O}_{26}\text{S}_{1}$ (Figure 1.8). As highlighted in Figure 1.8, the peptide exhibits 6 isotopic peaks, with the most abundant ones being $M$, $M+1$ and the $M+2$ peaks. The $M+1$ peak is formed due to a particular heavy isotope contributing an extra $\sim 1$ Da to the peptide mass. This additional 1 Da can be provided by $^{15}\text{N}$ (+0.99704 Da), $^{33}\text{S}$ (+0.99939 Da), $^{13}\text{C}$ (+1.00336 Da) or $^{2}\text{H}$ (+1.00628 Da) as displayed in Figure 1.8 B. Each isotope's discrete abundance contributes to the total abundance of the $M+1$ peak in Figure 1.8. Likewise, the $M+2$ peak derives from either a single isotope adding $\sim 2$ Da or two isotopes that provide $\sim 1$ Da. Due to the number of carbon atoms in the peptide and the natural abundance of $^{13}\text{C}$ (1.07 %), carbon has the most noticeable impact in Figure 1.8.

Table 1.2: Nominal- and monoisotopic mass and mass defect of elements found in proteins and iodine.

<table>
<thead>
<tr>
<th>Atom</th>
<th>Nominal mass/Da</th>
<th>Monoisotopic mass/Da</th>
<th>Mass defect/Da</th>
</tr>
</thead>
<tbody>
<tr>
<td>$^{12}\text{C}$</td>
<td>12</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>$^{1}\text{H}$</td>
<td>1</td>
<td>1.007 825 0322</td>
<td>0.007825032</td>
</tr>
<tr>
<td>$^{14}\text{N}$</td>
<td>14</td>
<td>14.003 074 004</td>
<td>0.003074004</td>
</tr>
<tr>
<td>$^{16}\text{O}$</td>
<td>16</td>
<td>15.994 914 619</td>
<td>-0.005085381</td>
</tr>
<tr>
<td>$^{32}\text{S}$</td>
<td>32</td>
<td>31.972 071 174</td>
<td>-0.027928826</td>
</tr>
<tr>
<td>$^{127}\text{I}$</td>
<td>127</td>
<td>126.904 47(3)</td>
<td>-0.095527</td>
</tr>
</tbody>
</table>

The high-resolution acquisition is required to observe this small peptide's isotopic fine structure displayed in Figure 1.8 (>600,000). Predictably, as the peptide mass increases, it becomes less probable that isotopic fine resolution will be adequately resolved, and the fusion of heavier stable isotope merge will rise in complexity. Generating the statistical display of isotopic peaks, we note in the isotopic distribution (Figure 1.8).
Figure 1.8: Isotopic fine structure of a small peptide. The theoretical mathematical model of a small molecular weight polypeptide chain \([C_{69}H_{114}N_{20}O_{26}S_1]^{3+}\) (A) and a closer shot of both the M+1 (B) and the M+2 peak (C) to demonstrate the isotopic fine structure of the peptide.
1.2 Mass Resolution and Resolving Power

MS's mass resolution is the capacity to resolve between two adjacent peaks in a mass spectrum. An MS's resolving power is defined as $m/\Delta m$, where $m$ is the nominal mass for a particular peak in the mass spectrum, and $\Delta m$ is peak width at half maximum of the peak height and is known as the full-width half maximum (FWHM).\textsuperscript{52} The shorter the width band of the adjacent peaks, the better separation can be achieved. This is demonstrated in Figure 1.9, where the chemical structure of a low molecular weight protein, ($C_{769}H_{1212}N_{210}O_{218}S_2$)$^{10+}$, is displayed. It is apparent from Figure 1.9 that as the spectral resolution improves, the spectrum changes from a single peak Gaussian distribution (blue line) to the isotopic distribution (black line), although both describe the same polypeptide pattern. At the lowest resolution (blue line), the peak's average mass can be only determined.

![Graph of varying resolution of a protein](image)

**Figure 1.9: Varying resolution of a protein.** At varying resolutions, the information deduced from a spectrum varies. Lower resolution (5000, blue line) generates broader peaks from which an average mass can be determined. As the resolution improves (12,000, green line; 20,000, dashed line), the isotopic peaks become evident. Further increase in resolution (100,000, black line) enables the isotopic distribution to be fully resolved.

Furthermore, this peak has a significant FWHM implying that two species at adjacent $m/z$ may be overlapping as a single stretched or modified distribution making assignment more
challenging. As the resolution improves to the highest resolution (black line), the monoisotopic peak is visible, enabling the polypeptide mass's accurate measurement. Each peak of the isotopic distribution is regularly spaced peaked and can give information about the charge state, and various configurations that partially overlay can be determined separately.

The resolving power achievable can be influenced by several factors, such as the detector capabilities of the MS employed, the mass of the analyte, the charge state distribution of the analyte and the presence of attached salt adducts. As mentioned earlier, adducts like sodium and potassium ions result from trace salt concentrations in the ESI buffer. They induce an increase in the protein mass and generate several smaller secondary distributions, overlaying the primary protein distribution. This is critical for higher molecular weight species, as the distribution for both the proteoform and the adduct may be observed as a singular, merged distribution in the spectrum. Thus, the calculated average mass would not accurately reflect the protein species' mass and will ultimately limit the protein analysis if undetected.
1.2.1 Mass analysers

A mass analyser separates ions within an MS according to their $m/z$ ratio. Mass analysers vary in their ability to separate ions of comparable or near-identical $m/z$ (resolving power), the range of ions detected, and the last transmission efficiency. Several types of mass analysers are employed in modern proteomics environments since each has its strengths and weaknesses in various applications.\textsuperscript{54}
1.2.1.1 Quadrupoles

A quadrupole is possibly the most commonly used mass analyser for ion separation according to \(m/z\). It consists of four parallel parabolic rods, connected as opposite pairs of electrodes, to which either radio frequency (RF) or direct current (DC) voltages are applied (Figure 1.10). Under a set electric field, only certain ions within a narrow \(m/z\) window will have a stable trajectory to facilitate transmission through the quadrupole, whilst all the rest ions with unstable paths will collide with the rods and walls of the mass analyser.

*Figure 1.10: A schematic of quadrupole electrodes.* The poles are arranged as two pairs of electrodes, as illustrated in the electric circuit. The potential differences established between the electrodes leads to a hyperbolic field. Ions are attracted towards the oppositely charged electrodes, but the applied field results in potential switching of the electrodes and thus, ions oscillate in the \(x\) and \(y\) directions while travelling in the \(z\)-direction.

A quadrupole mass analyser can utilise multiple modes. In the RF-only mode, a wide range of \(m/z\) ions traverse the quadrupole (in effect, the device acts as an ion guide). Operated in the static electric field, a quadrupole only enables a narrow window of ions through. By altering
the voltages applied to the quadrupole, a wide range of \( m/z \) ions can be determined over a short period of time. Quadrupole mass analysers usually have unit mass resolution and low sensitivity when in MS scan mode but high sensitivity and broader dynamic range when employed in a targeted analysis for distinct compounds.\(^{55,56}\)

Newton's motion equations can be employed to determine the charged ions' motion in an electric field, and Mathieu's equation can be employed to conclude which \( m/z \) rates will be transmitted for the applied voltages.

\[
\frac{d^2 x}{dt^2} = -\left(\frac{e}{m}\right) \frac{[U + V\cos(\omega t)]}{r_0^2} x
\]

Equation 1

\[
\frac{d^2 y}{dt^2} = \left(\frac{e}{m}\right) \frac{[U + V\cos(\omega t)]}{r_0^2} y
\]

Equation 2

\[
\frac{d^2 z}{dt^2} = 0
\]

Equation 3

Where \( e \) is the magnitude of electron charge, \( m \) is the mass, \( u \) is the DC voltage, \( V\cos(\omega t) \) is the radio frequency voltage, and \( r_0^2 \) is the distance of separation of rods.

Only ions of a specific \( m/z \) will have a steady trajectory to facilitate transmission through the quadrupole for any given applied voltage - the residual ions will have weak trajectories and will be moved into the rods. This enables specific \( m/z \) to be chosen from complex mixtures of ions to deduct more experiments.

1.2.1.2 Time-of-flight Mass Spectrometry

Time-of-flight (ToF) mass analyser relies on the separation of ions based on the time taken to travel through a field-free drift zone to a detector.\(^{57,58}\) Pulses of ions are expelled and accelerated by electric potential difference applied between an electrode and the extraction grid to give them equivalent kinetic energies. The ions then enter the field-free drift region where the spread of the velocities of the ions, due to their various \( m/z \) ratios, lead to separation of the ions (Figure 1.11).\(^{59}\)
**Figure 1.11: A schematic diagram of linear time-of-flight mass analyser demonstrating ions with varying m/z ratios.** The ion with a smaller m/z ratio (orange) arrives at the detector before the ion with the higher m/z ratio (green) due to its higher velocity.

The m/z of analytes is defined by measuring the total time that ions need to pass through the field-free drift zone between the source and the detector. Before it enters this region, an ion with mass (m) and total charge (q=ze) is accelerated by the applied potential (Vs) to give the kinetic energy (Ek):

\[ E_k = \frac{mv^2}{2} = qV_s = zeV_s \]

Equation 4

The velocity of the particle leaving the ion source is given by simple rearrangement of Equation 4 into:

\[ v = \left(\frac{2zeV_s}{m}\right)^{1/2} \]

Equation 5

After the initial acceleration, the ion moves in a straight line at a steady velocity towards the detector. The time (t) needed to travel the distance (L) and reach the detector is given as:

\[ t = \frac{L}{v} \]

Equation 6

Substituting the value of v in Equation 5 gives:
Confirming that $m/z$ can be determined from a measurement of $t$ and that the higher the ion's $m/z$ ratio, the longer it needs to reach the detector.

The major shortcoming of the linear ToF mass analyser is its poor mass resolution. Variables such as the length of the ion formation pulse and the variation of initial $E_k$ generate a population of flight times for ions with the equivalent $m/z$ ratio. The heterogeneous primary energies of the ions before acceleration leads to ions with equal $m/z$ passing through the field-free drift zone at varying velocities and reaching the sensor at slightly different times. Reflectrons can be attached to the drift sections of ToF instruments to boost the mass resolution (Figure 1.12). The reflectron device extends the length of the drift zone and reverses the course of ion velocities, adjusting for shifts in the kinetic energies. Ions with higher primary kinetic energy penetrate further into the reflectron and thus need more time to be reflected.  

![Diagram of a reflectron ToF mass analyser](image)

**Figure 1.12: Schematic graph of a reflectron ToF mass analyser.** Ions with a higher primary kinetic energy (yellow) enter further into the reflectron, whilst ions with less kinetic energy (green) are reflected faster. This raises the mass resolution achieved.
ToF-MS instruments give high sensitivity and relatively fast acquisition time. Modern ToF systems regularly have a mass resolving power of up to ~50,000 and mass accuracy of 1–5 ppm after calibrating the system.\textsuperscript{60}

**1.2.1.3 Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS)**

In 1974, Melvin B. Comisarow and Alan G. Marshall invented a new type of mass spectrometer, FT-ICR MS.\textsuperscript{61} The technique relies on determining an ions $m/z$ by monitoring its cyclotron frequency in an established magnetic field.\textsuperscript{62}

There are several advantages to this technique. Most notably, it offers the highest resolving power and mass accuracy compared to all types of mass spectrometers. In addition, it can be adjusted for both analytical and physical chemistry measurements. Moreover, high-resolution mass spectra from ions formed by almost every ionisation method can be obtained.\textsuperscript{63–67} The technique is also suitable for a wide range of tandem mass spectrometry approaches, and gas-phase fragmentation experiments can be performed to study ion chemistry as well as photochemistry.\textsuperscript{68–72} All the above features made FT-ICR a valuable tool for the study of various samples such as proteomics, metabolomics, petroleomics.\textsuperscript{73}

In FT-ICR, the $m/z$ of an ion is determined employing the defined cyclotron frequency ($\omega$). Traditionally with protein samples, nESI is employed to generate gaseous ions. The ions are accelerated by increasing the vacuum until restrained within the Penning trap detection cell\textsuperscript{74} held within a superconducting magnet cooled with liquid helium. At this point, the ions’ movement is considered to be within a homogeneous static magnetic field (B), within a voltage-gated cell that contains the ions. Therefore, we can reconstruct Newton's second law of motion (Equation 7) in an electromagnetic field to generate the left section of Equation 8, where $q$ and $v$ are charge and velocity, respectively.\textsuperscript{75} As ions are free to move in the $xy$ axis, the Lorentz force is perpendicular, causing an angular acceleration to the ions, defined in the right section of Equation 8 (Figure 1.13).

$$F = m \times a$$

**Equation 8**
Where \( m \) is mass and \( a \) is acceleration.

\[
F = qv_{xy}B = m\frac{v_{xy}^2}{r}
\]

Equation 9

**Figure 1.13: The ICR detection cell’s orientation and direction of the electromagnetic field \( B \) is shown on the left side.** In this magnetic field, ions will orbit the \( z \)-axis along the \( xy \) axis as displayed to the right; ions of counter-charges will move in a circular motion in reverse directions. This Lorentz motion is associated with the \( m/z \) and sets the basis of FT-ICR MS.

Equation 10 illustrates the 'ion's cyclotron angular frequency, \( \omega_c \), without applying any other electric field:

\[
\omega_c = \frac{qB}{m}
\]

Equation 10

In reality, ions need to be constrained in a defined area (for instance, an ICR cell) for frequency measurement. For this reason, the application of an additional electric field to the ions is inevitable (as a trapping voltage). Therefore, the ions are enclosed axially and radially by a trapping electric field and the Lorentz force, respectively. In a standard closed cylindrically ICR cell, the ions are introduced into the cell in the axial direction. Therefore, Equation 10 is not valid to determine the ions' movement when a trapping potential is applied.
The ions' movement under the quadrupolar potential is due to three modes. These are an axial vibration parallel to B with frequency $\omega_z$, a cyclotron movement with diminished frequency $\omega^+ \text{ and a rotational motion vertical to B with magnetron frequency, } \omega^-$. The cyclotron motion with decreased frequency is measured to determine the $m/z$. The following equation illustrates this movement:

$$\omega_{\pm} = \frac{\omega_c}{2} \pm \frac{\sqrt{(\omega_c^2 - \omega_t^2)}}{2}$$

Equation 11

As ions enter into and are trapped in the detection cell experience a reduced cyclotron motion ($\omega^+$). It is essential to remember that $\omega^+$ will not generate a detectable signal within the detection cell. To detect the ions, an oscillating electric field $E(t)$ in the $y$-direction, described by Equation 12, is applied within the two counter electrodes, as illustrated in Figure 1.14.

$$E(t) = E_0 \cos \omega_c t j$$

Equation 12

$$E_0 = \frac{2V_0}{d} = \frac{V_{p-p}}{d}$$

Equation 13

$$E(t) = E_R(t) + E_L(t)$$

Equation 14

This electric field is generated by applying voltages of $+V_0$ and $-V_0$ to the two parallel conductive plates separated by a distance of $d$ meters (Equation 13). The oscillation of $E(t)$ can be written instead as two alternating components (Equation 14) rotating in opposite directions, $E_R(t)$ orbiting in the same direction of the ion and $E_L(t)$ in the opposite. The $E_R(t)$ rotates with the same sense as the positive ions, so it will push the ions, while the $E_L(t)$ is off-resonance; thus, it will have little effect on the ion's motion. The post excitation radius is determined by employing Equation 15, and is independent of the $m/z$. Therefore, all ions can be excited to the same radius by a sweep across a range of radio frequencies.

$$r = \frac{V_{p-p} T_{\text{excite}}}{2dB_0}$$

Equation 15
Apart from the parallel electrodes which apply the oscillating electric field and excite the ions, the detection cell also has two detector plates (Figure 1.14). Ions move around the cell with the same post-excitation radius (Equation 15) but with different velocity. The discrete ion packets of all m/z values induce a signal known as the free induction decay (FID), from which individual frequencies are acquired through Equation 11 (Figure 1.14).  

**Figure 1.14: Detection of ions within an ICR detection cell.** Ions are excited with a radiofrequency pulse utilising a potential difference between the two detection plates (E) to a greater radius. At this point, the ions will induce a current across the detection plates (D) to generate a complicated overlap of several overlapping frequencies. These radio frequencies are separated by the Fourier transform function and converted into the m/z of the ion.
1.2.1.4 Orbitrap Mass Spectrometry

The orbitrap is a high-resolution mass analyser like FT-ICR. It is based on a primer ion storage device, the Kingdon trap, which engages orbital trapping in a constant electric field. A cylindrical ion trap with a central wire and a voltage applied to both advances elliptical oscillation of ions accelerated into the trap. In the orbitrap, a central spindle electrode is run coaxially through an outer coaxial (barrel-shaped) anode (outer electrode) (Figure 1.15). Ions are injected parallel and off-centre to the z-axis and are restrained between voltages implemented across the electrodes. Assuring the ions move forward to the inner electrode and oscillate around the central spindle.

![Figure 1.15: Oscillation of ions within the orbitrap detector.](image)

Ions are introduced to the orbitrap parallel to the z-axis and held between the voltages applied to the central spindle and outer barrel. The axial frequency ($\omega$) of the ion along the z-axis is dependent on the m/z of the ion and detected as a free induction decay (FID) and the axial frequency identified using a Fourier transform. From this the m/z ratio of the present ions can be calculated and shown within a mass spectrum.

$$\omega = \sqrt{\frac{k}{m/z}}$$

Equation 16
$z(t) = z_0 \cos(\omega t) + \sqrt{\left(\frac{2E_z}{k}\right)} \sin(\omega t)$

Equation 17

The ions exhibit a complex axial rotation around the spindle electrode (Equation 17), but fortunately, the defined axial frequency ($\omega$) is essentially the effect of the m/z of the ion (Equation 16). Likewise to FT-ICR, the move of the ions is detected as they advance towards the sensors (Figure 1.15), forming a complex FID, identifying the axial frequency ($\omega$) utilising the Fourier transform and employing (Equation 16) can determine the m/z of the ions within the detector.

1.2.1.5 Hybrid Instruments

The combination of quadrupole mass analyser in the same apparatus as the FT-ICR and ToF mass analysers enables the selective isolation of ions of specific m/z for further gas phase manipulation and detection. This process is well known as tandem Mass spectrometry (MS/MS). By employing gas-phase fragmentation techniques after isolation – fragment ions can be linked to parent m/z values. In protein/peptide mass spectrometry, this technique allows amino acid sequence information to be ascertained. Employing a hybrid system can assist in protein identification and allows de novo sequencing of species with unknown primary sequences. Furthermore, this process minimises the number of ions in the higher-resolution mass analysers raising the m/z resolution obtained as charge-charge interactions are reduced.

A hybrid instrument, the Bruker Daltonics Solarix-Qe FT-ICR MS, was employed extensively to record the MS and MS/MS spectra displayed in this dissertation. Figure 1.16 displays a simplified overview of the main segments on this instrument. Ions are formed by ESI and travel through the flight tube to the ICR cell through several sets of ion optics, including ion funnels and steering plates. The ICR detection cell is located within a 12 T superconducting magnet. A mass resolving quadruple and collision cell hexapole reside within the source and the transfer optics. For MS/MS, ions of a particular m/z can be isolated in the quadrupole followed by CID in the hexapole or ECD in the ICR cell. The ability to precisely determine the m/z of
large multiply charged precursors combing with the fragment ions makes FT-ICR MS a compelling tool for proteins characterisation.

Figure 1.16: Simplified overview of the Bruker Daltonics SolariX-Qe FT-ICR mass spectrometer. Regions highlighted in orange indicate where MS/MS may take place. Ions of a particular m/z can be selected and isolated in the quadrupole prior to CID in the collision cell hexapole, or ECD in the ECD cathode.

1.2.1.6 Ion Mobility Mass Spectrometry

Ion mobility coupled to mass spectrometry is an approach employed to investigate the structure of proteofoms. However, unlike MS, the separation of ions is based on their size, shape and charge, in other words, their mobility. Ions are entered in pulses and moved by the action of a linear electric field through a drift tube filled with an inert gas. Larger ions undergo more collisions with the gas molecules and slowly cross through the drift tube. Then, the separated ions are injected into the MS and are further studied. Utilising IM can provide information about ion geometry in CCS (Ω) values. CCS values are the average area of each analyte ion that physically interacts with an inert gas molecule. IM-MS can be exploited to interrogate the gas-phase stability of protein-ligand ions. By CIU-IMMS, the CCS values evidence the unfolding pathway of the gas-phase complexes. Typically, the internal energy of ions rises via collisions with the drift gas. To obtain a CIU fingerprint of a complex, the collision energy voltage is increased stepwise to monitor its influence on the gas-phase conformation of ions. The increase in the activation energy often leads to the observation of higher CCS values, indicating an unfolding event.
1.3 Protein Mass Spectrometry

In the last two decades, MS has found extensive application in proteomics - the analysis of an entire population of proteins in biophysical systems. Proteomes are highly complex systems, and thus there are various experimental workflows for their study, broadly classified into two main strategies, which have been termed as (i) bottom-up and (ii) top-down mass spectrometry.

1.3.1 Bottom-Up Mass Spectrometry

A conventional approach and the mainstay of proteomic workflows, the bottom-up strategy can be used to analyse large numbers of proteins simultaneously (shotgun proteomics). In this workflow, proteins are enzymatically digested into smaller peptides using a protease such as trypsin prior to MS (Figure 1.17). Proteases cleave the polypeptide chains into predictable peptide fragments relying on the amino acid sequence and cleavage site of the enzyme employed. Afterwards, the masses of generated peptides and fragmentation data are run against a database to enable the identification of the parent protein for each peptide.

The standard bottom-up workflow is known as data-dependent acquisition (DDA). In a DDA approach, peptide masses are determined in one MS scan and as many of the peptides identified in this scan as feasible are then fragmented in the MS to acquire detailed sequence data. This can be performed in a single sub-second cycle, indicating a complete peptide study can be carried out online with LC. It is mainly by this process that vast amounts of data have been produced, including mapping of almost the entire human proteome. At the point of writing, the Human Proteome Map database contains 30,057 identified proteoforms from 17,294 genes, all resolved from the determination of 293,700 peptide sequences.

This method has been developed to address a crucial issue in protein MS, the large size of the intact polypeptide chains, making intact mass analysis challenging. As the polypeptide chain increases its molecular weight, the intensity of the observed peaks drops since high molecular weight species will have broad isotopic distributions, spreading the signal across several peaks. Proteolysis before MS makes the bottom-up MS impartial to protein size and highly sensitive, which is one of the significant advantages of this approach. Due to these advantages, more recent attention has focused on providing bottom-up applications for de novo peptide sequencing in which enzymatic digestion is performed on intact proteins.
However, this method's benefit is marred by the disadvantages associated with the partial sequence coverage of the protein and, thus, no information regarding the presence of a post-translational modification (PTM) within polypeptide chains. When mapping PTMs using a bottom-up approach, it can be feasible to localise the modification based on the amino acid residues present alone; for instance, glutathionylation is only observed on cysteines. Nevertheless, if numerous potential localisation sites exist, for instance, a single glutathionylation is observed on a tryptic peptide carrying two cysteine residues, it is impossible to locate the modified amino acid residue following a classical bottom-up approach. Furthermore, when studying differentially mutated proteins, the combination that PTMs occurs in cannot be defined because the peptides' mass cannot be directly correlated to the mass of the parent proteoform. These shortcomings can be overcome employing tandem mass spectrometry for protein and PTMs assignment.

**Figure 1.17: Bottom-up approach.** In the Bottom-up method, enzymatic digestion is employed to cleave intact proteins into peptides. Peptides are determined through MS, and protein determination can occur by identifying peptides matched to a database. Peptides can be fragmented further to generate extensive sequence information and locate a post-translational modification to one distinct amino acid residue.
1.3.2 Top-Down Mass Spectrometry

Top-down MS is an emerging process for the analysis of intact proteins. The term was originated as a contrast with the bottom-up approach. In this approach, the entire mass of the protein is first defined. A particular intact protein ion of interest can be isolated within the instrument prior to fragmentation. In contrast to the bottom-up approach, the fragmentation of the protein occurs within the MS (Figure 1.18).

**Figure 1.18: Top-down approach.** In the Top-Down approach, intact proteins are introduced straight to the MS. The precise mass of the intact macromolecule is determined, followed by direct fragmentation. The generated precursor and fragments are then matched to potential sequences from a database. This provides the potential for the complete assignment of PTMs.

The exact mass of an intact proteoform or protein complex can provide vital information such as the nature of PTMs and their respective stoichiometries. Knowledge of the molecular weight combined with the capacity to isolate individual modification states or isoforms within the MS suggests that the relationship between the parent polypeptide and fragment ions is known. Top-down MS, hence, enables the unambiguous assignment of protein PTMs states to be resolved.
Moreover, top-down MS leads to far more extensive cleavage of inter-residue bonds opposed to bottom-up workflow because the entire protein sequence is available for investigation rather than solely several peptides. Still, the number of cleaved inter-residue bonds regularly diminishes with rising protein mass due to the lowered dissociation efficiency and fragmentation over the internal region of higher molecular weight proteins (> ~30 kDa) is regularly not adequate to determine the site of a PTMs unambiguously (knowledge of the intact mass does, though, imply their presence). Over a molecular mass of ~50 kDa, the parent protein fragmentation becomes highly problematic due to the increased complexity of the molecular ions gas-phase tertiary structure. The intact protein's cleavage produces a complicated MS/MS spectrum with numerous overlapping, multiply-charged fragment ions (Figure 1.19). Thus, high-performance mass spectrometers that provide superior resolving power and mass accuracy, such as FT-ICR instruments (1.2.1.3), are required for the top-down study of proteoforms. Apart from the protein size, the instruments high cost is a limiting factor in top-down MS protein research. These deficiencies of protein size can be overcome utilising the middle down approach outlined below.
Figure 1.19: Top-Down MS/MS spectrum of a 13 kDa protein recorded on an FT ICR MS with isolated precursor ion highlighted (♦). The FT ICR MS high-resolution allows for distinguishing the isotope distributions of the generated fragment ions. The inset zoom illustrates the overlapping distributions of the $z_{21}^{3+}$, $z_{30}^{4+}$ and $c_{43}^{6+}$ ions and the separate ion distributions are highlighted by the modelled overlaid theoretical distributions.
1.3.3 Middle-Down Mass Spectrometry

Middle-down MS is a process that intends to merge the advantages of bottom-up and top-down strategies while reducing their limitations, as mentioned earlier. Here, likewise to bottom-up, polypeptide chains are digested, yet, limited proteolysis is applied to generate peptides more amenable to MS, yet large enough (usually > 3 kDa) to boost the average size distribution of the generated peptides. The induced peptides can be then separated, followed by precise mass measurement of the precursor ion and fragmented to produce a moderately complex MS/MS spectrum comparable to top-down MS. The whole process of a standard middle-down analysis is illustrated in Figure 1.1.

![Middle-down Workflow](image)

**Figure 1.1: Middle-down approach.** The middle-down workflow to PTMs assignment by MS. The intact protein is proteolysed into large peptides utilising enzymes or a chemical cleaving reagent. The large polypeptide chains may contain several PTMs sites yet be small enough to enable extensive fragmentation, thus allowing univocal localisation of PTMs. Middle-down MS is especially useful for providing extensive sequence coverage of a high molecular weight protein.

Despite its multiple advantages, the middle down approach suffers from several significant drawbacks. The generation of few and long fragments occurs when enzymes are utilised to cleave distinct or rare amino acids residues. However, much of the research up to now has consistently shown that no suitable natural enzymes are proteolysing at these sites. One
approach to address this problem involves using chemical-mediated cleavage (such as CNBr, 3-bromo-3-methyl-2-(2-nitrophenyl)sulfonylindole (BNPS-Skatole), and 2-nitro-5 thiocyanatobenzoic acid (NTCB acid)) to target these rare amino acid residues (Met, Trp, and Cys), for which no particular proteases are assigned. However, the key challenge with these applications is the generation of multiple peptides with overlapping isotopic distributions. In MS, broad isotopic patterns have been confirmed to be one of the principal barriers in experimentally assigning the masses of peptides with adequate precision.

1.4 Fragmentation

The fragmentation of biomacromolecules in the gas phase, whether caused by collisions, photons, or some sort of charge exchange method, is central to the field of proteomics. In addition to providing a means to probe structure, these reactions frame the base for peptide identification either de novo or through a library search. Specifically, each of these ion activation methods aims to induce specific fragmentation of the amide backbone at any of the three bonds backbone of the repeating polymer chain (Figure 1.2: A schematic of the fragments that may arise upon top-down fragmentation study of a polypeptide backbone. a, b and c ions are N-terminus fragments and x, y and z ions are C-terminus fragments. The subscript indicates the number of amino acid residues in the generated fragment, counting from the N-terminus for a, b and c ions or the C-terminus for x, y and z ions.

The lowercase number indicates the position of the fragmentation in the polypeptide sequence. Sequence coverage values are the percentage of fragmentation sites observed relative to the amino acid sequence’s length. As the most extensively applied fragmentation process, collision-induced dissociation (CID) has attributed a central role in the proteomics field, but several other techniques, such as ECD/ETD have drawn significant attention in recent years.

) while avoiding cleavage of the side-chain functional groups or any side-chain PTMs.
Figure 1.2: A schematic of the fragments that may arise upon top-down fragmentation study of a polypeptide backbone. $a$, $b$ and $c$ ions are $N$-terminus fragments and $x$, $y$ and $z$ ions are $C$-terminus fragments. The subscript indicates the number of amino acid residues in the generated fragment, counting from the $N$-terminus for $a$, $b$ and $c$ ions or the $C$-terminus for $x$, $y$ and $z$ ions.

The lowercase number indicates the position of the fragmentation in the polypeptide sequence. Sequence coverage values are the percentage of fragmentation sites observed relative to the amino acid sequence’s length. As the most extensively applied fragmentation process, collision-induced dissociation (CID) has attributed a central role in the proteomics field,\textsuperscript{104–107} but several other techniques, such as ECD/ETD\textsuperscript{108,109} have drawn significant attention in recent years.

1.4.1 Collision Induced Dissociation

CID, or alternatively known as collisional activated dissociation (CAD), is the most common fragmentation process for tandem MS. In CID, the parent ion’s internal energy is increased by energetic collisions with neutral gas atoms (typically argon, helium or nitrogen). In the case of peptide or protein fragmentation, this increase in internal energy induces peptide backbone’s cleavage and fragment ions formation.\textsuperscript{110–113} Depending on the type of instruments used for tandem MS either higher-energy C-trap dissociation (HCD) (on TOF and sector instruments), or low-energy CID (usually performed on multipole e.g., quadrupole and ion trap apparatus), can be carried out. The CID experiments performed here were executed in a hexapole collisional cell, and ions were cleaved employing low-energy CID.
Low-energy CID belongs to the group of so-called, "slow-heating" procedures, along with several other MS/MS techniques, including IRMPD, sustained off-resonance irradiation CID (SORI-CID) and blackbody infrared radiative dissociation (BIRD).\cite{113-117} In these processes, cleavage is mainly a result of the internal excitation of the parent ion.\cite{118} In the case of CID, each collision with the inert neutral gas atoms increases the internal energy until sufficient internal energy is accumulated within the ion to overcome the dissociation barrier, leading to dissociation. Fragmentation of proteoforms employing slow-heating systems results primarily in the backbone amide bond’s cleavage, generating $b$ and $y$ productions. Thus, fragmentation of the peptide backbone is usually not random, and particular fragment ions, for instance, those emerging from cleavage of the facile amide bond on the $N$-terminus side of proline may prevail the spectrum.\cite{119} Furthermore, the cleavage of a labile PTM, such as phosphorylation and small neutral losses such as $\text{H}_2\text{O}$ and $\text{NH}_3$ are also regularly observed.\cite{120}

CID is the most robust and readily used fragmentation process of proteoform ions in MS/MS because it is relatively simple (in terms of its application), efficient and reproducible.\cite{121} Yet, a shortcoming of CID that is particularly vital to top-down MS (concerning the study of intact proteins) is the inverse correlation among molecular weight and dissociation efficiency. Literature appears to validate that the rise of the macromolecule mass leads to the rise in the number of degrees of freedom. This implies that many more energy channels exist whereby an excited ion’s internal energy may be dissipated.\cite{102,114,121} The net consequence is a drop in the yield of generated ions as molecular mass increases, and hence low amino acid sequence coverage is obtained as proteoform size rises, thus application in top-down fragmentation studies is limited. Furthermore, the tendency for CID to result in losses of side-chain (especially with high-energy CID) is less than ideal for the analysis of protein PTM.\cite{92}

1.4.2 Electron Capture Dissociation and ETD

To address the issues outlined above, alternative ion activation approaches, electron-capture dissociation (ECD) and electron transfer dissociation (ETD) (collectively known as ExD) were introduced.\cite{122} ExD methods function via the move of a singular low energy electron to a multiply protonated proteoform. The electron is transmitted either instantly (ECD) or from a previously generated radical anion (ETD).\cite{123} Compared with the conventional slow-heating activation processes as CID, ExD fragments peptide backbone $N$–$\text{Ca}(c$ and $z$ type ions) and disulfide bridges, while preserving the labile PTM to a significantly higher degree.\cite{124} Thus, this makes ExD methods a powerful tool towards the structural study of several types of biomacromolecules, including proteoforms, carbohydrates, oligonucleotides, and others.\cite{122}
For the product ions of ECD to be visible by MS, the parent (positive) ion needs to be at least doubly charged since the capture of an electron by a singly charged positive ion would end in a zero charge.\textsuperscript{125} The cleavage of the radical species occurs very rapidly, and the possibility of fragmentation of the N-Cα bond is hardly affected by the adjacent amino acid residue. Hence, the method is non-ergodic and does not favour specific fragments. Fragmentation is only unlikely to happen at the N-terminal side of proline residues due to the ring structure.\textsuperscript{126} Thus, even though ECD's efficiency is lower, a higher sequence coverage can often be achieved compared to CID.\textsuperscript{125} While in ETD, electrons are conveyed to the polypeptide ion by radical anions. This reaction can be conducted in an ion trap compatible with most high-resolution mass analysers, such as Orbitrap FTMS or TOF MS.\textsuperscript{124} In contrast, ECD has predominantly been conducted only on FT-ICR MS due to the need for sufficient overlap among the protein ion cloud and the electrons formed by the cathode, which is enabled by carrying out ECD runs in the FTICR detection cell.\textsuperscript{127}
2 Complete mapping of disulfide connectivity in disulfide-rich proteins by high-resolution FT-ICR MS combined with pepsin and trypsin digestion

2.1 Biological Redox Signalling

2.1.1 The Chemistry of Cysteine

The role of redox processes in cellular physiology has always been a key issue of study. There is a rapidly growing literature in the biochemical field, which indicates that oxidation of organic molecules and shift of electrons to molecular oxygen can deploy vast amounts of free energy utilised by cells for the generation of energy. However, quite early, it was also determined that excess of by-products of aerobic oxidation, particularly reactive oxygen species (ROS), might induce deleterious effects on cells, which are offset by multiple physiological antioxidant defence systems.

Much of the current literature on ROS pays particular attention to describing the role of biochemical radicals in redox signalling within several physical processes regarding extracellular stimuli. The amount of ROS is subtly controlled by many enzymatic and non-enzymatic activities to maintain redox homoeostasis for conventional cell processes. Specifically, ROS serve as agents in biological signal transduction, and many prokaryotic and eukaryotic polypeptide chains are now perceived to have their functions regulated via ROS oxidation arising on critical sulfur-containing amino acids cysteine (Cys) and methionine.

Cys is the strongest nucleophile and the rarest amino acid in proteins, contributing only to 1.9% of all natural amino acid residues. Cys residues can undergo thiol-modifications due to the ability of sulfur atom to adopt a wide range of oxidation states (-2 to +6) and distinct chemical forms (chemotypes) (Error! Reference source not found.). Reactivity of the Cys thiol group towards ROS depends on the accessibility and the pKa value, shaped by microenvironmental conditions such as pH and neighbouring residues. Specifically, Cys thiols with neutral pKa values of 8.4 to 8.6 are protonated and not redox-sensitive under oxidative stress. In contrast, Cys thiols with low pKa are present in the deprotonated thiolate anion form that is redox-sensitive to undergo thiol oxidation. For instance, Cys pKa...
values of 3.5 and 10 were determined for two active site Cys residues of the DsbA disulfide oxidoreductase, but only the low pKa Cys was involved in thiol-oxidation of the substrate.\textsuperscript{139,140} Thus, the determination of the Cys pKa is crucial to reveal the redox-sensing Cys residues in proteins and their function in redox signalling processes. Redox-sensitive Cys residues with low pKa values can undergo different reversible and irreversible post-translational thiol-modifications under oxidative stress. Reversible thiol-disulfide switches are involved in redox signalling processes and protect the thiol group against overoxidation.\textsuperscript{139}

\textbf{Figure 2.3: Oxidative thiol modifications.} Sulfur occupies a unique position in biology because of its ability to adopt a broad range of oxidation states (-2 to +6) and distinct chemical forms (chemotypes). This oxidative diversity, along with the inherent reactivity of the thiol and thiolate modifications, drive to the amino acid cysteine being susceptible to a diverse range of redox reactions which can lead to an array of protein covalent modifications.

2.1.2 The importance of disulfide bonds

In the literature, the S-S bond refers to a post-translational oxidative modification that imposes a defined distance restraint via a covalent bond within two distant Cys residues (1.8 to 3. Å) in the polypeptide, and it is this essence that makes the disulfide bond interesting (Error! Reference source not found.).\textsuperscript{141,142} Disulfides serve as (essentially permanent) molecular “staples” that direct and preserve the three-dimensional formation of proteins and define the distance and angle restraints between the connected cysteine residues, sustaining the folded
state in respect to the unfolded form. In addition, they can also take an active role in enzyme activity, such as in redox-active polypeptides that are involved in thiol-disulfide interconversion (e.g., the thioredoxin superfamily). In those cases, the disulfide formed on oxidation of adjacent thiols is frequently found in high-energy strained formations (e.g., the so-called +/-RH Hook and –RH Staple configurations), which are believed to facilitate rapid reduction back to the di-thiol form. Such redox-active disulfides have been demonstrated to be essential for regular cellular function, with perturbations assumed to be implicated in pathological conditions that are defined by abnormal/altered redox states, including aging, cardiovascular disorder, asthma, rheumatoid arthritis, cystic fibrosis, and multiple neurodegenerative diseases (e.g., Alzheimer’s, Parkinson’s, Huntington’s, Creutzfeldt-Jakob disease).

Anfinsen was the first to identify the crucial role that disulfide bonds play in ribonuclease by conducting a set of experimental trials where he denatured, chemically reduced, oxidized, and renatured the protein. His research confirmed that when thiols are oxidized to disulfide bonds under denaturing conditions, the ribonuclease’s activity has been remarkably diminished compared to thiols oxidized under renatured conditions. The data yielded by this study provide convincing evidence that that native disulfide chains need to be present for a biomacromolecule to function properly as non-native disulfide bonds form an utterly altered tertiary structure.

Since then, redox biologists have investigated different targets, and it has become clear that the effect of disulfide linkages on the protein folding, complex formation, and protein function is can not be generalised. For example, several reports support the assertion that disulfide bonds do not always stabilize their resident protein. Specifically, some disulfides provide little support to stabilization, while others may destabilize their resident proteoform.

More recently, there has been a rapid growth of protein-based therapeutics, which are regularly enriched with disulfide linkages to increase stability and support their biophysical functions. Consequently, the complete characterisation of disulfide connectivity (disulfide

Figure 2.4: Cellular bond disulfide formation. Schematic representation of electron transfer oxidation and reduction of disulfide bonds.
mapping) has become crucial in the pharmaceutical industry to ensure safe drug products and chemical consistency.\textsuperscript{150–152} These findings highlight the urgent need for comprehensive disulfide mapping techniques, as stated by Trivedi and colleagues.\textsuperscript{150}

2.1.3 Conventional Methods for Disulfide Bond Assignment

As noted, knowledge of the disulfide pattern is a significant issue in the biochemical field. However, defining the disulfide connectivity for a given proteoform is a complicated problem. Thus, several methods for disulfide bond analysis in proteins have emerged, employing a variety of analytical tools to address this issue; these include structural techniques such as Nuclear magnetic resonance (NMR)\textsuperscript{153–155}, and X-ray crystallography\textsuperscript{156,157} and sequencing approaches such as Edman degradation\textsuperscript{158,159} and mass spectrometry.

The use of structural techniques, despite successes, suffers from certain weaknesses. NMR and crystallography approaches require relatively large amounts of highly pure protein sample. Furthermore, X-ray crystallography uses the proximity of neighbouring thiols to determine disulfide bond bridges rather than measuring actual disulfide chains. Another major drawback to exploiting this system is the additional time needed for handling the data. After raw experimental data are obtained, both NMR and X-ray crystallography require significant amounts of time and effort. Lastly, these approaches, especially NMR, are not amenable to large proteins.\textsuperscript{153–157}

Edman Degradation can be useful if the studied sample is a peptide, usually less than 5 kDa in molecular weight. The reaction efficiency of each degradation step is high which minimises sample losses. However, the main shortfall of this approach is the need of a significant quantities of chemical reagents and time throughout the experiment. Furthermore, it can only interpret one pure peptide at a time, and the unit needs to have a free N-terminus to initiate degradation reaction.\textsuperscript{158,159} In contrast, mass spectrometry is an alternative analytical technique that circumvents some of these shortcomings.\textsuperscript{2,160,161}

2.1.4 Current Methods for Assigning Disulfide Linkages in Proteins by MS/MS

Bottom-up MS is the most widely utilised method for mapping disulfide connectivity patterns in biomacromolecules. The standard bottom-up proteomic strategies in MS-based disulfide mapping includes three key steps (Figure 2.5). The protein target is enzymatically digested,
usually using the enzyme trypsin, under nonreducing conditions to generate peptide fragments. A subset of digest products containing the disulfide linkages exist as bridged peptides, and analysis by proteomic LC/MS/MS workflows allow identification of peptide sequences and assignment of specific cysteine pairs involved in disulfide bonds formation. An ideal sample preparation process would release many disulfide-linked peptides and minimize the formation of artefactual disulfide bonds. Afterwards, the disulfide-linked species are separated by an LC system and are dissociated with appropriate MS/MS fragmentation processes that are powerful enough to fragment the peptide backbones, yet soft enough to preserve the disulfide chains integrity. Then, the MS/MS spectra are run by automated software tools to distinguish disulfide bridged peptides and localize the disulfide bond sites. A helpful tool can assign disulfide bonds with high sensitivity and accuracy and enable simple validation to eliminate false positives. Therefore, several software platforms (Kojak, Maxquant, ProteinMetrics, etc) have been developed that are capable of allocating disulfide connectivity based on the MS/MS fragmentation patterns of proteolytic digests. 

**Figure 2.5: Bottom-up proteomics workflow for disulfide connectivity mapping.** In the Bottom-up method, enzymatic digestion is employed to cleave intact proteins into peptides. Peptides are then chemically reduced and run by MS, and peptides are identified by running them against a database. Afterwards, oxidised peptides are run by LC-MS/MS to generate extensive sequence information and locate the disulfide bonds.
The appealing features of the bottom-up method involve the availability of a wide range of proteases to proteolysis high molecular proteins into small fragments (peptides bearing intact disulfide chains) that are easier to analyse, compatible fragmentation processes, and the capability to couple MS with LC to separate the proteolytic digests before injecting to the mass spectrometer.\textsuperscript{166}

However, despite recent technological advances, routine disulfide bridges mapping using this approach is still challenging, especially for the analysis of disulfide-rich proteins (Figure 2.6).\textsuperscript{166–168} This is due to several complicating factors. Firstly, the presence of structural disulfide bonds in the target protein often dramatically reduces the enzymatic efficiency upon protease treatment, especially if this step is performed in solution.\textsuperscript{152} Thus, insufficient proteolytic digestion of oxidized cysteine-rich protein is common. Although this can be mitigated by chemical denaturation before digestion, this additional step in sample preparation is time-consuming and can be accompanied by unwanted chemical modification of protein sidechains, hampering downstream data interpretation. Secondly, depending on the protein primary amino acid sequence of the target protein, the production of the desired simple digest products that exclusively contain a single disulfide bond may not be feasible. This is often encountered when using a high specificity protease, such as trypsin, when an arginine or lysine residue is not found between two cysteine residues. Again this becomes commonplace in cysteine-rich proteins, containing nested disulfide bridges with closely spaced (or even adjacent) cysteine residues.\textsuperscript{169–172} Thirdly, proteolysis is routinely performed in the pH range of 7.5–8.5 which is often necessary for optimal digestion. However, at this slightly basic pH, disulfides are susceptible to rearrangement, termed disulfide scrambling.\textsuperscript{168} This undesirable phenomenon can be minimised by adjusting the pH to acidic conditions; yet, this will adversely affect digestion specificity and efficiency.\textsuperscript{168} Thus, with more than a single disulfide bond present in the species to be analyzed, other strategies need to be developed for the routine comprehensive assignment of the disulfide bond connectivity in cysteine-rich proteins.\textsuperscript{169}
Figure 2.6: The key challenges related to routine disulfide connectivity mapping are enzymatic digestion, closely spaced cysteine residues, and S-S scrambling.

To overcome this challenge, some research groups have employed pepsin to identify disulfide bonds. Pepsin is an acidic protease whose activity is greatly dependent on its pH. Pepsin has its highest proteolytic activity at a pH within the 1.8 and 2.0 range. It remains stable, yet highly active even when the pH falls to as low as 1.0. Moreover, previous studies have shown that pepsin will begin to lose activity around pH 5, and it converts to an irreversibly inactive form at a pH of around 7. Pepsin is increasingly set to become a vital tool in the proteomics field due to its broad specificity. Previous studies appear to suggest that it preferentially cleave after bulky hydrophobic amino acid residues. Because of its broad specificity, pepsin generates a lot of peptides during digestion. The various cleavage sites imply that the peptides generated are often low molecular weight, around 3 to 30 residues in length.

This strategy is attractive for multiple reasons: the ability to perform efficient digestion at acidic pH helps minimise disulfide bond reshuffling; while also ensuring that proteoforms
conformations are likely to be unfolded and accessible to cleavage, allowing more efficient proteolysis within closed-nested Cys residues of the substrate protein\textsuperscript{,165,170,175,176} Pepsin, which behaves optimally at low pH, is a proper choice for disulfide bond analysis, but its broad specificity can hinder the assignment. Finally, pepsin generates complex digests that contain overlapping peptides due to nonspecific cleavage, which thus leads to complicated spectra\textsuperscript{,165,170,175–177} Although this characteristic affords the opportunity to cleave between Cys residues which contain no conventional proteolytic site, it makes the assignment of the peptide fragments challenging.

Another strategy that has been considered is the application of a variety of fragmentation techniques for the MS/MS analysis of peptides. Collision induced dissociation (CID) is the most prevailing dissociation technique for the study of S-S peptides\textsuperscript{,178} CID usually triggers the preferential cleavage of the peptide backbone (amide) bonds, while leaving the S-S chain intact, thus generating $b$ and $y$ ions that carry the disulfide link, as well as ions that do not carry it. Although the assignment of produced CID ions of disulfide-bonded peptides was based before on the presumption that only a single peptide bond is fragmented during CID, Clark et al. recently confirmed that double cleavage of peptide bonds is typical during CID of disulfide-joined peptides\textsuperscript{,179} One of the major drawbacks to adopting this system is that CID dissociation is not ideal for distinguishing disulfide-linked peptides since it does not cleave the disulfide bridges and generates difficult-to-interpret product mass spectra of the disulfide-joined peptides. In addition, in cases where closed nested disulfide linkages are involved, information concerning each disulfide bond in the protein formation is lost\textsuperscript{,180–182} Efforts have been directed to tackle these obstacles by employing novel fragmentation techniques to determine protein identity rapidly and provide comprehensive molecular connectivity information.
Figure 2.7: Fragmentation patterns for persulfide and dehydroalanine fragment ions formed by C-S bond cleavage reactions, and for cysteine and cysteine thioaldehyde fragment ions formed by S-S bond cleavage reactions.

Electron-capture dissociation (ECD), electron transfer dissociation (ETD) are alternative means of fragmentation that can be employed to provide disulfide connectivity information. The key advantage of these dissociation techniques is that they are more efficient than CID in fragmenting S–S bonds, hence enabling studying disulfide-protected regions. However, data analysis can be tedious because ECD and ETD typically produce backbone cleavages and cleavages resulting from the cleavage of a backbone bond in conjunction with either C–S or S–S bond cleavage, leading to complex tandem mass spectra. Figure 2.7 shows the four potential fragmentation products of disulfide linkages. The generating ions produced by the latter type of cleavage involve mass shifts matching to cysteine thioaldehyde (−2 Da), cysteine persulfide (+32 Da), or dehydroalanine (−33 Da). Generated fragment
ions carrying more than one cysteine may have a mixture of these variants, thus further complicating data interpretation.

Jiang et al.\textsuperscript{189} described a process that combines altering ETD/CID dissociation of the disulfide-chained peptides and CID-MS\textsuperscript{3} fragmentation of the cysteine-bearing fragment ions that emerged from the ETD-MS\textsuperscript{2} cleavage of disulfide linkages. This approach is appropriate to map complex disulfide patterns, such as nested disulfides and cysteine clusters,\textsuperscript{190} along with peptides fully cyclized by a disulfide chain.\textsuperscript{191} However, due to the tremendous volume of data produced due to the extra CID step, the process may not require samples that do not contain complicated disulfide patterns. In his recent study, Massonnet demonstrated a similar method for mapping disulfide chains of peptides carrying two disulfide bonds. After the cleavage of the disulfide links by ETD, the produced species are separated by ion mobility, and then the separated ions are identified by CID fragmentation.\textsuperscript{192} Clark et al.\textsuperscript{178} demonstrated a simpler bottom-up ETD-based approach that is suitable for samples carrying simple interchain disulfide-bonded peptides, involving disulfides with several disulfide-chained strings. The system does not require an MS\textsuperscript{3} step, and rather than searching the m/z values of all identified disulfides, and potential shuffled disulfide chains, the extracted ion chromatograms of all Cys carrying species are produced, and the ETD spectra of the peaks in the extracted ion chromatograms are studied to assign the disulfide patterns.\textsuperscript{178,193} The determination of disulfide patterns is made when the fragments of cysteine-carrying peptides are assigned in the same ETD spectrum, and the sum of their masses minus 2 Da (mass of 2H) corresponds to the molecular weight of the expected disulfide. The c and z ions emerging from the amide split of the attached peptides can be utilised to further affirm the identities of the disulfide chains.\textsuperscript{193} This method is suitable to map the disulfide links in proteoforms whose disulfide connectivity is unknown and confirm non-native or alternative disulfide bonds promptly. The extracted chromatograms of cysteine-bearing peptides can also provide information about any free thiol in the polypeptide without the need for alkylation.\textsuperscript{194,195}

Besides CID and ETD, higher-energy collisional dissociation (HCD)\textsuperscript{196} and a dual fragmentation method known as electron-transfer and electron-transfer/higher-energy collision dissociation (EThcD)\textsuperscript{197,198} are recently described fragmentation methods available in specific Orbitrap instruments, and they are attaining more and more attention in disulfide bond analysis.\textsuperscript{168,172,199} HCD exhibits only b and y ions in a manner similar to CID\textsuperscript{200}, while EThcD spectra generate ions visible in both ETD and CID spectra, and its initial studies suggest its performance may exceed ETD in comparative experiments.\textsuperscript{168}
Despite the exceptional versatility and broad adoption of these methods, there are extensive efforts to develop alternative ion activation systems, such as ones that favour disulfide bond cleavages. Examples of these include the rise of photodissociation processes (UVPD; in 157 nm\textsuperscript{201}, 193 nm\textsuperscript{160}, and 266 nm\textsuperscript{202–204}. The potential of employing UVPD at 157 nm to cleave both inter- and intrachain disulfide links without charge reduction was initially demonstrated by Zubarev and coworkers.\textsuperscript{201} Agarwal et al. used 266 nm UVPD to selectively favour fragmentation at S-S bonds while leaving fifteen other bonds intact, allowing for the determination of disulfide chains patterns. Figure 2.8 demonstrates this homolytic S-S bond cleavage at 266 nm. Research has provided ample support for the assertion that adjacent chromophores such as tyrosine and tryptophan can increase photodissociation yield of the disulfide linkage. A characteristic spectrum comprising three peaks is regularly observed: the parent ion and two flanking peaks correspondent to the individual peptides.\textsuperscript{202–204}

![Disulfide linkage](image)

*Figure 2.8: Selective homolytic cleavage of the disulfide linkage with 266 nm wavelength.*

### 2.2 Research Outline

In this study, we described an efficient workflow (Figure 2.9) for both intra- and inter- disulfide bridge assignment, including a combination of pepsin and trypsin proteolysis and the usage
of dual fragmentation of ECD and CID without the need of a chromatographic system. As others have highlighted, challenges associated with disulfide assignment by LC-MS of nonreduced enzymatic digests are to resolve complex disulfide connectivity and ensure confident data analysis.\textsuperscript{167}

**Figure 2.9: Developed proteomics workflow for disulfide connectivity mapping.** In this method, enzymatic digestion is employed to cleave intact proteins into peptides. Peptides are chemically reduced and run by FT-ICR-MS, and peptides are identified by running them against a database. Afterwards, oxidised peptides are run by FT-ICR-MS and fragmented by CID and ECD to generate extensive sequence information and locate the disulfide bonds.

As mentioned above, pepsin has the benefits of generating non-specific cleavages under acidic conditions. However, the exceptionally high search space makes assignment based on the mass of peptides alone quite challenging. On the other hand, trypsin protease only cleaves after K and R; therefore, search space is significantly smaller. However, employing trypsin alone is not adequate either since it does not provide comprehensive coverage of S-S bonds. To address the issue thus, we have decided to combine both proteases. One of the simplest ways of tackling this problem is to exclude the chromatographic system and achieve efficient
separation of the fragments based only on the resolving power provided by the analytical mass spectrometer. The main analytical mass spectrometer with such high resolving power is FT-ICR-MS. The significantly high resolution of the FT-ICR-MS allows us to determine the isotope fine structure of peptide sequences in pepsin digest and be used as an additional confirmation of their identity. Thus, the aim of this work is to develop a chromatography-free protocol for easy and efficient mapping of disulfide connectivity of disulfide-rich proteins on high field FT-ICR platforms.

As a test protein, bovine serum albumin (BSA, P02769) was enzymatically digested, by trypsin or pepsin, under nonreducing conditions to generate peptides. Bovine serum albumin (BSA) is a model protein in multiple pieces of research. It is structurally well-characterized and readily available in its native conformational state. The primary structure of BSA was first published in 1971 and later updated in 1990. BSA contains 583 amino acids, of which one free Cys thiol and seventeen disulfide bonds. The generated peptides were then chemically reduced and analysed by high-resolution FT-ICR-MS by direct infusion, and the reduced peptides were assigned from database searches. Afterwards, the same process was repeated without reducing the disulfide bonds prior to analysis. Ions unique to the non-reduced peptide mixture, indicating the presence of disulfide linkages, were subjected to CID and ECD to generate extensive sequence information and locate the disulfide bonds. This framework overcomes issues induced by disulfide reshuffling through sample preparation and provides a confident assignment of both intra- and inter-disulfide bridges from MS/MS analyses. We evaluated the integrated workflow using BSA (Figure 2.10), a disulfide-rich standard protein, as a model protein and demonstrated that this approach could certainly and robustly locate disulfide bridges on proteins.
Figure 2.10: Bovine Serum Albumin primary (a) and secondary (b) structures, the latter containing its native SS cross-links. Based on Majorek et al.44 (Protein Data Bank code 3V03) and Huang et al. (Uniprot Code:PO2769).

2.3 Materials and Methods

2.3.1 Reagents

Bovine Serum Albumin (BSA) was purchased from Sigma Aldrich (Cas Number: 9048-46-8, St. Louis, MO). Ammonium bicarbonate and TCEP-HCl (purity ≥98 %) were purchased from Thermo, USA. Trypsin reagent and pepsin were purchased from Thermo, Germany. Acetonitrile (ACN) (≥99.9 %) was purchased by VWR Chemicals (USA). Water (LC/MS Grade) used for all experiments was obtained from Thermo Fisher Scientific, UK. Formic Acid (FA) (≥96 %) was obtained by Sigma Aldrich (Germany).

2.3.2 Mass Spectrometry

Mass spectrometry was carried out using a 12 T SolariX FT-ICR instrument mass spectrometer (Bruker Daltonics, Billerica, MA, USA) with ESI ionization. Proteins (0.5–10 μM) were prepared in denaturing solution conditions with ACN: H₂O: FA ratio of 49.95:49.95:0.1. Nitrogen was employed as the collision gas. MS experiments were performed in the broadband mode from m/z 300–2000. Precursor ions of single charge states...
were isolated by a quadrupole Q1 with a selection window size of 1–5 m/z. MS/MS experiments were performed with the following settings: ECD; pulse length and ion accumulation were modified as necessary; bias 1.5 V; ECD lens 15 V; CAD, collision energy 0–3 V. 50–100 scans were averaged for each MS experiment. 100 scans were averaged for each ECD experiment.

2.3.3 Enzymatic Digestion

Pepsin was dissolved in a formic acid aqueous solution (1 %) at an initial concentration of 1.64 mg/mL. BSA with an initial concentration of 1.64 mg/mL in water was digested with pepsin in a ratio of 20:50 (weight of pepsin: weight of BSA) at 37°C overnight. After digestion, excess salts were removed prior to mass spectrometry analysis with ZipTip Bond Elut OMIX C18 Ziptip (Agilent Technologies).

Trypsin was dissolved in a 50 mM ammonium bicarbonate aqueous solution at an initial 1.64 mg/mL concentration. BSA with an initial concentration of 1.64 mg/mL in water was digested with trypsin in a ratio of 20:50 (weight of trypsin: weight of BSA) at 37°C overnight. After digestion, excess salts were removed prior to mass spectrometry analysis with Bond Elut OMIX C18 Ziptip (Agilent Technologies).

2.3.4 Chemical Reduction of the Peptide Mixture

The peptides mixture was controlled at pH = 5 by adding HCl or NaOH. Afterwards, TCEP.HCl (10 mM) was added to the mixture and left incubating at 37 °C for 2 hours to ensure that all the peptides were reduced. After reduction, excess salts were removed prior to mass spectrometry analysis with ZipTip Bond Elut OMIX C18 Ziptip (Agilent Technologies).

2.3.5 Data Analysis

FT-ICR spectra were analysed using DataAnalysis software (Bruker). Peak lists were generated by the Sophisticated Numerical Annotation Procedure (SNAP) used as part of the Bruker DataAnalysis software. This algorithm relies on matching observed isotope patterns to that of the theoretical amino acid "averagine", which has the molecular formula $C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$, calculated to be the average composition of an amino acid in the human genome. To retrieve every generated ion SNAP identified, we set the quality factor...
threshold at 0.3, the signal-to-noise ratio (S/N) threshold at 4, the relative intensity threshold (base peak) at 0.01 %, the absolute intensity threshold at 0 and the maximum charge state to 20.

Lists of proteolytic peptide monoisotopic masses were used to match potential peptides employing the MS-Bridge software, part of the “Protein Prospector” program. The link search type was chosen disulfide, and maximum link molecules was set to 4 and 0 for oxidised and reduced protein, respectively. For both reduced and oxidised, the error tolerance was set to 2 ppm error. All CID/ECD spectra were calibrated internally by quadratic correction to multiple confirmed fragments manually. The quadratic correction was selected since the generated disulfide bridged peptides had a broad molecular weight range and many Cys that affect their isotope pattern.

Fragmentation spectra were analysed using Prosight Light\textsuperscript{205} (Northwestern University). Matches to a fragment ±5 Da were accepted, and all signals matched to a fragment from each peptide spectra were manually checked for each spectrum. Similarly, if an expected fragment was not detected, the spectrum was manually checked to ensure all fragments were detected.

All isotope distributions were modelled using the IsotopePattern program (Bruker).
Results and discussion

2.4.1 Sequence mapping of BSA by combining trypsin and pepsin digestion

2.4.1.1 Setting the Optimum Conditions

In the initial stage, the protein was successfully identified using a bottom-up approach with two different proteases: trypsin and pepsin. The generated peptide fragments were chemically reduced and analysed by direct infusion FT-ICR MS. The signals provided by both mass spectra were analysed, and a database search was performed using the online proteomics tool Protein Prospector (v6.2.1), developed by the Mass Spectrometry Faculty in the University of California, San Francisco. The data yielded by the trypsin and pepsin digestion showed the generation of 403 and 739 peptides, respectively (Table Appendix 1 and 2). To determine the optimum settings, the monoisotopic masses generated by trypsin digestion were searched against the database using an error tolerance of 2, 5, 10, 20, and 50 ppm error (Figure 2.11). Bioinformatically, trypsin has more constraints, so it should be easier to assign confidently since trypsin cleaves only on lysine and arginine residues. The maximum number of missed cleavages was set to 4.

By looking at Figure 2.11, there appears to be a clear upward pattern in the total number of hits as the set error rises. However, given that our findings are based on a mass match, the results from such analyses should, therefore, be treated with the utmost caution because they might be due to false positives.

To examine this theory, the same peptide list was searched against a shuffled BSA sequence generated from the https://onlinetoolz.net/shuffle-letters website. As illustrated in Figure 2.12, we found that the total number of hits dropped to significantly lower values than the previous. Hence, these results provide confirmatory evidence that the sequence coverage can be achieved confidently only simply based on the software’s prediction.
Figure 2.11: Total number of hits of reduced BSA species (right sequence) using an error tolerance of 2, 5, 10, 20 and 50 ppm error. From the graph, we can conclude that there is a strong correlation between the set error and the total number of hits.

Figure 2.12: Total number of hits of reduced BSA species (shuffled sequence) using an error tolerance of 2, 5, 10, 20, and 50 ppm error. We can see a strong correlation between the set error and the total number of hits from the graph.
Afterwards, the process was repeated. However, the generated fragments were introduced to the FT ICR MS without prior reduction during this time. The produced fragments were then searched against a database using an error tolerance of 2, 5, and 100 ppm error (Figure 2.13). As before, all the variable modifications were removed since there are no known other redox modifications. The software predicted disulfide bridged peptides whose mass matched with the theoretical disulfide pattern. Since the disulfide connectivity is well known, the suggested peptides were tested manually. As anticipated, our experiments demonstrated that the number of false hits rose with the increase of error.

Furthermore, as shown in Figure 2.13, the increased error led to a higher number of false positives. The exact process was repeated with the shuffled sequence. From these results, the best ratio of correct predictions against the total number of hits was given by setting an error tolerance of 2 ppm error.

![Graph showing total number of hits of oxidised BSA species (correct sequence) using an error tolerance of 2, 5, and 10 ppm error. The correct and false positives were illustrated with blue and orange, respectively. The graph above shows that as the error of tolerance is set higher, false hits rise significantly. We thus chose 2 ppm error as the optimum since the ratio of correct positives to total hits was the highest.](image)

**Figure 2.13: Total number of hits of oxidised BSA species (correct sequence) using an error tolerance of 2, 5, and 10 ppm error.** The correct and false positives were illustrated with blue and orange, respectively. The graph above shows that as the error of tolerance is set higher, false hits rise significantly. We thus chose 2 ppm error as the optimum since the ratio of correct positives to total hits was the highest.
Then, the same process was repeated, employing the pepsin data. As before, to determine the optimum settings, the monoisotopic masses generated by pepsin digestion were searched against the database using an error tolerance of 2, 5, 10, 20, and 50 ppm error (Figure 2.14). However, the analysis is more bioinformatically challenging since our findings are based on a mass match; the results from such analyses should, therefore, be treated with the utmost caution because they might be due to false positives. To determine this, the same peptide list was searched against a shuffled BSA sequence. As illustrated in Figure 2.15, we found that the total number of hits leapt to much higher values than the previous. Hence, these results provide confirmatory evidence that the sequence coverage cannot be achieved confidently only based on the software’s prediction. Further experiments need to be run to verify the predictive outcome.

![Figure 2.14: Total number of hits of reduced BSA species (correct sequence) using an error tolerance of 2, 5, 10, 20, and 50 ppm error. From the graph, we can note a strong correlation between the set error and the total number of hits. Specifically, setting the error of tolerance to higher values leads to a higher number of total hits.](image-url)
Figure 2.15: Total number of hits of reduced BSA species (shuffled sequence) using an error tolerance of 2, 5, 10, 20, and 50 ppm error. From the graph, we can see a strong correlation between the set error and the total number of hits. Specifically, setting the error of tolerance to higher values leads to a higher number of total hits.

Afterwards, the process was repeated. However, the generated fragments were introduced to the FT ICR MS without prior reduction during this time. The produced fragments were then searched against a database using an error tolerance of 2, 5, and 100 ppm error (Figure 2.16). As before, all the variable modifications were removed since there are no known other redox modifications. As before, all the variable modifications were not taken into account since there are no other redox modifications. The software predicted disulfide bridged peptides that their mass matched with the ones from the given list. Since the disulfide connectivity is well known, the suggested peptides were tested manually. As anticipated, our experiments demonstrated that the number of false hits rose with the increase of error.

Moreover, as shown in Figure 2.16, the increased error led to more false positives. The exact process was repeated with the shuffled sequence. From these results, the best ratio of correct predictions against the total number of hits was given by setting an error tolerance of 2 ppm error.
Figure 2.16: Total number of hits of reduced BSA species (shuffled sequence) using an error tolerance of 2 and 5 ppm error. The correct and false positives were illustrated with blue and orange, respectively. The graph above shows that as the error of tolerance is set higher, false hits rise significantly. We thus chose 2 ppm error as the optimum since the ratio of correct positives to total hits was the highest.

2.4.2 De novo sequencing of disulfide bridged peptides

In this section, the question under discussion is whether a high-resolution FT ICR approach combined with trypsin and pepsin digestion can study complex disulfide bridged peptides without employing an LC system. BSA was chosen as a model protein in this study, because its disulfide connectivity has been extensively described, making this polypeptide a suitable choice for validating the developed method. Our approach to enhancing the production of the disulfide-bound peptides is to digest the protein without reduction/alkylation. Trypsin and pepsin digestion of BSA prior to reduction and alkylation cleaves BSA into many peptide products containing the 17 disulfide bonds of BSA and other conventional tryptic and peptic peptides that do not contain any disulfide bonds. Various single, double and triple peptides have been isolated and fragmented by CID/ECD. Here, selected peptides from each category have been selected and presented (Figure 2.17).
2.4.2.1 A single peptide chain containing one intrachain disulfide bond

A peptide generated after pepsin digestion displayed a mass of 1745.771 3 Da. This mass was selected because when the sample was reduced, the peak shifted, indicating that it had a disulfide bond. This peak corresponds to the peptide AKTC53VADESHAGC62EKSL (C70H115N21O27S2) with a single disulfide bond (Figure 2.18, Table Appendix 1). A specific charge state of the oxidised peptide ([M+3H]3+) was isolated and studied alongside a simulation assuming one disulfide bond. As displayed in Figure 2.18, the obtained isotope distribution matched well with the simulation, and a monoisotopic mass error of 0.5 ppm was achieved in addition to the assignment based on monoisotopic mass error and isotopic distribution, isotope fine structure analysis achieved further confidence in the assignment.

To add an even higher level of confidence, the third peak of the isotopic distribution was zoomed in and compared to the theoretical isotopic composition. The zooming demonstrates that apart from a slight discordance, the result matches the prediction. Thus, the present findings confirm that FT ICR MS resolution (Rs = 49391) is so high that the isotopic composition is visible, and it can be used as an extra level of confidence. These primary findings confirm that this peptide is indeed AKTC53VADESHAGC62EKSL with one disulfide bond.
Figure 2.18: Comparison of the ion \([M+3H]^{3+}\) with mass of 1745.7713 with the simulated isotope distribution of \(\text{AKTC}_{53}\text{VADESHAGC}_{62}\text{EKSL}\) with one disulfide bond. The simulated isotope distribution is plotted over the obtained spectrum as a scatterplot. To add an even higher level of confidence, the third peak of the isotopic distribution was zoomed in and compared to the mathematical simulation. The zooming shows that FT ICR MS resolution is so high that the isotopic composition is similar to the theoretical one. This comparison confirms that the oxidised peptide is \(\text{AKTC}_{53}\text{VADESHAGC}_{62}\text{EKSL}\) with one disulfide bond.

The following step was to sequence the peptide by CID and ECD. Initially, it was presumed that the CID would not be as efficient at reducing the disulfide bonds as ECD. To demonstrate this, CID and ECD experiments were run on the oxidised peptide, and their spectra (CID; (Figure 2.19; Table Appendix 3), ECD; (Figure 2.21; Table Appendix 4)) fragmentation maps (CID; Figure 2.39, ECD; Figure 2.45) are displayed.
From Figure 2.19, we can see that the most abundant peaks in the CID spectra are products of electron capture with no dissociation of the precursor ion and several \( b \) and \( y \) fragments close to the edges. As anticipated, our experiments show that the CID fragments from the edges of the peptide are significantly more intense than those from the central region. The single most conspicuous observation to emerge from the data was that by applying 12 V, low-intensity fragments with reduced disulfide bond emerged. This confirms previous findings in the literature saying that CID might lead to detectable sequence fragment ions of the protein backbone inside a disulfide bridge. The application of 12 V resulted in a sequence coverage of 100%, and all the identified ions were assigned within a 1.5 ppm error (Figure 2.39). These results offer compelling evidence for confirming our assignment.
Figure 2.39: The fragment map displays fragment b and y ions resulting from CID fragmentation of AKTC$_{53}$VADESHAGC$_{62}$EKSL. Several peaks have been generated, allowing 100% sequence coverage.

From the Figure 2.21, we can see that the most abundant peaks in the ECD spectra are products of electron capture with no dissociation of the precursor ion and several c and z fragments.

Figure 2.21: ECD spectra for the [M+3H]$^{3+}$ charge state of AKTC$_{53}$VADESHAGC$_{62}$EKSL with pulse length set to 100 ms. The fragments with black colour are the ones with the disulfide bond intact, while the red ones result from the reduction of disulfide bonds by ECD.

A closer look at the data indicates that ECD is more efficient in cleaving disulfide bonds and generating high-intensity fragments from across the sequence. Specifically, a series of c-ions (c4-c11) and z-ions (z7-z13) are observed. These fragments are internal to the disulfide bond between Cys53 and Cys63. Studying the observed c-type ions in the ECD spectra of oxidised AKTC$_{53}$VADESHAGC$_{62}$EKSL, it is clear that several c-ions derived from oxidised AKTC$_{53}$VADESHAGC$_{62}$EKSL are products of the cleavage of a backbone bond in conjunction
with either C–S or S–S bond cleavage (Figure 2.41, Figure 2.42, Figure 2.43, Figure 2.44). To add an even higher level of confidence, the third peak of each isotopic distribution was zoomed in and compared to the theoretical isotopic composition. The zooming demonstrated that FT ICR MS resolution is that high that it can distinguish between the different isotopes. This suggests that ECD-induced disulfide bond reduction results in a Cys–S• sidechain on Cys53 (i.e. the neutral radical species).

Figure 2.41: Analysis of c₆ fragment ion of oxidised AKT₅₃VADESHAGC₆₂EKSL. The c₆ fragment ion was compared with a mathematical simulation assuming a c₆ fragment from the given sequence with a loss of a proton. To add an even higher level of confidence, the third peak of the isotopic distribution was zoomed in and compared to the mathematical simulation. The zooming of the third peak shows that the data matches the theoretical isotopic distribution even at this high level of resolution. All these findings confirm that the c₆ ion in the oxidised AKT₅₃VADESHAGC₆₂EKSL ECD spectra has lost a hydrogen suggesting a cleavage of the SS– bond between Cys53 and Cys62.
Figure 2.42: Analysis of $c_7$ fragment ion of oxidised AKTC$_{53}$VADESHAGC$_{62}$EKSL. The $c_7$ fragment ion was compared with a mathematical simulation assuming a $c_6$ fragment from the given sequence with a proton loss. This comparison hints that it matches. The zooming of the third peak shows that the data matches the theoretical isotopic distribution even at this high level of resolution. All these findings confirm that the $c_7$ ion in the oxidised AKTC$_{53}$VADESHAGC$_{62}$EKSL ECD spectra has lost a hydrogen suggesting a cleavage of the SS – bond between Cys53 and Cys62.
Figure 2.43: Analysis of $c_{14}$ fragment ion of oxidised AKTC$_{53}$VADESHAGC$_{62}$EKSL. The $c_{14}$ fragment ion was compared with a mathematical simulation assuming a $c_{14}$ fragment from the given sequence with the disulfide bond intact, and this comparison implies that it matches. The zooming of the third peak shows that the data matches the theoretical isotopic distribution even at this high level of resolution.
Figure 2.44: Analysis of $c_{14}$ fragment ion of oxidised $\text{AKT}_{53}\text{VADESHAGC}_{62}\text{EKSL}$. The $c_{12}$ fragment ion was compared with a mathematical simulation, and this comparison suggests that it matches. The $c_{12}$ ion in the oxidised $\text{AKT}_{53}\text{VADESHAGC}_{62}\text{EKSL}$ ECD spectra has gained a sulphur suggesting a cleavage of the SS – bond between Cys$_{53}$ and Cys$_{62}$.

The 100 ms pulse length application in combination with 3 V collisional voltage led to 87 % sequence coverage. Furthermore, all the identified ions were assigned within a 3 ppm error (Figure 2.45). To sum up, the data yielded by this study provide convincing evidence that the isolated peptide is $\text{AKT}_{53}\text{VADESHAGC}_{62}\text{EKSL}$, and there is an intramolecular disulfide bond between Cys$_{53}$ and Cys$_{62}$. 
Figure 2.45: The fragment map displays fragment ions resulting from ECD fragmentation of AKT$_{52}$VADESHAGC$_{62}$EKSL. Several peaks have been generated, allowing 82% sequence coverage.

2.4.2.2 Single Chain connected with two disulfide bonds

A peptide fragment peak generated after pepsin digestion with molecular weight of 3584.7837 Da. This peak corresponds to NRLC$_{460}$VLHEKTPVSEKVTKC$_{474}$C$_{475}$TESLVNRRPC$_{486}$F (C$_{152}$H$_{253}$N$_{47}$O$_{45}$S$_{4}$) peptide with two disulfide bonds (Figure 2.46, Table Appendix 1). In detail, a specific charge state of the oxidised peptide ([M+5H]$^{5+}$) was isolated and studied alongside a simulation assuming two disulfide bonds. As displayed in Figure 2.46, the obtained mass spectra fit pretty well with the mathematical simulation assuming two disulfide bonds. These primary findings confirm that the peptide is NRLC$_{460}$VLHEKTPVSEKVTKC$_{474}$C$_{475}$TESLVNRRPC$_{486}$F indeed, with two disulfide bonds intact.
Figure 2.46: Comparison of the ion \([M+5H]^{5+}\) with a mass of 3584.7837 with the mathematical simulation assuming 
\(\text{NRLC}_{460}\text{VLHEKTPVEKTKC}_{474}\text{C}_{475}\text{TESLVNRRPC}_{486}\text{F}\) with two disulfide bonds. This comparison suggests that the oxidised peptide is 
\(\text{NRLC}_{460}\text{VLHEKTPVEKTKC}_{474}\text{C}_{475}\text{TESLVNRRPC}_{486}\text{F}\) with two disulfide bonds.

The following step was to sequence the disulfide bridged peptide by ECD. ECD experiment was run on the oxidised peptide, where the spectrum (Figure 2.47, Table Appendix 5) and the fragmentation map (Figure 2.48) are displayed.

In Figure 2.47, we can see that the most abundant peaks in the ECD spectra are several \(c\) and \(z\) fragments from across the region of the disulfide bridged peptide. These fragments are internal to the region between \(\text{Cys}_{460}\) and \(\text{Cys}_{484}\). Studying the observed \(c\)-type ions in the ECD spectra of oxidised \(\text{NRLC}_{460}\text{VLHEKTPVEKTKC}_{474}\text{C}_{475}\text{TESLVNRRPC}_{486}\text{F}\), several \(c\)- and \(z\)-ions derived from oxidised \(\text{NRLC}_{460}\text{VLHEKTPVEKTKC}_{474}\text{C}_{475}\text{TESLVNRRPC}_{486}\text{F}\) are products of the cleavage of a backbone bond in conjunction with either C–S or S–S bond.
cleavage. These fragments resulted in 77 % sequence coverage. Furthermore, it is worth mentioning that all the identified ions were assigned within a 5.5 ppm error (Figure 2.48).

Figure 2.47: ECD spectra for the [M+ 6H]$^{6+}$ charge state of NRLC$_{460}$VLHEKTPVEKVTKC$_{474}$C$_{475}$TESLVNRRPC$_{486}$F with pulse length set to 100 ms.

These findings have further strengthened our confidence in the hypothesis that disulfide bridged peptide is NRLC$_{460}$VLHEKTPVEKVTKC$_{474}$C$_{475}$TESLVNRRPC$_{486}$F indeed and that there are two intramolecular disulfide bonds. Furthermore, it was confirmed that Cys460 binds either to Cys474 or Cys475 since several c- fragments (c$_{20}$-c$_{26}$) have a disulfide bond intact. In addition, there are several z- fragments (z$_{13}$-z$_{26}$) with a disulfide bond intact. Hence, it can be conceivably hypothesized that there is an intramolecular disulfide bond between Cys486 with either Cys474 or Cys475. However, one negative factor regarding our methodology was that it could not further assign the disulfide pattern. Despite these limitations of this method, our findings suggest the region where the fourteenth and fifteenth disulfide bond is placed.

Figure 2.48: The fragment map displays fragment ions resulting from ECD fragmentation of NRLC$_{460}$VLHEKTPVEKVTKC$_{474}$C$_{475}$TESLVNRRPC$_{486}$F. Several peaks have been generated, allowing 77 % sequence coverage.
2.4.2.3 Two Chains connected with two disulfide bonds

A peptide fragment peak generated after trypsin digestion with molecular weight of 2586.1087. This peak corresponds to VHKEC\textsubscript{244}C\textsubscript{245}HGDLLEC\textsubscript{251}ADDR – C\textsubscript{199}ASIQK disulfide bridged peptide with two disulfide bonds (Figure 2.49, Table Appendix 2). In detail, a specific charge state of the oxidised peptide ([M+4H]\textsuperscript{4+}) was isolated and studied alongside a simulation assuming two disulfide bonds. As displayed in Figure 2.49, the obtained mass spectra fit fairly well with the mathematical simulation assuming two disulfide bonds. This primary finding confirms that this peptide is VHKEC\textsubscript{244}C\textsubscript{245}HGDLLEC\textsubscript{251}ADDR – C\textsubscript{199}ASIQK indeed with two disulfide bonds.

![Figure 2.49: Comparison of the ion with a ion \([M+4H]^{4+}\) with of 2588.1240 with the mathematical simulation assuming VHKEC\textsubscript{244}C\textsubscript{245}HGDLLEC\textsubscript{251}ADDR – C\textsubscript{199}ASIQK with two disulfide bonds. This comparison suggests that the oxidised peptide is VHKEC\textsubscript{244}C\textsubscript{245}HGDLLEC\textsubscript{251}ADDR – C\textsubscript{199}ASIQK with two disulfide bonds.](image-url)
The following step was to sequence the disulfide bridged peptide by CID and ECD. CID and ECD experiments were run on the oxidised peptide, and their spectra (CID; Figure 2.50, Table Appendix 6, ECD; Figure 2.52, Table Appendix 7) fragmentation maps spectra (CID; Figure 2.51, ECD; Figure 2.55) are displayed.

In Figure 2.50, we can see that the most abundant peaks in the CID spectra are several $b$ and $y$ fragments from the peptide edges. As foreseen, our experiments show that the CID led to generating fragments from the edges of the bridged peptide while there are no internal fragments within Cys244 and Cys251.

CID of the oxidised bridged peptide produced a series of $b$-ions ($b_{14}$-$b_{18}$, A chain and $b_{23}$-$b_{15}$, B chain). These fragments are external to the disulfide bonds, which suggests that the central region is not susceptible to combined reduction and backbone cleavage under these applied CID conditions due to a potential disulfide bond. The most striking result to emerge from the data is the $y_7$ fragment from the A chain. This fragment is a result of an amide cleavage in conjugation with an S-S cleavage. The application of 12 V resulted in a sequence coverage of 56 %, and all the identified ions were assigned within a 3 ppm error (Figure 2.51). These results offer compelling evidence for confirming our assignment.

![CID spectra](image)

**Figure 2.50:** CID spectra for the $[M+4H]^4+$ charge state of VHKEC$_{244}$C$_{245}$HGDLLEC$_{251}$ADDR – C$_{199}$ASIQK with the collisional voltage set to 12V.
Figure 2.51: The fragment map displays fragment b and y ions resulting from CID fragmentation of VHKEC\textsubscript{244}C\textsubscript{245}HGDLLEC\textsubscript{251}ADDR – C\textsubscript{199}ASIQK. Several peaks have been produced, allowing 56 % sequence coverage.

It is apparent from Figure 2.52 that the application of 120 ms pulse length in combination with 3 V collisional voltage followed the generation of several c and z fragments from across the region of the disulfide bridged peptide. The data gathered in the pilot study implied that the ECD led to the production of two types of fragments; the ones that come with the intermolecular disulfide bond intact (Figure 2.53) and the ones with the intramolecular disulfide bond cleaved (Figure 2.54). Combing the two types of fragments lead to 87 % sequence coverage. Furthermore, it is worth to be mentioned that all the identified ions were assigned within a 3 ppm error (Figure 2.55).

These findings have further strengthened our confidence in the hypothesis that disulfide bridged peptide is QEC\textsubscript{167}C\textsubscript{168}QAEDKGAC\textsubscript{176}L – C\textsubscript{123}DEF indeed and that there are two disulfide bonds; one intramolecular and one intermolecular. The study of Figure 2.53 confirmed that Cys123 does not bound to Cys176 since there is an A chain b\textsubscript{5}, B chain c\textsubscript{3} fragment with the intramolecular disulfide bond intact. Hence, it can be conceivably hypothesized that there is an intramolecular disulfide bond between Cys199 with either Cys244 or Cys245. However, one negative factor regarding our methodology was that it could not further assign the disulfide pattern. Despite these limitations of this method, our findings imply the region where the fourth and fifth disulfide bonds are placed.
**Figure 2.52:** ECD spectra for the \([M+4H]^{4+}\) charge state of VHKEC\(_{244}C_{245}HGDLLEC_{251}ADDR – C_{199}ASIQK\) with pulse length set to 120 ms.

**Figure 2.53:** The fragment map displays fragment ions resulting from ECD fragmentation of VHKEC\(_{244}C_{245}HGDLLEC_{251}ADDR – C_{199}ASIQK\) assuming the intramolecular disulfide bond intact.
Figure 2.54: The fragment map displays fragment ions resulting from ECD fragmentation of VHKEC_{244}C_{245}HGDLLEC_{251}ADDR – C_{199}ASIQK assuming the intramolecular disulfide bond reduced.

Figure 2.55: The fragment map displays fragment ions resulting from ECD fragmentation of VHKEC_{244}C_{245}HGDLLEC_{251}ADDR – C_{199}ASIQK. Several peaks have been produced, enabling 87% sequence coverage.

2.4.2.4 Three Chains connected with two disulfide bonds

A peptide fragment peak generated after trypsin digestion rose with a molecular weight of 3815.8090. This peak corresponds to LKEC_{277}C_{278}DKPLLEK-SHC_{288}IAEVEK-YIC_{264}DNQDTISSIONK tripeptide with two disulfide bonds (Figure 2.56, Table Appendix 2). A specific charge state of the oxidised peptide ([M+6H]^{6+}) was isolated and studied alongside a simulation assuming two disulfide bonds. As displayed in Figure 2.56, the obtained mass spectra fit reasonably well with the mathematical simulation assuming two disulfide bonds. These primary findings confirm that the peptide is LKEC_{277}C_{278}DKPLLEK-SHC_{288}IAEVEK-YIC_{264}DNQDTISSIONK indeed, with two disulfide bonds intact.
Figure 2.56: Comparison of the ion [M+6H]^{6+} with a molecular weight of 3815.8090 with the mathematical simulation assuming LKEC_{277}DKPLLEK-SHC_{288}IAEVEK-YIC_{264}DNQDTISSK, with two disulfide bonds. This comparison suggests that the oxidised peptide is LKEC_{277}DKPLLEK-SHC_{288}IAEVEK-YIC_{264}DNQDTISSK with two intermolecular disulfide bonds.

A spectrum of oxidised LKEC_{277}DKPLLEK-SHC_{288}IAEVEK-YIC_{264}DNQDTISSK was obtained, and the same charge state ([M+6H]^{6+}) was isolated applying 65 pulse length in combination with 3 V collisional voltage (Figure 2.57, Table Appendix 8). The most abundant peaks in the spectrum are individual chains, several b and c fragments. In Figure 2.57, we can see that the most abundant peaks in the ECD spectra are several b/c fragments from the peptide edges and intact chains.
Figure 2.57: ECD spectra for the [M+ 6H]^{+}\text{ charge state of LKEC}_{277}\text{C}_{278}\text{DKPPLLEK-SHC}_{286}\text{IAEVEK-YIC}_{264}\text{DNQDTISSK with pulse length set to 65 ms.}

Our trials show that the ECD cleaves the disulfide bonds, leading to individual chains to be observed. The identification of individual chains gave an extra level of confidence to our assignment.
Further analysis showed that ECD led to the production of three types of fragments; the ones that come with both intermolecular disulfide bonds intact (Figure 2.58), the ones with a single intramolecular disulfide bond cleaved (Figure 2.59), and lastly, the ones with both disulfide bonds cleaved (Figure 2.60). Combining the three types of fragments together, the final achieved sequence coverage is 86%. Furthermore, all the identified ions were assigned within a 5 ppm error (Figure 2.61).

Figure 2.58: ECD spectra for the [M+ 6H]^{6+} charge state of LKEC_{277}C_{278}DKPLLEK-SHC_{288}IAEVEK-YIC_{264}DNQDTISSK with pulse length set to 65 ms.

Figure 2.59: ECD spectra for the [M+ 6H]^{6+} charge state of LKEC_{277}C_{278}DKPLLEK-SHC_{288}IAEVEK-YIC_{264}DNQDTISSK with pulse length set to 65 ms.
This result has further strengthened our confidence in the hypothesis that disulfide bridged peptide is LKEC\textsubscript{277}C\textsubscript{278}DKPLLEK-SHC\textsubscript{288}IAEVEK-YIC\textsubscript{264}DNQDTISSK indeed and that there are two intramolecular disulfide bonds. Based on the available evidence from Figure 2.58, it is fair to suggest a disulfide bond between Cys264 and either Cys277 or Cys278; which thus indicates that there is another disulfide bond between Cys288 and with the other Cys. However, one downside regarding our methodology is that it cannot distinguish further the disulfide pattern. Although the performance was not optimal, we still believe it has managed to successfully identify the second and the third disulfide bonds and limit the range of potential cysteine combinations.

**Figure 2.60:** The fragment map displays fragment ions resulting from ECD fragmentation of LKEC\textsubscript{277}C\textsubscript{278}DKPLLEK-SHC\textsubscript{288}IAEVEK-YIC\textsubscript{264}DNQDTISSK, assuming both intermolecular disulfide bonds reduced.

**Figure 2.61:** The fragment map displays fragment ions resulting from ECD fragmentation of LKEC\textsubscript{277}C\textsubscript{278}DKPLLEK-SHC\textsubscript{288}IAEVEK-YIC\textsubscript{264}DNQDTISSK. Several peaks have been generated, allowing 86% sequence coverage.
2.5 Conclusions

In summary, we report a new approach utilizing high-resolution FT ICR MS combined with pepsin and trypsin digestion for sequence disulfide-rich proteins and assigning their disulfide connectivity pattern. BSA was enzymatically digested, by trypsin or pepsin, under nonreducing conditions to generate fragments. These fragments were then chemically reduced by TCEP and injected into the FT ICR MS to be separated based on their synchrotron frequency. Afterwards, the fragments were assigned by running them against a database. Afterwards, the same process was repeated without chemically reducing the disulfide bonds prior to injection to the mass spectrometer analyser. Separated ions are online subjected to CID and ECD to generate extensive sequence information and locate the disulfide bonds spectra of each species. Combining the sequence coverage after trypsin and pepsin digestion, the total sequence coverage achieved was 97%.

Regarding the ability to study complex disulfide bridged peptides without employing an LC system, peptides with single, double, and triple chains were successfully studied. The approach in this chapter could be applied to study crosslinked or disulfide linkages in recombinant proteins.
3 Selective N-terminal mass defect labelling of proteins for improved de novo top-down sequencing

3.1 Introduction

3.1.1 De novo sequencing applications: Biopharmaceuticals

*De novo* peptide sequencing is a strategy for determining the amino acid sequence of a protein or peptide without prior knowledge of the amino acid series. This strategy relies on using tandem MS to generate a nested set of cleavages differing by one amino acid residue. Thus, each residue's identity is determined by calculating the mass difference of fragments differing from one residue. Therefore, this process can achieve the peptide series without a protein database, overcoming the shortcomings of database-dependent policies.\(^{207}\)

With the ever-rising number of complete genomes sequenced, one might assume that there is less need for *de novo* protein sequence assignment from MS.\(^{208}\) However, the genome sequences of most species are still unexplored. Even for known, modifications such as PTM events may hinder the assignment of all or part of the protein sequence, or at least the identity of the modifications. Thus, a comprehensive characterisation of the protein primary sequence often entails determination of the protein structure with the minimum assistance from genomic analysis—*de novo* protein sequencing.\(^{209}\)

Biopharmaceutical proteins such as monoclonal antibodies (mAb) are one such example. Their variable regions carrying the complementarity determining domains (CDRs) responsible for antigen-binding specificity are different for each mAb with no similar sequences in the protein databases. Therefore, *de novo* sequencing is needed to fully characterise this important class of proteins (Figure 3.62).\(^{210,211}\) Applications of the method range from drug discovery to quality control with thorough protein determination needed at each stage to ensure the reagents' efficiency, safety, stability, and batch-to-batch consistency.\(^{211,212}\) The high molecular weight of mAb (ca. 150 kDa for IgG), in addition to the presence of intrachain disulfide bridges, makes them a challenging target to study using a top-down approach with a record quoting sequence coverages only up to 30 %.\(^{211,213}\) Utilising a 'middle-down' strategy, the Fab and Fc fragments can be cleaved by a GingisKHan and IdeS specific proteases and then reduced, separating the heavy and light chains to generate smaller polypeptides (ca. 25 kDa).\(^{214-}\)
The size of the polypeptides chains generated using this strategy better meets the performance of prevailing MS systems. However, even then, sequencing is not trivial. Structural hindrances in most MS/MS spectra hinder accurate and full-length de novo sequence assignment. In the past years, the challenges of de novo peptide and protein sequencing from MS/MS spectra have garnered a lot of attention, which has triggered the development of many fragmentation techniques such as collision-induced dissociation (CID), higher-energy collisional dissociation (HCD), electron capture dissociation (ECD), electron transfer dissociation (ETD), infrared multiphoton dissociation (IRMPD) and Ultraviolet photodissociation (UVPD).

Figure 3.62: Diagram outlining structural aspects of a human antibody and how it can be broken down for middle down MS. The first step in the workflow involves separating by proteolytic cleavage of the IgG CDR-containing Fab from the constant Fc utilizing the IdeS/GingisKHAN enzyme, which cleaves just underneath the disulfides bonds (yellow) among the heavy chains. Afterwards, the generated peptides are chemically reduced.

3.1.2 The Need of De Novo Sequencing

De novo protein sequencing is one of the significant challenges associated with MS-based proteomics, particularly for novel proteoforms such as monoclonal antibodies where genome information is either restricted or not available most of the time. Furthermore, limitations in peptides cleavages and coverage combined with spectra interpretation uncertainties make
complete *de novo* assembly, even of small proteins, laborious. To demonstrate this, top-down *de novo* sequencing of bovine Ubiquitin (Ubiq, P0CG53) was attempted using ECD fragmentation.

**Figure 3.63:** A) Mass spectra of bovine Ubiquitin (Ubiq). The charge state distribution of Ubiq. B) The MS spectra of the [M+8H]$^8+$ charge state of Ubiq. The Ubiq was compared with the mathematical simulation assuming a molecular formula of $C_{378}H_{629}N_{105}O_{118}S_1$ (8559.616712 Da).

ESI analysis of Ubiq produced a mass spectrum containing [M+11H]$^{11+}$ to [M+6H]$^6+$ charge states (Figure 3.63, A, Table Appendix 9). A specific charge state of Ubiq ([M+8H]$^{8+}$) was then isolated, and the isotopic distribution was shown to be in good agreement with the predicted distribution of the 76 amino acid protein (Figure 3.63, B).

After verifying the molecular formula of Ubiq, ECD fragmentation was conducted on the isolated [M+11H]$^{11+}$ charge state. After SNAP deconvolution of the data set and comparison of the resulting ion mass list against the known sequence of Ubiq, almost full sequence coverage of the protein was obtained (Figure 3.64). As illustrated in Figure 3.64, the most abundant peaks in the ECD spectrum are several c- and (53 ions) and z-ions (44 ions) fragments (Table Appendix 10). The ECD data resulted in a sequence coverage of 91 %, in which the identified ions were assigned within a 3-ppm error (Figure 3.65).

Interestingly, no fragmentation is observed in the N-terminal region between aa36-39—a highly conserved proline region, suggesting that this closely nested proline region is not susceptible
to backbone cleavage under these applied ECD conditions. Thus, the generated data is consistent with previous findings and further support the idea that the only cleavage site that is not typically observed in ECD is the N-terminal side of proline due to its cyclic structure.\textsuperscript{223}

\textbf{Figure 3.64: ECD fragmentation spectrum of standard Ubiq.} The bottom spectrum shows a magnified region of the fragmentation spectrum with all assigned signals labelled.

\textbf{Figure 3.65: The fragment map displays fragment ions resulting from ECD fragmentation of Ubiquitin.} Red hooks indicate identified fragments. Specifically, ECD
fragmentation led to 91% sequence coverage, of which 34% of the identified ions were assigned within a 3 ppm error.

Figure 3.66: The potential protein sequences of Ubiq determined de novo employing conventional ECD fragmentation. Ambiguous sequences are highlighted in Red and unknown amino acid acids with green.

Next, this data set was used in an attempt to determine the primary sequence using conventional top-down de novo sequencing methodology, i.e. assuming no prior knowledge of the sequence. By employing a 3-ppm error threshold on the previously collected ECD data, it was feasible to define 89% of the candidate primary sequence (Figure 3.66) since most amino acids were successfully identified. Due to the dataset's high resolution and mass accuracy, it is essential to note that it was viable to sequence up to two amino acids confidently. As is well known, the amino acid masses are based on the sum of the elemental masses, thus leading to a unique mass for each of them. Therefore, adding two amino acids creates a unique number that corresponds precisely to this combination of residues giving us the ability to determine them confidently. For instance, residues 56 and 57 (top sequence) or 20-21 (bottom sequence) were correctly assigned as LS/SL, yet the correct order could not be defined as there was no internal fragment between the two residues. As anticipated, one downside regarding proteomics methodology is the incapability to distinguish between L and I residues as in most tandem MS case studies. To determine L/I accurately, additional runs would be required, collected from MS experiments leading to internal cleavages, such as hot ECD.²²⁴
However, it is worth mentioning that it is impossible to differentiate between the $c$ and the $z$ ion series from this single dataset without any prior knowledge. Thus, since the method could not distinguish between the $N$-termini and the $C$-termini ions, two potential sequences emerge, illustrated in Figure 3.66. These two sequences are the correct (forward) protein primary sequence, which has reversed the incorrect alternative.

To generate as many fragments as possible and thus to improve the identification coverage, researchers have employed a combination of fragmentation methods such as CID, HCD, and ECD. Although this approach is interesting, it is quite laborious and time-consuming. Characteristics that are not appealing, especially if we consider the rising need for sequencing antibodies fast and efficiently. Therefore, the findings supported the urgent need for novel methods for *de novo* sequencing. This research seeks to address this and a way to solve this issue by incorporating an MD tag on the $N$-terminus of proteins. Such a label would enable easy distinction between $N$- and $C$-terminal fragments from top-down MS data, thereby aiding fragment assignment.

### 3.1.3 Mass Defect

In this study, we explore the possibility of using MD to aid protein sequencing by allowing the identification of $N$- and $C$- terminal fragment ion series and thus simplifying the protein sequencing. Proteoforms are formed by a small subset of elements, the most abundant of which $^{12}\text{C}$ (MD = 0.000), $^{1}\text{H}$ (MD = +0.007825), $^{14}\text{N}$ (MD = +0.003074) and $^{16}\text{O}$ (MD = -0.005085) have low MD compared to $^{12}\text{C}$ (Table 3.3) meaning that we anticipate proteins to have an overall slightly negative MD.
Table 3.3: Nominal- and monoisotopic mass and mass defect of elements found in proteins and iodine and bromine.43

<table>
<thead>
<tr>
<th>Atom</th>
<th>Nominal mass/Da</th>
<th>Monoisotopic mass/Da</th>
<th>Mass defect/Da</th>
</tr>
</thead>
<tbody>
<tr>
<td>$^{12}$C</td>
<td>12</td>
<td>12</td>
<td>0.007825032</td>
</tr>
<tr>
<td>$^{1}$H</td>
<td>1</td>
<td>1.007 825 0322</td>
<td>0.007825032</td>
</tr>
<tr>
<td>$^{14}$N</td>
<td>14</td>
<td>14.003 074 004</td>
<td>0.003074004</td>
</tr>
<tr>
<td>$^{16}$O</td>
<td>16</td>
<td>15.994 914 619</td>
<td>-0.005085381</td>
</tr>
<tr>
<td>$^{32}$S</td>
<td>32</td>
<td>31.972 071 174</td>
<td>-0.027928826</td>
</tr>
<tr>
<td>$^{127}$I</td>
<td>127</td>
<td>126.904 47(3)</td>
<td>-0.095527</td>
</tr>
<tr>
<td>$^{79}$Br</td>
<td>79</td>
<td>78.918 338</td>
<td>-0.081662</td>
</tr>
</tbody>
</table>

If we study a polymer of repeating monomer units, the MD plotted against increasing monomer mass will show a linear relationship. This is illustrated in Figure 3.67, which demonstrates the MD of three polymers, poly(ethene), poly(phenylethene) and poly(chloroethene) (PE, PS and PVC), as their polymer chain length increases. PE is a simple long saturated hydrocarbon chain. Adding a phenyl group into the repeat unit (as in PS) increases the oxidation state, raising the MD, as introducing an element with a high MD (chlorine, as in PVC). Similarly, if we consider proteins as polymers, we should also see a trend in MD of increasing protein length size. This is described by considering the theoretical repeating monomer unit of polypeptides, polymer poly(averagine), (Figure 3.67, green line) and displays the expected MD along with a primary protein structure. According to the literature, averagine refers to the theoretical amino acid, with the molecular formula (C_{4.938}H_{7.758}N_{1.3577}O_{1.4773}S_{0.0417}), is defined to be the average composition of an amino acid in the human genome.226,227
The diversity of twenty amino acid sidechains unavoidably leads to variations in the monomeric units' elemental composition within proteoforms. Hence, MD with a rising nominal mass of polypeptides differs from the theoretical straight line illustrated by poly(averagine). This drift also differs among proteins and even within singular domains of the same protein. However, this variation is relatively small because the natural amino acids' elemental diversity is limited to CHNOS. This is described in Figure 3.68, which demonstrates the MD of the polypeptide sequence of both the C- and N-terminal domains of the tumour suppressor protein p53, along with the full length of AS. In these diagrams, the MD is represented as the molecular mass function as the polypeptide array raises by adding each sequential amino acid. Residues 1-96 of AS incorporated several repeating units and predominantly hydrocarbon and nitrogen sidechains, indicated in the uniform and a relatively steep reduction in the MD. However, the C-termini domain is profoundly acidic and induces a relative rise in MD due to a high number of oxygen atoms, which have a positive MD. Thus, the C-terminal region follows the poly(averagine) model throughout its length. Finally, the profoundly acidic N-terminus of p53 is determined to have a higher MD than averagine. Within this domain is a 19-residue "proline-rich" region containing only proline, alanine and valine residues, which can be identified by a steady drop in MD seen in the plot.

Figure 3.67: MD against nominal mass along the length of three polymers along with the hypothetical poly(averagine).
However, as illustrated in Figure 1.7, more prominent (non-organic) atoms have significantly higher negative MD, and this fundamental measurable difference is what is employed in the MD labelling approach.

3.1.4 Mass Defect Labelling in MS

Incorporating one or more chemical elements with atomic numbers within the 17 (Cl) and 77 (Ir) range into the bio-molecules of interest forms an evident difference among the masses of modified and unmodified species of the same nominal mass. The stable nuclei of these elements have extensively higher absolute MD values than those standard to biomolecules (carbon, hydrogen, oxygen, nitrogen, sulfur, and phosphorus). The difference in MD displays itself as a resolvable mass shift in most high-resolution mass spectrometers. As defined by Bajrami et al., the goal of the process is to shift a given protein, peptide, or peptide fragment ion into the ‘forbidden zone’ of the spectrum by altering its mass. Thus, the peptides are moved into a less noisy spectral space, making their detectability more likely and allowing effective identification of desired modified species (Figure 3.69).
Figure 3.69: Mass spectra of protein treated without (top) and with (bottom) mass defect tag. The c and z ions are closely nested on the top, leading to potentially overlapping species that would be hard to distinguish. The signals have been spread out across the m/z range on the bottom, allowing more straightforward and more confident detection.

This method is desirable for complex bottom-up proteomics studies that result from a polypeptide’s proteolysis by a particular enzyme and therefore matched to potential sequences employing a spectral protein database. In the literature, there are numerous examples of MD labelling combined with MALDI-MS to simplify spectral complexity since usually with MALDI as all ions are 1+. Notably, this approach has many practical applications in the field of chemical crosslinking.

Chemical cross-linking combined with MS has risen as a promising tool for determining protein conformation and protein-protein and protein-ligand interactions. This has led to a diverse range of commercially available cross-linking reagents, and well-developed protocols and MD labelling has been successfully incorporated into crosslinker design. MD tagging of the synthesised crosslinker allows prompt, straightforward assignment of cross-linked peptides from unreacted ones. Hernandez et al. recently reported a novel cysteine MD tag, 2,4-dibromo-(2'iodo) acetonilide (Figure 3.70), to improve the specificity of peptide identification. This is because adding an MD tag to the cysteine-containing peptides was found to improve both their detectability and identification specificity. It also improves their applications with
HPLC MALDI-FTICR MS. The MD linker tag integrates a unique mass shift and isotope signature and thus allows a more promising combination for automated, high-speed, and confident assignment of cross-linked products, thus enhancing proteoform detection specificity.

![Chemical Structure of dibromo-(2'-iodo) acetanilide.](image)

*Figure 3.70: Chemical Structure of dibromo-(2'-iodo) acetanilide.*

In this study we will develop an innovative MD labelling approach for top-down de novo sequencing. Selectively modifying a single terminus of a protein with a chemical tag containing a high MD (such as iodine, bromine, chlorine) will generate a characteristic shift in MD to all fragments from this terminus. This will aid confident and rapid N- and C-terminal ion series assignments in top-down fragmentation analysis without prior knowledge, thereby aiding fragment assignment. We hypothesise that this strategy will allow increased fragmentation coverage in top-down proteomic workflows.

### 3.1.5 Mass Defect plots using averagine scaling analysis

To convert the relative mass shift induced by the MD label into a form that enables straightforward distinction between labelled and unlabelled fragments in an MS/MS spectrum, the data needs to be displayed in a so-called MD plot. For hybrid mixtures dominated by hydrocarbons, the CH$_2$ repeating unit is given as a fixed mass of 14.183. This allows the Kendrick mass (KM) to be calculated from the exact mass of the compound by multiplying it by (14/14.0157), or the NM of the repeating CH$_2$ unit divided by the accurate mass of CH$_2$. The Kendrick mass defect (KMD) is defined by deducting the KM from the NM. Ultimately,
compounds can be constructed by drafting a KMD plot, in which NM is plotted against the KMD. Ions with the equivalent levels of saturation or functional group content will be on a horizontal line across the plot, and thus if one ion along this series can be identified, the chemical formation for others may be assumed to be comparable. For instance, saturated hydrocarbons will all have a nearly zero KMD; conjugating a carbonyl group will increase the KMD due to the oxygen atom's MD and increased oxidation state of the molecule. However, while CH$_2$ is a significant component of proteins, it is often better to refer to the repeating monomer unit of polypeptides as the averagine amino acid. Hence, it provides a more suitable mass reference for the MD study of proteoforms.$^{226,227}$

The conversion of masses to an averagine-based mass scale (Da$_{avg}$) is achieved by multiplying the monoisotopic masses delivered from the Dalton mass scale by a scaling factor of 0.999493894; determined as the ratio of the nominal to the monoisotopic mass of averagine given on the Dalton mass scale by the following equation:

$$\frac{NM_{Dalton}(averagine)}{MIM_{Dalton}(averagine)} = \frac{110 \text{ Da}}{111.0543052 \text{ Da}} = 0.999493894$$

This scaling sets the MD of averagine on the averagine mass scale to zero, and any peptide fragments with elemental compositions close to that will have approximately zero MD. As demonstrated by the example in Figure 3.71, constructing an averagine-scaled MD figure for a standard provides protein does not provide much information. As it can be seen, it is impossible to distinguish between $c$ and $z$ ions without prior knowledge as there is no consistent difference in mass defect between the two ion series. However, it would be helpful if there was a way to differentiate between the two series to build up the primary sequence. In theory, adding a mass defect offset to one of the ion series would separate the two in the mass defect plot, allowing them to be easily distinguished. This study explores the possibility of adding this mass defect offset by selectively adding a mass defect tag to the protein N-terminus.

The theoretical advantage of this strategy is shown in Figure 3.72. Figure 3.72 demonstrates that the construction of an averagine-scaled MD figure for a protein labelled selectively with an MD tag on the N-termini provides confirmatory evidence for this graphical approach's usefulness. It is apparent from Figure 3.72 that the addition of this tag leads to the generation of two series of fragment ions that are clearly distinguished in the MD plot. One series with
close to zero MD consisted of the untagged fragments (C-terminal z-ion fragments) and a row of points with distinct negative MD set by the fragments carrying the MD label (N-terminal c-type fragments).

A closer look at the data indicates that the lines are near-horizontal due to the averagine scaling, but they are not straight. The twenty amino acid sidechains' diversity unavoidably ends with variations in the elemental composition of the monomeric units within proteins. However, despite the variations in the elemental composition of the monomeric units, two trends are clearly illustrated. Thus, the trend of MD with rising NM of proteins deviates from the theoretical straight line illustrated by polyaveragine. The separation between the rows is roughly equivalent to the effective (averagine-scaled) MD of the label and should be high enough to distinguish the two trends confidently. Therefore, the MD plot allows for easy resolution of tagged from untagged fragments and, in the case of a terminus-selective label, among N- and C-terminal fragments.  

In principle, either the N- or C-terminus could be used as a potential location for incorporating an MD tag. However, due to its potential reactivity, in this study, N-terminal labelling was investigated.
Figure 3.71: Mass defect analysis illustrated using the peptide GYSKEASAL. The MS software provides the list of the monoisotopic masses (MIM). Afterwards, this mass is corrected by the averagine factor to standardize it for every protein (MIM$_{avg}$). The nominal mass (NM$_{avg}$) is calculated by rounding the previously estimated number, and finally, the MD is calculated by subtracting from the MIM$_{avg}$ the NM$_{avg}$. The MD plot was constructed by plotting MIM against MD. As illustrated, c and z ions overlap with each other, thus making it impossible to distinguish between the two classes without any prior knowledge.
Figure 3.72: Mass defect analysis illustrated using the peptide GYSKEASAL labelled on the N-termini with an MD tag (C7H5I, monoisotopic mass: 215.9436). As before, the MS software provides the list of the monoisotopic masses (MIM). Afterwards, this mass is corrected by the averagine factor to standardize it for every protein (MIM$_{avg}$). The nominal mass (NM$_{avg}$) is calculated by rounding the previously estimated number, and finally, the MD is calculated by subtracting from the MIM$_{avg}$ the NM$_{avg}$. The MD plot was constructed by plotting MIM against MD. A closer look indicates two trends corresponding to c (red) and z ions (blue ions). Therefore, labelling selectively either the N- or C-terminus allows for easy resolution of tagged from untagged fragments and, in the case of a terminus-selective label, among N- and C-terminal fragments.
3.1.6 Chemical Biology Strategies for Selective Chemical Modification of the Protein N-terminus

For this strategy to succeed, N-terminal labelling must be highly site-specific to the N-terminus, and sidechains labelling should be avoided; otherwise, the assignment cannot be done confidently.

Current efforts in the protein conjugation field have predominantly focused on site-specific techniques to append a single synthetic molecule to a specific site on a protein. Such labelling reactions are valuable because they provide well-defined conjugates with uniform features. However, those reactions demand a high degree of selectivity towards certain functional groups while proceeding in aqueous solutions under moderate pH and temperature conditions to preserve the protein's structural integrity and biological activity. This restriction provides an exceptionally demanding setting for organic reactions, as the high abundance of most amino acid residues on proteins narrows the point to which site-specificity can be reached. Hence only a finite set of biochemical reactions have been demonstrated for site-specific modification. Conventional methods for bioconjugated complex proteoforms in single sites include the alkylation of introduced Cys residues, the introduction of unnatural amino acid residues with distinct reactivity into protein sequences, native chemical ligations, and proteolytic labelling methods. Although all these processes are highly beneficial, they require significant degrees of protein engineering, leading to diminished protein expression levels and disrupted protein function. Moreover, these approaches cannot be easily implemented to a broad range of proteins.

As a complement to these techniques, there has been growing interest in targeting the protein N-termini as this site is uniquely reactive and solvent-accessible on most expressed proteoforms. Conventional approaches to address this issue exploit the slight difference in pKa between the N-terminal α-amines (pKa = 6–8) and lysine amines (pKa ~10.5) to selective modify the N-terminus. This minor difference favours the N-terminus' selective acylation and alkylation when conducting chemical reactions at a low-to-neutral pH range. Although these systems are appealingly straightforward to perform, they are unfortunately limited to cases in which few lysine side chains compete; thus, total site-specificity is hardly ever achieved.
Figure 3.73: Standard strategies for modifying N-terminal amino acid residues. (a) Thioesters react with N-terminal Cys by native chemical ligation. (b) Thiazolidines are formed by the reaction between aldehydes with N-terminal Cys. (c) Sodium periodate is employed to form aldehydes from N-terminal serine and threonine residues. (d) Pictet-Spengler reactions are undergone between N-terminal tryptophans and aldehydes. (e) The N-terminal amino group was converted to a carbonyl group by transamination reagents. (f) N-terminal prolines react with aminophenols by an oxidation reaction (g) 2-pyridinecarboxaldehyde (2PCA) was employed to label the N-termini selectively, the nitrogen of the adjacent amino acid cyclises on the imine to form an imidazolidinone product at the N termini.

Several methods rely on participation from the N-terminal residues amino acid side chains for efficient chemical modification. For instance, N-termini Cys residues have been successfully modified with thioesters, and reactive aldehydes have been employed to form N-termini thiazolidines (Figure 3.73 a, b). In addition, reactive aldehydes can be formed through
periodate oxidation of N-terminal serine and threonine amino acid residues for consequent oxime ligation or strain aided to alkyne-nitrone cycloaddition, and N-terminal tryptophans can be conjugated via Pictet–Spengler reactions (Figure 3.73 c, d). A recent study by the Francis group has demonstrated a site-specific transamination chemical reaction that introduces reactive ketones or aldehydes at the N-termini along with an oxidative coupling reaction among aminophenols and N-terminal proline residues (Figure 3.73 e,f). After imine condensation, the amide nitrogen of the adjacent amino acid cyclises on the imine to form an imidazolidinone product at the N termini (Figure 3.73, g). Notably, the lysine sidechain lacks the nearby amide group and thus does not produce stable products. This study underlines that the N-terminus has the sole ability to react in combination with its side chain group, often enabling selective incorporation through the form of cyclic intermediates and products-conjugates. This reaction was demonstrated for 12 different proteins, including the soluble domain of the human estrogen receptor. These results offer indisputable evidence that the reaction is independent of the N-terminal residue's sidechain.

Furthermore, they applied their approach to label antibodies, which preserved their ability to bind targets even after labelling. Although this approach is more selective, it fails to modify proteins with proline as the first amino acid and doubly conjugate polypeptides with glycine residues at the N-termini, hence lacking ubiquitous sequence compatibility. Effective as they are, all of these systems restrain the particular N-terminal amino acid residues that can be present and need multi-steps. Consequently, a simple, one-step method competent for modifying a single site on a broad range of proteins would be of high value to the field of chemical biology.

3.2 Research Outline

Given the drive for innovative strategies that increase fragment ion assignment efficiency in top-down mass spectra and therefore maximise the protein sequence coverage obtained, we have developed a strategy (Figure 3.74) for utilising the reductive alkylation reaction described by several researchers to the attachment of N-terminal selective MD tags (Table 3.4) into proteins. Such a tag aids classification between N- and C-terminal fragments from top-down MS data, thus simplifying fragment assignment.
Table 3.4: Halogen-based MD tags employed in this research to increase fragment ion assignment efficiency in top-down mass spectra and therefore maximise the protein sequence coverage obtained. Such a tag aids classification between N- and C-terminal fragments from top-down MS data, thus simplifying fragment assignment.

<table>
<thead>
<tr>
<th>Name</th>
<th>Structure</th>
<th>Formula</th>
<th>Molecular Weight</th>
<th>Δmass</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-Iodobenzaldehyde</td>
<td><img src="image1" alt="Image" /></td>
<td>C\textsubscript{7}H\textsubscript{5}IO</td>
<td>231,9385</td>
<td>215,9436</td>
</tr>
<tr>
<td>5-Bromo-2-pyridinecarboxaldehyde</td>
<td><img src="image2" alt="Image" /></td>
<td>C\textsubscript{6}H\textsubscript{4}BrNO</td>
<td>186,0059</td>
<td>142,9371</td>
</tr>
</tbody>
</table>

![Image](image3)

Figure 3.74: The Top-down workflow in combination with a mass defect labelling. Protein is selectively labelled on the N-terminus with a covalent MD label before top-down fragmentation. The large MD allows the classification of fragment ions into N-terminal (c-ions) and C-terminal (z-ions) fragment series without prior sequence information.

In this context, the aim was to develop a method for the N-terminal selective attachment of halogen-based MD tags into proteins. For the MD label to be a valuable tool for the proposed
application of *de novo* sequencing, the labelled conjugate yield needs to be high enough to efficiently isolate and fragment in the mass spectrometer. Finally, the label must display high selectivity toward the *N*-terminus to minimise false positives.

Several proteins were used for the development of this strategy. Initial experiments were performed using Insulin (Ins, P01317) was selected as a study protein due to the successful incorporation of benzaldehyde previously demonstrated by Chen et al.\textsuperscript{258}, and MacDonald et al.\textsuperscript{259}. Additionally, its small size (5.7 kDa) should make it easier to study, and the fact that it only contains a single lysine residue placed towards the *C*-terminal end of its B-chain would greatly simplify top-down analysis (Figure 3.75, a\textsuperscript{47}). With the ultimate goal of applying the tagging method to various proteins, the technique will also need to verify that the method applies to other proteins. Bovine Ubiquitin (Ubiq, P0CG53) was also chosen in part due to its small size (8.5 kDa) and its well-characterised top-down fragmentation profile (Figure 3.75, b)\textsuperscript{260}.

On the other hand, bovine Myoglobin (Myo) was chosen as a more challenging protein with a mass of 16.9 kDa (Figure 3.75, c)\textsuperscript{261}. Furthermore, for the system to be valuable in the characterisation of PTMs by top-down MS, one would also need to study the effect of the *N*-terminal alkylation reaction on PTMs. Therefore, Bovine Ribonuclease A (RNase A, P4557) was chosen as a study protein with a mass of 13.7 kDa since it has 4 disulfide bonds (Figure 3.75, d)\textsuperscript{262}. 
Figure 3.75: Structural features of a) insulin$^{263}$, b) ubiquitin, myoglobin$^{261}$ and Ribonuclease A$^{262}$. a) Primary structure of the structure of porcine insulin (5.7 kDa). The protein contains 6 cysteine residues, which form 3 disulfide bonds. The disulfide connective is highlighted. b) Primary structure of the structure of Ubiquitin (8.5 kDa). c) Primary structure of the structure of Myoglobin (16.9 kDa). d) Primary structure of the structure of bovine Ribonuclease A (13.7 kDa). The protein contains 8 cysteine residues, which form 4 disulfide bonds. The disulfide connective is highlighted.
3.3 Materials and Methods

3.3.1 Reagents

Ultrapure water, acetonitrile and DMSO was purchased from Thermo Fisher. All solvents were LCMS grade. Tris(2-carboxyethyl) phosphine (TCEP) was purchased from Thermo Fisher. The remaining chemicals (5-Bromo-substituted 2-pyridinecarboxaldehyde 4-iodobenzaldehyde, trisodium citrate, citric acid, sodium cyanoborohydride, disodium phosphate, monosodium phosphate, guanidine-HCl (2 M), formic acid (≥ 96 %), ammonium acetate) and proteins (insulin from bovine pancreas, Myoglobin from equine (horse) heart, Ribonuclease A from bovine pancreas and Ubiquitin from bovine erythrocytes) were purchased from Sigma Aldrich. Zip-tips (OMIX C4, 100 μL) and size-exclusion spin columns (Micro Bio-Spin) were obtained from Agilent Technologies and Bio-Rad, respectively. Ultrapure water was used for all reactions and MS experiments unless otherwise stated.

3.3.2 Incorporation of a 4-iodobenzyl IMDT via Reductive Alkylation

Selective N-terminal labelling was performed by adapting a published reductive alkylation reaction presented by Chen et al.\textsuperscript{258} Ins and Myo samples (1 mM) were dissolved in ultrapure water and N-terminally labelled by addition of sodium cyanoborohydride (5 mM) and 4-iodobenzaldehyde (1 mM). The reaction was performed in citric acid buffer (0.1 M, pH 6.1) and allowed to proceed in the dark with gentle mixing at 47 °C and 37 °C for Ins and Myo, respectively (Figure 3.76). Afterwards, TCEP.HCl (50 μL, 20 mM) was added to the Insulin samples and left incubating at 37 °C for 60 min to ensure that all the disulfide bonds were reduced.
Figure 3.76: Reaction scheme for the N-terminal selective incorporation of 4-iodobenzaldehyde into Ins and Myo proteins.
3.3.3 Incorporation of a Bromo-substituted 2-pyridinecarboxaldehyde Mass Defect Tag via imine condensation

Selective N-terminal labelling was performed by adapting a published transamination strategy. Ubiq, Myo and RNase samples (50 μM) were dissolved in ultrapure water, and N-terminally labelled by adding 5-Bromo-2-pyridinecarboxaldehyde (4 mM). The reaction was performed in phosphate buffer (0.1 M, pH 7.9) and allowed to proceed in the dark with gentle mixing at 37 °C overnight (Figure 3.77). Afterwards, TCEP.HCl (15 μL, 100 mM) was added to modified RNase A and was incubated at 90 °C for 5 min. N-Ethylmaleimide (NEM) (30 μL, 100 mM) was added to the sample and left incubating at 37 °C for 20 min.

![Reaction scheme for the N-terminal selective incorporation of 5-Bromo-substituted 2-pyridinecarboxaldehyde into proteins.](image)

**Figure 3.77:** Reaction scheme for the N-terminal selective incorporation of 5-Bromo-substituted 2-pyridinecarboxaldehyde into proteins.

3.3.4 Time-course experiments at varying temperatures

Time course reactions to monitor reaction rate were performed at 17 °C (measured room temperature), 37 °C, 47 °C and 67 °C. Aliquots were taken every hour for insulin and every second hour for Myoglobin up to 6 hours (day 1) and then after 23 hours (day 2).
3.3.5 LCMS to Confirm Conjugation and Assess Yield

The yield of the reactions was assessed by LCMS analysis of the crude reaction mixture (diluted to 10 μM) on a Synapt G2 UPLC-Q-ToF (Waters) instrument with a reverse-phase C4 column. Relative yields of the unmodified protein and singly and multiply modified conjugates were evaluated from relative abundances by considering each peak's contribution to the sum of peaks in the charge state.

3.3.6 Sample Preparation for Top-Down Mass Spectrometry

Samples were desalted by Bond Elut OMIX C4 Ziptip (Agilent Technologies) or using size-exclusion spin columns eluting in acetonitrile 50 % v/v, 0.1 % v/v formic acid or 100 mM ammonium acetate (buffer exchange).

3.3.7 Top-Down Mass Spectrometry

Mass spectrometry was carried out using a 12 T SolariX FT-ICR instrument mass spectrometer (Bruker Daltonics, Billerica, MA, USA) with ESI ionisation. Proteins samples (10 μM) were prepared in denaturing solution conditions with ACN: H₂O: FA ratio of 49.95:49.95:0.1. Nitrogen was employed as the collision gas. MS experiments were performed in the broadband mode from m/z 300–2000. A quadrupole Q1 isolated precursor ions of single charge states with a selection window size of 1–5 m/z. Fragmentation was performed, and a spectrum was recorded, accumulating 150-200 scans to improve the signal-to-noise ratio. ECD fragmentation was performed with a 25 ms ECD pulse length, and CID was performed using a suitable voltage depending on the degree of activation desired. The raw data was processed in Compass Data Analysis (Bruker) using the built-in SNAP tool for peak picking and deconvolution of isotopic distributions. Fragment assignment was performed by comparing fragment monoisotopic masses to the known protein sequence in ProSight Lite²⁶⁴ (Dr Neil Kelleher, University of Illinois) with mass errors within ± 5 ppm.
3.3.8 Mass Defect Analysis and De Novo Sequencing

In the case of iodine-based mass defect tag (IMDT; delta mass: $\text{C}_7\text{H}_5\text{I}$), a peak list was provided by the Sophisticated Numerical Annotation Procedure (SNAP) used as part of the Bruker DataAnalysis software. This algorithm relies on matching observed isotope patterns to that of the theoretical amino acid "averagine", which has the molecular formula $\text{C}_{4.9384}\text{H}_{7.7583}\text{N}_{1.3577}\text{O}_{1.4773}\text{S}_{0.0417}$, calculated to be the average composition of an amino acid in the human genome. To retrieve every generated ion SNAP identified, we set the quality factor threshold at 0.3, the signal-to-noise ratio (S/N) threshold at 4, the relative intensity threshold (base peak) at 0.01%, the absolute intensity threshold at 0 and the maximum charge state to 20. The repetitive building block was set to $\text{C}_{4.9384}\text{H}_{7.7583}\text{N}_{1.3577}\text{O}_{1.4773}\text{S}_{0.0417}$. SNAP returned a matrix with columns for $m/z$, charge, intensity, resolving power, and molecular mass. However, this was not possible with the bromine-based mass defect tag (BrMDT; delta mass: $\text{C}_4\text{H}_2\text{NBr}$) since bromine has a distinct isotope pattern. Thus, two generated peak lists were incorporated together; one relied on the molecular formula $\text{C}_{4.9384}\text{H}_{7.7583}\text{N}_{1.3577}\text{O}_{1.4773}\text{S}_{0.0417}$ as before, and one on the molecular formula $\text{C}_{5.3384}\text{H}_{7.9583}\text{N}_{1.4577}\text{O}_{1.4773}\text{S}_{1.4773}\text{Br}_{0.1}$. Afterwards, monoisotopic fragment masses were scaled using the averagine scaling factor of 0.999493894 and rounded to the nearest integer. The scaled MD was calculated as the difference between the scaled monoisotopics mass and the integer value. MD plots were constructed by plotting the scaled mass defects against the monoisotopic masses of the fragments. As identified from the MD plot, the $N$- and $C$-terminal ions were split into separate ion series, and the two-ion series were sequenced individually. Starting at the termini, ion series were extended by considering the mass difference between the n and n+1 ions, where n was the last identified ion in the series and n+1 was the following ion in the data set and matching against a table of amino acid masses. If no amino acid match was found, the process was repeated for the n and n+j, j = 2, 3, 4, ..., ions until a match was found. The amino acid identified was accepted if the ion’s experimental mass error compared to the theoretical mass of the candidate sequence was less than ± 5 ppm.
3.4 Results and Discussion

3.4.1 Chemical Labelling of Insulin using reductive alkylation incorporating an N-terminal selective iodine-based mass defect tag

**Figure 3.78**: Reaction scheme for the N-terminal selective incorporation of a) benzaldehyde and b) 4-iodobenzaldehyde into proteins.
This study aimed to extend the utility of the reductive alkylation described by the Hung-Chieh Chou group to incorporate an N-terminal selective iodine-based mass defect tag (IMDT; delta mass: C7H5I, monoisotopic mass: 215.9436) into the N-terminus of proteins (Figure 3.78). Initially introduced as a strategy for site-selective N-terminal conjugation of native proteins by Chen et al. in 2017, the reductive alkylation reaction utilises the variation in pKa among the N-terminal amine (pKa ≈ 8) and lysine primary amines (pKa ≈ 10) to accomplish selectivity for the N-termini. The authors studied various buffers and pH conditions to achieve the highest selectivity employing benzaldehyde for conjugation. They achieved >99:1 selectivity for the peptide XYSKEASAL where X varies over the 20 amino acid residues, leading to an incredible conversion (≥78 %) except the peptide bearing N-terminal cysteine (42 %). Furthermore, the group illustrated that the label could be conjugated into various proteins, including insulin, with reasonable conversion rates.258

Initial optimisations of the conditions revealed that increasing the amount of reducing agent from 5 to 15 equivalents led to an increase in the reaction yield by 16 %. Thus, IMDT tags were successfully incorporated into Ins, as demonstrated by LCMS analysis (Figure 3.79, A). As illustrated in Figure 3.79, Ins is a 51-residue protein (5733 Da) with two polypeptide chains connected by one intrachain and two interchain disulfide bonds.

On these grounds, we can hypothesise that Ins will be double modified since there are two available N-terminus to be chemically modified. The data yielded by this study provide convincing evidence that Ins has been successfully doubled modified as anticipated (Figure 3.79, A). A specific charge state of oxidised Ins ([M+5H]^5+) was then isolated and studied alongside mathematical simulations of the unmodified and modified protein.
Figure 3.79: A) Mass spectra of insulin (Ins). Top: Controls showing the charge state distributions of the unmodified protein. Bottom: Spectra of the crude reaction mixture with a combination of unmodified protein and species with higher mass. B) Observed isotopic distributions of Ins overlaid with simulated isotopic distributions for the a) unmodified (C_{254}H_{377}N_{65}O_{75}S_6, red), b) singly modified (C_{261}H_{382}N_{65}O_{75}S_{61}, blue) and di-modified (C_{268}H_{387}N_{65}O_{75}S_{62}, yellow) protein as represented by coloured rings.
3.4.2 Time-course experiments of reductive alkylation incorporating N-terminal selective iodine-based mass defect tag at varying temperatures of Insulin

Next, the impact of temperature on reaction progression and yield was determined by performing a set of time-course experiments at various temperatures (17, 37 and 47 °C) (Figure 3.80).

As highlighted in Figure 3.80, the unmodified proteoform disappeared throughout the time as it was used up in the chemical reactions to generate the modified species. Furthermore, the relative yield of the di-modified conjugate of Ins had a steady increase over the whole time range without reaching a plateau, and the same can be observed for the tri- and tetra-modified conjugates. The most striking observation from the data comparison was that rise was already visible at the 1 hr time point. The highest yield of the di-modified species was 31 %, and it was achieved after 23 hours of reaction time at 47 °C.
Figure 3.80: Relative yields from Ins within a time course experiment at varying temperatures. As illustrated, the unmodified proteoform disappeared throughout the time as it was used up in the chemical reactions to generate the modified species. Furthermore, the relative yield of the di-modified conjugate of Ins had a steady increase over the whole time range without reaching a plateau, and the same can be observed for the tri- and tetra-modified conjugates. The most striking observation from the data comparison was that rise was already visible at the 1 hr time point. The highest yield of the di-modified species was 31 %, and it was achieved after 23 hours of reaction time at 47 °C.
3.4.3 Evaluating Selectivity of the Chemical Labelling of Insulin using a reductive alkylation incorporating N-terminal selective iodine-based mass defect tag

Figure 3.81: ECD fragmentation spectrum of Ins modified with iodobenzaldehyde. The charge state distribution of Ins, 5+ charge state isolated (middle) and with ECD (bottom). The bottom spectrum shows a magnified region of the fragmentation spectrum with all assigned signals labelled.

The selectivity level for N-termini amines over lysine residues was the most crucial variable for evaluating the chemical reaction's suitability for the intended top-down sequencing applications.
In previous studies, to determine the $N$-terminal selectivity of the reaction, trypsin digestion and MS analysis was performed. However, this approach may be unreliable since adopting a bottom-up approach may lead to missing information on nonspecific modifications. Thus, top-down fragmentation was performed on the modified species to determine the location of the resulting chemical modification. As mentioned above, Ins comprises two chains bound by intermolecular disulfide. To simplify the data analysis, Ins was chemically reduced after the incorporation of the conjugation. Surprisingly, upon MS analysis of the reduced Ins, only the $B$-chain was detected. Although it would have been ideal for analysing the $A$-chain as well, this discrepancy is negligible since the $A$-chain does not contain any lysine residues, and thus, only $N$-termini labelling would be expected. ECD fragmentation was conducted on the isolated $[M+5H]^5+$ peak of the singly modified $B$-chain conjugate, accomplishing almost full sequence coverage of the chain (Figure 3.81).

**Figure 3.82:** Sequence of the Ins $B$-chain with an MD tag on the $N$-terminus (top) or Lys-29 (bottom) matched against the ECD fragmentation data for Ins modified under initial conditions. Red hooks indicate identified fragments.

Figure 3.82 displays that the most abundant peaks in the ECD spectrum are several tagged $c$- and (here termed $c^*$-ions, 37 ions) and untagged $z$-ions (22 ions) fragments. Both of which imply the conjugation was incorporated on the $N$-terminus. Only two potential untagged $c$-ions were observed, $c_{24}$ and $c_{26}$, suggesting that either the label had dissociated during fragmentation or the modification has occurred on the Lys-29 residue. As illustrated in Figure 3.82, two tagged $z$-ions were also seen, $z_6$ and $z_9$, implying the tag on Lys-29.
Figure 3.83: Plot showing the intensities of all fragments assigned from ECD fragmentation data for Ins modified under initial conditions. The displayed ratios highlight the relatively low intensity of the fragments suggesting tagging of Lys-29 compared to their counterparts, suggesting an N-terminal tag.

Table 3.5: Identified fragments sorted according to ion type and whether they are consistent with a tag on the N-terminus or Lys-29. The ratios indicate selectivity based on the summed intensity of fragments suggesting N-terminal labelling versus fragments suggesting labelling of Lys-29.

<table>
<thead>
<tr>
<th>Fragmentation method</th>
<th>Tag on N-term</th>
<th>Tag on Lys-29</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$b/c$</td>
<td>$y/z$</td>
</tr>
<tr>
<td>ECD</td>
<td>37</td>
<td>22</td>
</tr>
<tr>
<td>CID</td>
<td>27</td>
<td>19</td>
</tr>
</tbody>
</table>
Table 3.6: Comparison of specific Identified indicating a tag on the N-terminus or Lys-29. The ratios indicate selectivity based on the summed intensity of fragments suggesting N-terminal labelling versus fragments suggesting labelling of Lys-29.

<table>
<thead>
<tr>
<th>Tag on N-term</th>
<th>Tag on Lys-29</th>
<th>Ratio (Tag on N-term/Tag on Lys-29)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fragment</td>
<td>Error</td>
<td>Intensity (%)</td>
</tr>
<tr>
<td>c*24 -0.35</td>
<td>0.01365</td>
<td>0.3615</td>
</tr>
<tr>
<td>c*26 -1.11</td>
<td>0.27046</td>
<td>0.00481</td>
</tr>
<tr>
<td>z6 -0.45</td>
<td>0.63368</td>
<td>0.00406</td>
</tr>
<tr>
<td>z9 -0.24</td>
<td>0.51513</td>
<td>0.0051</td>
</tr>
</tbody>
</table>

It is well known that ion intensity needs to be interpreted with caution since it is not a quantitative measure of exact abundances, and whether an ion is observed in MS relies on various elements such as solubility and number of ionisable groups. However, fragment intensities were compared for structurally similar ions to evaluate selectivity since they have comparable ionisation characteristics (Figure 3.83). The single most striking observation to emerge from the comparison of the modified and unmodified fragment ion intensities was that the ions implying tagging of Lys-29 were of significantly lower intensity than those proposing the incorporation of the N-termini (Table 3.5). Similarly, the study at all generated fragments and their relative intensity, the fragmentation data proposes that the MD tag shows high selectivity for the N-termini (Table 3.6).

3.4.4 De Novo Sequencing of Insulin based on a reductive alkylation incorporating N-terminal selective iodine-based mass defect tag

As stated in the Introduction, we initiated this study to demonstrate MD analysis in a de novo sequencing context; Ins was sequenced using a simple de novo top-down (CID and ECD) sequencing methodology after being selectively modified on the N-termini, assuming no prior knowledge.
3.4.4.1  *De Novo* Sequencing of Insulin based on CID fragmentation and a reductive alkylation incorporating *N*-terminal selective iodine-based mass defect tag

CID fragmentation was conducted on the isolated [M+5H]^{5+} after applying 10 V collisional voltage (Figure 3.84, A). After a mass list was constructed using SNAP, an MD plot was produced employing an averagine scaling factor for the detected fragment ions from the Ins CID fragmentation dataset (Table Appendix 12), as previously explained. Inspection of this mass defect plot clearly illustrates that the fragments formed two well-defined series with the series centred on mass defect ca 0.2 corresponding to the *N*-terminal modified *b*-ions ions and the series centred around MD ca 0 representing the unmodified *C*-terminal *y*-ions (Figure 3.84, B).

After successfully separating *N*- and *C*-termini fragment ions, the two separate data sets were *de novo* sequenced individually. The difference between two consecutive fragments corresponds to the mass of specific amino acid, and this process is illustrated in Figure 3.85. The mass difference between the \(b_{n+1}\) and \(b_n\) corresponds to the mass of histidine (H). If we now turn to the mass difference between the \(b_{n+2}\) and \(b_{n+1}\), the result suggests that the following amino acid residue is either a leucine (L) or isoleucine (I). Thus, this data suggests that either L or I follow an H. By repeating this process for all the obtained data, this protocol seeks to define the peptide's primary sequence.
Figure 3.84: A) CID fragmentation spectrum of Ins B chain modified with iodobenzaldehyde. B) Mass defect plot derived employing averagine scaling correcting factor for the assigned fragment ions from Ins CID fragmentation dataset.
By employing a 3-ppm error threshold, it was not only feasible to confidently assign the right candidate primary sequence by considering adjacent amino acids, but it was also viable to sequence sections of up to two amino acids accurately. Therefore, residues 3 and 4 were rightly assigned as NQ/QN, yet the correct order could not be defined as there was no internal fragment between the two residues. Using this mass defect aided \textit{de novo} sequence process, 90% of the primary sequence was confidently assigned (Figure 3.86). All but residues 28-30 were not assigned due to the lack of fragmentation within this region.

\textbf{Figure 3.85}: On the top, a mass defect plot derived employing averagine scaling correcting factor for the assigned fragment ions from Ins CID fragmentation dataset and on the bottom a zoomed region (600-1400 Da) of this plot.
One downside regarding our methodology is that there have been 104 fragments that have not been determined. A reasonable explanation for this outcome may be that these are internal cleavages and/or sidechain losses, a phenomenon common to CID. Nevertheless, 90% of the primary sequence was assigned using this analysis. However, in higher molecular weight systems, internal cleavages and/or sidechain losses will likely limit the viability of this technique. One possible way to avoid this issue is to use ECD instead of CID. ECD is known to be more specific for backbone fragmentation, generating \( c \) and \( z \) ions, with little neutral loss and sidechain cleavage.

![Protein sequence](image)

**Figure 3.86: The protein sequence of Ins B chain determined de novo combining an N-termini MD tag and CID cleavage.** Although the performance was not ideal, we nevertheless managed to determine 90% of the primary sequence was assigned. Ambiguous sequences are highlighted in Red and unknown amino acids with green.

### 3.4.4.2 De Novo Sequencing of Insulin based on ECD fragmentation, and a reductive alkylation incorporating \( N \)-terminal selective iodine-based mass defect tag

ECD fragmentation was conducted on the isolated \([M+5H]^{5+}\) peak of the singly modified B-chain conjugate with a pulse length of 25 ms (Figure 3.87, A, B, Table Appendix 13). As before, the SNAP algorithm was employed, and an MD plot was derived employing averagine scaling correcting factor then for the fragment ions from Ins ECD fragmentation dataset (Figure 3.87, B). Again, the fragment ions formed two well-defined trends with the modified \( c \)-ions, corresponding to the \( N \)-terminal ions, displaying a mass defect offset of ca. 0.15 and the unmodified \( z \)-ions displaying MD around 0.
Figure 3.87: A) ECD fragmentation spectrum of Ins B chain modified with iodobenzaldehyde. B) Mass defect plot derived employing averagine scaling correcting factor for the assigned fragment ions from Ins ECD fragmentation dataset.
Figure 3.88: On the bottom, a mass defect plot derived employing average scaling correcting factor for the assigned fragment ions from Ins ECD fragmentation dataset and on the right, a zoomed region (300-800 Da) of this plot.

After successfully separating N- and C-termini fragment ions, the two separate data sets were sequenced individually. As it is already established, each fragment is illustrated by a specific bullet point. The difference between two consecutive fragments corresponds to the mass of specific amino acids, and this process is illustrated in Figure 3.88. The difference between the $c_{n+1}$ and $c_n$, $c_{n+2}$ and $c_{n+1}$, and $c_{n+3}$ and $c_{n+2}$ corresponds to valine (V), asparagine (N) and glutamine (Q), respectively. Thus, this data implies that a V is followed by N and Q.
By employing a 3-ppm error threshold, it was feasible to define the right candidate primary sequence. Again, it was viable to sequence sections of up to two amino acids accurately. Therefore, residues 22 and 23 were rightly assigned as RG/GR, yet the correct order could not be defined as there was no internal fragment between the two residues. However, amino acid residues 27-28 were rightly sequenced as threonine and proline, considering ECD does not fragment on the N-terminal side of prolines. The simplified de novo sequence process described here led to the determination of 100% of the primary sequence (Figure 3.89).

![Protein Sequence](image)

**Figure 3.89: The protein sequence of Ins B chain determined de novo combining an N-termini MD tag and ECD fragmentation. Ambiguous sequences are highlighted in Red.**

3.4.4.3 De Novo Sequencing of Insulin based on CID, ECD fragmentation, and a reductive alkylation incorporating N-terminal selective iodine-based mass defect tag

By merging the obtained sequences from CID (Figure 3.86) and ECD (Figure 3.89) fragmentation, the correct amino acid residues series have been successfully defined with a high level of confidence (Figure 3.90). However, it is vital to consider the small peptide size and the high quality of the fragmentation data; sequencing should have been relatively easy even if performed on the unsorted data set. However, this constitutes soundproof of principle for the MD tagging strategy, and the advantages should be more noticeable for higher molecular weight systems.
3.4.5 Chemical Labelling of Myoglobin using reductive alkylation incorporating N-terminal selective iodine-based mass defect tag

As before, the initial studies were executed using the conditions described by Chen et al. MD tags were successfully incorporated into Myo as demonstrated by LCMS analysis (Figure 3.91, A). [M+28H]^{28+} to [M+11H]^{11+} ions were observed. A specific charge state of oxidised Myo ([M+19H]^{19+}) was then isolated and studied alongside mathematical simulations of the unmodified and modified protein (Figure 3.91, B, Table Appendix 14). For Myo, the desired conjugate is singly modified. The relative yield was 26 % for singly modified Myo.
Figure 3.91: A) Mass spectra of Myo. Top: Controls showing the charge state distributions of the unmodified protein. Bottom: Spectra of the crude reaction mixture with a combination of unmodified protein and species with higher mass. B) Observed isotopic distributions of Myo overlaid with simulated isotopic distributions for the unmodified (red) and singly-modified (IMDT) (blue) protein as represented by coloured rings.
3.4.6 Time-course experiments of reductive alkylation incorporating N-terminal selective iodine-based mass defect tag at varying temperatures of Myoglobin.

Figure 3.92: Relative yields from Myo within a time course experiment at varying temperatures.
As previously, the effect of temperature on reaction progression and yield was investigated by performing a series of time-course experiments at several temperatures (17, 37 and 47 °C) (Figure 3.92).

Figure 3.92 reveals that there has been a gradual decline in the unmodified protein as the singly modified species were formed. Specifically, the yield rose from 17 °C to 37 °C but dropped going from 37 °C to 47 °C, probably due to lower thermal stability. The highest yield of singly conjugated species (33 %) was reached within 6 hours at 37 °C, slightly higher than the 31 % yield after 2 hours at the equivalent temperature.

3.4.7 Evaluating Selectivity of the Chemical Labelling of Myoglobin using a reductive alkylation incorporating N-terminal selective iodine-based mass defect tag

As before, to utilise the IMDT for de novo sequence purposes, we had to verify that the chemical reaction is selective to the N-terminal amines. Again, top-down fragmentation was employed and ECD fragmentation was conducted on the isolated [M+20H]^{20+} peak of the singly modified Myo conjugate, accomplishing almost full sequence coverage of the chain (Figure 3.93). Figure 3.93 shows the sequence coverage observed for top-down ECD of IMDT modified myoglobin. In detail, in total, 365 ions were detected where 243 of them were successfully assigned. The most abundant peaks in the ECD spectrum are an extended sequence of tagged c-ion fragments (115 ions) and untagged z-ions (142 ions) fragments (Table Appendix 15). Both of which imply the conjugation was successfully incorporated on the N-terminus. However, a series of untagged c-ions (94) was also observed, suggesting that either the label had dissociated during fragmentation or a significant amount of non-selective modification had taken place.
Figure 3.93: ECD fragmentation spectrum of Myo modified with iodobenzaldehyde. The charge state distribution of Myo, 20+ charge state isolated (middle) and with ECD (bottom). The bottom spectrum shows a magnified region of the fragmentation spectrum with all assigned signals labelled.
Table 3.7: Identified fragments sorted according to ion type and whether they are consistent with a tag on the N-terminus or unselective. The ratios indicate selectivity based on the summed intensity of fragments suggesting N-terminal labelling versus fragments suggesting unselective labelling.

<table>
<thead>
<tr>
<th>Fragmentation method</th>
<th>Tag on N-term</th>
<th>Unselective Labelling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>b/c</td>
<td>y/z</td>
</tr>
<tr>
<td>ECD</td>
<td>115</td>
<td>142</td>
</tr>
</tbody>
</table>

As illustrated in Figure 3.94, thirteen tagged $z$-ions were also assigned, suggesting that most probably some nonspecific modification has taken place probably on a Lysine residue since it has primary amine.

Comparing the ion intensities for these series of ions (Figure 3.95), it is clear that the $z^*$ ion series is of significantly lower intensity. Furthermore, contrary to expectations, we noticed that $c$ and $c^*$ ions have almost the same intensity. Based on the available evidence, it seems fair to suggest that the reaction appears to be more selective towards the $N$-termini over the lysines. However, there is a strong presence of undesired products with modified lysines that complicate de novo sequencing. Therefore, this reaction is not ideal since it generates a lot of undesired byproducts. Thus, there is the need to investigate other chemical reactions that would be more selective towards the $N$-terminal amines.
Figure 3.94: Sequence of the Myo with an MD tag on the N-terminus (top) or the C-terminus (bottom) matched against the ECD fragmentation data for Myo modified under initial conditions. Red hooks indicate identified fragments.
Figure 3.95: Plot showing the intensities of all fragments assigned from ECD fragmentation data for Myo modified under initial conditions. The displayed ratios highlight the relatively low intensity of the fragments suggesting tagging of an internal Lysine compared to their counterparts which suggest an N-terminal tag.

3.4.8 Incorporation of a Bromo-substituted 2-pyridinecarboxaldehyde Mass Defect Tag via Imidazolidinone Formation on Ubiquitin

Since the selective labelling approach proposed by Chen$^{62}$, was not proven to be sufficiently selective, we decided to adapt a reductive alkylation strategy recently outlined by MacDonald$^{259}$. We used this to incorporate a bromine-based mass defect tag (BrMDT; delta mass: C$_4$H$_2$NBr, monoisotopic mass: 142.9371) to the N-terminus of proteins.$^{259}$ Initially, this strategy was developed to modify the N-terminus of native proteins in 2015 through imidazolidinone formation (Figure 3.96, a). The authors studied various buffers and pH conditions to achieve the highest selectivity employing benzaldehyde for conjugation. They achieved >99:1 selectivity for the protein RNase A, leading to an incredible conversion (≥90 %). Furthermore, the group illustrated that the label could be conjugated into various proteins, including aldolase, lysozyme, bovine serum albumin, with reasonable conversion rates.$^{259}$
Figure 3.96: Reaction scheme for the N-terminal selective incorporation of a) of 2PCA-biotin reagent and b) 5-bromo-substituted 2-pyridinecarboxaldehyde into proteins.

This recent report achieved selective labelling of a range of proteins by using pyridinecarboxaldehyde derivatives in good yield. However, in our initial experiments, using 5-bromo-2-pyridinecarboxaldehyde and the published reaction conditions (Figure 3.96, b), only low covalent labelling of Ubiquitin was observed. LCMS analysis revealed that the yield of the reaction was less than 10%. For the BrMD to be valid for the intended application of de novo sequencing, the yield of the tagged conjugate must be high enough to be amenable to fragmentation. Therefore, the reaction conditions were then optimised in an attempt to increase the reaction yield.

We postulated that polypeptides' three-dimensional structure might shield the N-terminal amine and inhibit efficient interaction with the bulky Bromo-substituted pyridinecarboxaldehyde moiety. Therefore, we investigated if denaturing the protein prior to derivatisation would increase the reaction yield. A series of experiments, including the denaturing agent guanidine-HCl added to the reaction mixture, was performed (Figure 3.97).
Using this approach, the reaction yield rose dramatically to greater than 50% (Figure 3.97). A specific charge state of modified Ubiq ([M+8H]^{8+}) was then isolated and studied alongside mathematical simulation of the modified protein (Figure 3.97). These primary findings confirmed that Ubiq has successfully been modified. Furthermore, it was encouraging to notice no double or triple modified species signs, indicating that the reaction is likely selective to the N-termini.

**Figure 3.97:** a) The mass spectra of Ubiquitin (Ubiq) treated with 5-Bromo-substituted 2-pyridinecarboxaldehyde and denaturing agent. Observed isotopic distributions of Ubiq overlaid with simulated isotopic distribution for the b) unmodified (red) and c) singly-modified (BrMDT) (blue) protein as represented by coloured rings.
3.4.9 Comparison of top-down ECD fragmentation of Unmodified and BrMDT-modified Ubiquitin.

Having demonstrated that modification of Ubiquitin could be achieved in yields suitable for top-down fragmentation, it was then necessary to determine if N-terminal modification had any influence on top-down ECD fragmentation efficiency.

A series of ECD experiments were performed on the same charge states of both unmodified and modified Ubiq using similar conditions. The analysis carried out in this report uses a 12 T FT-ICR SolariX (Bruker) equipped with ECD. The electrons are produced using a heated filament source situated behind the ICR cell. Electrons are introduced to the trapped protein ions within the ICR cell, so all resultant fragments ions, in theory, should be detectable. The fragmentation spectra achieved by applying ECD fragmentation to the \([\text{M+11H}]^{11+}\) charge state of unmodified (Table Appendix 17) and BrMDT-modified (Table Appendix 18) Ubiq are shown in Figure 3.98.

As highlighted in Figure 3.98, the fragmentation pattern of unmodified (Figure 3.98, top) and modified (Figure 3.98, bottom) have surprisingly different fragmentation spectra. It is clear that, for unmodified Ubiq, the most abundant ions in the ECD spectrum are several \(c\)- and \(z\)-ions (37 ions) fragments throughout the protein sequence, thus providing a high coverage of the protein amino acid sequence (Figure 3.99).

In contrast, the highest abundance peaks observed upon ECD of BrMDT-modified Ubiq are precursor \([\text{M+11H}]^{11+}\) ions and a series of ions consistent with electron capture with no dissociation (ECnoD) (e.g. \([\text{M+11H}]^{10+}\)). Although several \(z\)-ions are also evident, it was surprising that no \(c\) fragment ions were observed under these conditions. This somewhat unexpected result may be due to the presence of the tag. These findings, therefore, provide confirmatory evidence that the modification presence affects the generations of \(c\) ions and thus drops the sequence coverage of the polypeptide remarkably.
Figure 3.98: ECD fragmentation spectra for Ubiq. Top-down ECD fragmentation of standard (top) and BrMDT-modified (bottom) protein. From the graph, it can be observed that higher fragmentation coverage is achieved for standard Ubiq.
Figure 3.99: The fragment map displays fragment ions resulting from ECD fragmentation of standard (top) and modified (bottom) Ubiquitin. Top-down FT ICR MS suggested that the N-termini modification leads to lower sequence amino acid coverage of the protein. Specifically, ECD fragmentation of the standard and modified led to 35 and 87% sequence coverage respectively within a 5 ppm error.

Figure 3.100: The number of c and z ions produced by standard and Modified Ubiquitin under the same conditions.
Table 3.8: Different tested ECD conditions. The impact of ECD bias (V), collisional voltage (V), and the gas flow control (%). The available evidence seems to suggest that the optimum conditions are the ones described by experiment 6.

<table>
<thead>
<tr>
<th>Experiment Number</th>
<th>Collisional Voltage (V)</th>
<th>ECD Bias (V)</th>
<th>Gas Control Flow (%)</th>
<th>Sequence Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>1.5</td>
<td>30</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>1.5</td>
<td>30</td>
<td>49</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>1.5</td>
<td>30</td>
<td>24</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>1.8</td>
<td>30</td>
<td>48</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>2.5</td>
<td>30</td>
<td>44</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>1.5</td>
<td>10</td>
<td>93</td>
</tr>
</tbody>
</table>

The data generated by top-down fragmentation of standard and modified Ubiq are reported in Figure 3.99. The evidence that emerges from comparing the number of hits of standard and modified Ubiq (Figure 3.100) suggests that the presence of modification leads to a significant drop in the observed fragment ions. Specifically, the ECD of the standard protein leads to extensive fragmentation throughout the protein sequence, while the ECD of the modified protein results in fragments predominantly in the C-terminal region. However, these results were not very encouraging since for the BrMDT to be valid for the intended application of de novo sequencing, the fragmentation of the tagged conjugate must be extended throughout the protein sequence. Therefore, optimisation of the ECD conditions was further explored in order to increase sequence coverage.

ECD bias and collisional voltage were chosen as the parameters that need to be set by the user to optimise the ECD experiments. The ECD bias sets the potential difference between the cathode and the effective potential in the centre of the ICR cell, while the collisional voltage sets the potential to activate the ions. Several ECD bias and collisional voltage levels were studied with the gas control flow effect, and those parameters combined with the obtained c and z ions are summarised in the following table (Table 3.8).

Initial studies were carried out at the same charge state ([M+12H]^{12+}), applying 3 V (Table Appendix 19), 5 V (Table Appendix 20), and 10 V (Table Appendix 21) collisional voltage, ECD bias of 1.5 (Table Appendix 20), 2.5 (Table Appendix 22) and 1.8 V (Table Appendix 23) and gas control flow of 30 % (Table Appendix 20) and 10 % (Table Appendix 24). The data generated by the top-down fragmentation of various conditions are reported in Error! Reference source not found. As mentioned above, the initial conditions (3 V) have led to 21
% sequence coverage (122 ions; Figure 3.101, experiment 1). As the collisional rose to 5 V, a significant climb to the number of c* (50) and z (47) ions was observed, corresponding to 49% sequence coverage; thus, providing greater coverage of the primary sequence than before (Figure 3.101, experiment 2). Thus, it can be reasonably assumed that the application of 5 V has activated the precursor ion to cause further fragmentation. However, setting the collisional voltage at 10 V led to a significant drop compared to the initial conditions. As illustrated in Figure 3.101 (experiment 3), setting the collisional voltage at 10 V has led to the disappearance of c ions and the generation of primarily b* (10) and y (15) ions, presumably through vibrational fragmentation dominating the fragmentation process. Therefore, considering these findings, a collisional voltage of 5 V was employed along with ECD in future experiments.

Figure 3.101: The sequence coverage obtained by employing 3 V, 5 V, and 10 V ECD bias of 1.5, 2.5, and 1.8 and gas control flow of 30 % and 10 %. This study indicates that the optimum collisional voltage, ECD bias, and gas control flow are 5 V, 1.5 V, and 10 %, respectively. Thus, the highest number of c* (58) and z ions (191) are generated at the optimum conditions.

After determining the need to apply a 5 V collisional voltage, the next step was determining the ECD bias’s optimum conditions. The data generated by the top-down fragmentation of various conditions are reported in Figure 3.101. As described in Figure 3.101 (Experiment 2, 4, 5), applying 2.5 V and 1.8 V has led to the generation of a lower number of c ions (33 and
19 ions, respectively) compared to the initial conditions where 50 c ions were produced. Furthermore, it is worth mentioning that neither the number of z ions nor the sequence coverage was significantly altered. Thus, the data gathered in the pilot study implied that the ECD bias has no meaningful impact on the fragmentation. Therefore, further investigation was not required.

Finally, as described in Figure 3.101 (experiment 6), setting up the gas flow at 10 % has led to a dramatic rise in forming a high number of c (55) and z ions (120). Furthermore, it is worth mentioning that almost full sequence coverage was significantly obtained (93 %). Thus, the data yielded by this study provide convincing evidence that the flow of gas has a meaningful impact on fragmentation. Therefore, the gas flow was set to 10 %.

3.4.10 Evaluating Selectivity of the Chemical Labelling of Ubiquitin incorporating a Bromo-substituted 2-pyridinecarboxaldehyde Mass Defect Tag

Since our main aim was to employ the BrMDT for de novo sequence purposes, the following step was to the chemical reaction's selectivity towards the N-terminal amines. To do that, we employed a top-down workflow as before. Thus, ECD fragmentation was conducted on the isolated [M+11H]11+ peak of the singly modified Ubiq conjugate with a pulse length of 25 ms and 5 V collisional voltage (Table Appendix 24), accomplishing almost full sequence coverage of the protein (Figure 3.102). In total, 887 ions were detected, and 216 of them were successfully assigned.

Figure 3.102 shows that the most abundant peaks in the ECD spectrum are several tagged c- and (28 ions) and untagged z-ions (120 ions) fragments. Both of which imply the conjugation was incorporated on the N-terminus. However, as illustrated in Figure 3.103, several untagged c-ions (8) were observed, suggesting that either the label had dissociated during fragmentation or placed on an internal Lysine residue. Furthermore, the absence of tagged z-ions implied that an internal Lysine was not modified (Table 3.9). Combining those findings provides enough support for the conceptual premise that the reaction is selective to the N-termini.
**Figure 3.102: ECD fragmentation spectrum of Ubiq modified with 5-Bromo-substituted 2-pyridinecarboxaldehyde.** The charge state distribution of Ubiq, 11+ charge state isolated (middle) and with ECD (bottom). The bottom spectrum shows a magnified region of the fragmentation spectrum with all assigned signals labelled.
Figure 3.103: Sequence of the Ubiq with an MD tag on the N-terminus (top) or C-terminus (bottom) matched against the ECD fragmentation data for Ubiq modified. Red hooks indicate identified fragments.

Table 3.9: Identified fragments sorted according to ion type and consistent with a tag on the N-terminus (selective) or an internal lysine (unselective). The ratios show selectivity based on the summed intensity of fragments suggesting N-terminal labelling versus fragments suggesting unselective labelling.

<table>
<thead>
<tr>
<th>Fragmentation method</th>
<th>Tag on N-term</th>
<th>Tag on Lysine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>b/c</td>
<td>y/z</td>
</tr>
<tr>
<td>ECD</td>
<td>29</td>
<td>65</td>
</tr>
<tr>
<td>CID</td>
<td>8</td>
<td>63</td>
</tr>
</tbody>
</table>
3.4.11 *De Novo* Sequencing of Ubiquitin based Imidazolidinone Formation incorporating *N*-terminal selective bromine-based mass defect tag

As stated in the Introduction, we initiated this study to demonstrate MD analysis in a *de novo* sequencing context; Ubiq was sequenced using a simple *de novo* top-down (ECD) sequencing methodology after selectively modified on the *N*-termini, assuming no prior knowledge.

ECD fragmentation was conducted on the isolated [M+11H]^{11+} and [M+8H]^{8+} peaks of the singly modified Ubiquitin conjugate with a pulse length of 25 ms and applying 5 V collisional voltage (Figure 3.104, A, Table Appendix 24). The, an MD plot was derived employing average scaling correcting factor for the fragment ions from Ubiq ECD fragmentation dataset Figure 3.104, B) as previously explained in 3.1.5. Utilising the approach to the fragment ions can be used to evaluate the method's performance in differentiating between *N-* and *C-* terminal fragment ions. The figure below clearly demonstrates that, in contrast to the CID analysis, the fragments formed two well-defined series with the modified *c*-ions in the top row corresponding to the *N*-terminal ions and the unmodified *z*-ions corresponding to the bottom row *C*-terminal ions (Table Appendix 26, Table Appendix 27).

After successfully separating *N-* and *C-* termini fragment ions, the two separate data sets were sequenced individually. By employing a 5 ppm error threshold, it was feasible to define most of the candidate primary sequence since most of the amino acids were successfully identified. The high sequence coverage was achieved due to the dataset's high resolution and mass accuracy; it was viable to sequence sections of up to two amino acids accurately. Therefore, residues 15 and 16 were rightly assigned as LE/EL, yet the correct order could not be defined as there was no internal fragment between the two residues. The simplified *de novo* sequence process described here led to 91% of the primary sequence being determined (Figure 3.105).
Figure 3.104: A) ECD fragmentation spectrum of Ubiq modified 5-Bromo-substituted 2-pyridinecarboxaldehyde. B) Mass defect plot derived employing average scaling correcting factor for the assigned fragment ions from Ubiq ECD fragmentation dataset.
Figure 3.105: The protein sequence of Ubiq determined de novo combing an N-termini MD tag and ECD fragmentation. Ambiguous sequences are highlighted in Red and unknown amino acid acids with green. The simplified de novo sequence process described here led to 91 % of the primary sequence being determined.

3.4.12 Incorporation of a Bromo-substituted 2-pyridinecarboxaldehyde Mass Defect Tag via Imidazolidinone Formation on Myoglobin

As established before, the initial studies were executed using the conditions described by the conditions recommended by Francis et al. MD tags were successfully incorporated into Myo as demonstrated by LCMS analysis (Figure 3.106, A, Table Appendix 28). [M+19H]^{19+} to [M+12H]^{12+} ions were observed. A specific charge state of Myo ([M+19H]^{19+}) was then isolated and studied alongside mathematical simulations of the unmodified and modified protein (Figure 3.106, B). For Myo, the desired conjugate is singly modified. The relative yield was 70 % for singly modified Myo. These primary findings support previous research confirming that the tag has successfully been incorporated into Myo. Furthermore, it was encouraging to notice no signs of double or triple modified species, indicating that the reaction is probably selective to the N-termini.
3.4.13 Evaluating Selectivity of the Chemical Labelling of Myoglobin incorporating a Bromo-substituted 2-pyridinecarboxaldehyde Mass Defect Tag

As before, to implement the BrMDT for protein studies, we had to confirm that the chemical reaction is selective to the N-termini. As earlier, we chose a top-down approach to verify that. ECD fragmentation was conducted on the isolated [M+20H]\(^{20+}\) peak of the singly modified Myo conjugate, accomplishing high sequence coverage of the protein (Figure 3.107, Table...
Appendix 29). In total, 445 ions were detected where the 163 of them were successfully assigned.

![ECD fragmentation spectrum of Myo modified with BrMDT](image)

**Figure 3.107: ECD fragmentation spectrum of Myo modified with BrMDT.** The bottom spectrum shows a magnified region of the fragmentation spectrum with all assigned signals labelled.

Figure 3.107 indicates that the most abundant peaks in the ECD spectrum are several tagged c- and (37 ions) and untagged z-ions (44 ions) fragments. Both of which imply the conjugation was incorporated on the N-terminus. In addition, there has been no sign of modified C-termini. However, it is worth mentioning that there is one untagged c-ion,
suggesting that the label had dissociated during fragmentation. Therefore it seems that the reaction is selective to the \(N\)-termini. As illustrated in Figure 3.108, the obtained sequence coverage was 54\%, in which 37\% of the identified ions were assigned within a 5 ppm error.

**Tag on N-terminus**

```
N  G  L  S  D  G  E  W  Q  Q  V  L  N  V  W  G  K  V  E  A  D  I  A  G  H  G  25
Q  E  V  L  I  R  L  F  T  G  H  P  E  T  L  E  K  F  D  K  F  K  H  L  K  50
T  E  A  E  M  K  A  S  E  D  L  K  K  H  G  T  V  V  L  T  A  L  G  G  I  75
L  K  K  K  G  H  H  E  A  E  L  K  P  L  A  Q  S  H  A  T  K  H  K  I  P  100
I  K  Y  L  E  F  I  S  D  A  I  I  H  V  L  H  S  K  H  P  G  D  F  G  A  125
D  A  Q  G  A  M  T  K  A  L  E  L  F  R  N  D  I  A  A  K  Y  K  E  L  G  150
F  Q  G  C
```

**Tag on C-terminus**

```
N  G  L  S  D  G  E  W  Q  Q  V  L  N  V  W  G  K  V  E  A  D  I  A  G  H  G  25
Q  E  V  L  I  R  L  F  T  G  H  P  E  T  L  E  K  F  D  K  F  K  H  L  K  50
T  E  A  E  M  K  A  S  E  D  L  K  K  H  G  T  V  V  L  T  A  L  G  G  I  75
L  K  K  K  G  H  H  E  A  E  L  K  P  L  A  Q  S  H  A  T  K  H  K  I  P  100
I  K  Y  L  E  F  I  S  D  A  I  I  H  V  L  H  S  K  H  P  G  D  F  G  A  125
D  A  Q  G  A  M  T  K  A  L  E  L  F  R  N  D  I  A  A  K  Y  K  E  L  G  150
F  Q  G  C
```

**Figure 3.108:** Sequence of the Myo with an MD tag on the N-terminus (top) or C-terminus (bottom) matched against the ECD fragmentation data for Myo modified under initial conditions. Identified fragments are indicated by red hooks.

### 3.4.14 De Novo Sequencing of Myoglobin based Imidazolidinone

Formation incorporating \(N\)-terminal selective bromine-based mass defect tag

As stated in the Introduction, we commenced this study to implement MD analysis in a *de novo* sequencing context; Myo was sequenced employing a simple *de novo* top-down ECD
sequencing system after being selectively modified \(N\)-termini, assuming no prior knowledge. ECD fragmentation was conducted on various isolated peaks of the singly modified Myo conjugate. One specific charge was selected to be plotted (Figure 3.109, A, Table Appendix 30) An MD plot was derived employing averagine scaling correcting factor for all the fragment ions from Myo ECD fragmentation datasets (Figure 3.109, B) as previously explained in the Mass Defect plots using averagine scaling analysis. Utilising the approach to the fragment ions can differentiate between \(N\)- and \(C\)-terminal fragment ions (Table Appendix 31). The figure below clearly illustrates that the fragments formed two well-defined series with the modified \(c\)-ions in the top row corresponding to the \(N\)-terminal ions and the unmodified \(z\)-ions corresponding to the bottom row \(C\)-terminal ions.

After successfully separating \(N\)- and \(C\)-termini fragment ions, the two separate data sets were sequenced individually. By employing a 5 ppm error threshold, it was feasible to define 78% of the candidate primary sequence since most of the amino acids were successfully identified. The high sequence coverage was achieved due to the dataset’s high resolution and mass accuracy; it was viable to accurately sequence sections of up to two amino acids. Amino acid residues 36-37 were rightly sequenced as histidine and proline, considering ECD does not fragment on the \(N\)-termini side of prolines. The simplified \textit{de novo} sequence process described here led to 70% of the primary sequence being assigned (Figure 3.110).

Figure 3.109: A) ECD fragmentation spectrum of Myo modified with 5-Bromo-substituted 2-pyridinecarboxaldehyde. B) Mass defect plot derived employing averagine scaling correcting factor for the assigned fragment ions from Myo ECD fragmentation dataset.
Figure 3.110: The protein sequence of Myo determined de novo combining an N-termini MD tag and ECD fragmentation. Ambiguous sequences are highlighted in Red and unknown amino acid acids with an “X”. The proposed simplified de novo sequence process described here led to 70 % of the primary sequence.

3.4.15 Incorporation of a Bromo-substituted 2-pyridinecarboxaldehyde Mass Defect Tag via Reductive Alkylation on Rnase A

This section demonstrates that our MD tagging strategy is compatible with chemical cysteine reduction methods required for top-down sequencing of Cys rich proteins. As before, the initial studies were executed using the conditions recommended by Francis et al. MD tags were successfully incorporated into RNase A as demonstrated by LCMS analysis (Figure 3.111, A, Table Appendix 31). [M+14H]$^{14+}$ to [M+7H]$^{7+}$ ions were observed. A specific charge state of oxidised RNase A ([M+8H]$^{8+}$) was then isolated and studied alongside mathematical simulations of the unmodified and modified protein (Figure 3.111, B). For oxidised RNase A, the desired conjugate is singly modified, and the relative yield was 40 % for singly oxidised RNase A. These primary findings support previous results confirming that the tag has successfully been incorporated into oxidised RNase A. Furthermore, it was encouraging to notice no double or triple modified species signs, indicating that the reaction is probably selective to the N-termini.
The question under scrutiny in this section is the best protocol to modify the N-terminus and reduce and alkylate the cysteines. To solve this issue, we developed two parallel workflows. The first approach includes reduction and alkylation of cysteines followed by modification of the N-termini. In the second one, the N-termini was modified, and then the cysteines were reduced and alkylated. Figure 3.112 illustrates the two workflows.

**Figure 3.112:** A) Mass spectra of oxidised RNase A. Top: Controls showing the charge state distributions of the unmodified protein. Bottom: Spectra of the crude reaction mixture with a combination of unmodified protein and species with higher mass. B) Observed isotopic distributions of oxidised Rnase A overlaid with simulated isotopic distributions for the a) unmodified (red) and b) singly-(blue) protein as represented by coloured rings.

Both processes were run simultaneously, and samples from each step were collected and analysed by MS to ensure that the process proceeded successfully. As illustrated in the Figure 3.113, protocol A has lead to the multi conjugation of Rnase A while by applying protocol B,
RNase was successfully reduced and alkylated by NEM and conjugated by a single modification. However, when we have tried to modify the \(N\)-terminus, several side products were generated. A buffer exchange step was added before the \(N\)-terminus labelling reaction to address this issue—however, no significant improvement in the spectrum quality was observed. This unexpected outcome could be due to a replacement reaction where BrMDT subsidised the NEM.

A similar approach is used in protocol B. The vital difference is that the \(N\)-terminus labelling occurs first compared to protocol A. The oxidised RNase was successfully modified, as established in section. However, as the tag is susceptible to rapid reaction with the nucleophilic cysteine thiol produced by chemical reduction, adding a buffer exchange step after \(N\)-terminal labelling before the disulfide bonds reduction was necessary. Using this approach, the cysteines were successfully modified with NEM. As illustrated in Figure 3.113, there are no signs of multiple modified species before the rise of new peaks corresponding to cysteine alkylation. A specific charge state of NEM and \(N_{\text{term}}\)-modified RNase (\([M+14H]^{14+}\)) was then isolated and studied alongside mathematical simulations of the modified protein (Figure 3.114). As it can be seen, the mathematical simulation fits quite well the isotopic distribution confirming that the RNase was labelled successfully.
Figure 3.112: The two protocols studied to modify the N-terminus and reduce and alkylate the cysteines selectively. In the A protocol, on the first step, cysteines are reduced, and then alkylated and on the final step the N-terminus is modified. On the B protocol, the N-terminus is first modified, and then cysteines are reduced and modified.
Figure 3.113: Top: Mass spectra of Modified RNase when protocol A is followed. In the A protocol, on the first step, cysteines are reduced, alkylated, and the N-terminus is modified on the final step. Bottom: Mass spectra of Modified RNase when protocol B is followed. On the B protocol, the N-terminus is first modified, and then cysteines are reduced and modified.

Figure 3.114: A) Mass spectra of Modified RNase A. B) Observed isotopic distributions of NEM and N-term-modified Rnase A overlaid with simulated isotopic distributions for the singly-(blue) protein as represented by coloured rings.
Evaluating Selectivity of the Chemical Labelling of RNase A incorporating a Bromo-substituted 2-pyridinecarboxaldehyde Mass Defect Tag

As previously explained, the following step was to study whether the BrMDT was selectively attached to the N-termini; thus, ECD fragmentation was conducted on the isolated $[M+15H]^{15+}$ peak of the singly modified RNase A conjugate, accomplishing high sequence coverage of the protein (Figure 3.115, Table Appendix 32). In detail, in total, 297 ions were detected where the 92 of them were successfully assigned.

![Figure 3.115: ECD fragmentation spectrum of RNase A modified with 5-Bromo-substituted 2-pyridinecarboxaldehyde. The bottom spectrum shows a magnified region of the fragmentation spectrum with all assigned signals labelled.](image)

**Figure 3.115:** ECD fragmentation spectrum of RNase A modified with 5-Bromo-substituted 2-pyridinecarboxaldehyde. The bottom spectrum shows a magnified region of the fragmentation spectrum with all assigned signals labelled.
Figure 3.115 indicates that the most abundant peaks in the ECD spectrum are several tagged c- and (27 ions) and untagged z-ions (37 ions) fragments. Both of which imply the conjugation was incorporated on the N-terminus. In addition, there has been no sign of modified C-termini. However, it is worth mentioning that a couple of untagged c-ion (18) suggests that either the label had dissociated during fragmentation (Table 3.10). This implies that careful control of fragmentation conditions must be considered to overcome fragment inhibition by tag. As illustrated in Figure 3.116, the obtained sequence coverage was 49 %, in which 31 % of the identified ions were assigned within a 5 ppm error.

**Tag on N-terminus**

![Sequence of the RNase A with an MD tag on the N-terminus (top) or C-terminus (bottom) matched against the ECD fragmentation data for Myo modified under initial conditions. Identified fragments are indicated by red hooks.](image)

**Tag on C-terminus**

![Sequence of the RNase A with an MD tag on the N-terminus (top) or C-terminus (bottom) matched against the ECD fragmentation data for Myo modified under initial conditions. Identified fragments are indicated by red hooks.](image)
Table 3.10: Identified fragments sorted according to ion type and consistent with a tag on the N-terminus or an internal lysine. The ratios indicate selectivity based on the summed intensity of fragments suggesting N-terminal labelling versus fragments suggesting unselective labelling.

<table>
<thead>
<tr>
<th>Fragmentation method</th>
<th>Tag on N-term ( b/c )</th>
<th>Tag on N-term ( y/z )</th>
<th>Tag on Lysine ( b/c )</th>
<th>Tag on Lysine ( y/z )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECD</td>
<td>27</td>
<td>37</td>
<td>18</td>
<td>0</td>
</tr>
</tbody>
</table>

3.4.17 De Novo Sequencing of RNase A based on ECD fragmentation, and alkylation described by MacDonald incorporating N-terminal selective bromine-based mass defect tag

As stated in the Introduction, we started this research to validate MD analysis in a de novo sequencing framework; RNase A was sequenced utilizing a simple de novo top-down ECD sequencing methodology after being selectively modified N-termini, assuming no prior knowledge. ECD fragmentation was conducted on various isolated peaks of the singly modified RNase A conjugate. To visualise this progress, one specific charge was selected to be plotted (Figure 3.117, A).

An MD plot was derived employing averagine scaling correcting factor for all the fragment ions from alkylated RNase A ECD fragmentation datasets (Figure 3.117, B) as previously explained in the Mass Defect plots using averagine scaling analysis (Table Appendix 33). Utilising the approach to the fragment ions can differentiate between N- and C-terminal fragment ions. The figure below clearly illustrates that the fragments formed three series of ions, probably due to tag cleavage during ECD. However, the top row ions were assigned as \( c \)-ions while the bottom series was identified as \( z \) and miscleavages and the middle ones as unknowns.

After successfully separating N- and C-termini fragment ions, the two separate data sets were sequenced individually. Amino acid residues 113-114 were sequenced as asparagine and proline, considering ECD does not fragment on the N-termini side of prolines. By employing a
Figure 3.117: A) ECD fragmentation spectrum of RNase A modified with 5-Bromo-substituted 2-pyridinecarboxaldehyde. B) Mass defect plot derived employing averaginescaling correcting factor for the assigned fragment ions from RNase A ECD fragmentation dataset.
5 ppm error threshold, it was feasible to define 79% of the candidate primary sequence since most of the amino acids were successfully identified (Figure 3.118).

**Figure 3.118:** The protein sequence of RNase A determined de novo combining an N-termini MD tag and ECD fragmentation. Ambiguous sequences are highlighted in Red and unknown amino acid acids with green. The proposed simplified de novo sequence process described here led to 79% of the primary sequence.

### 3.5 Conclusions

Our work has led us to conclude that MD tagging has potential uses in top-down de novo sequencing. However, extreme caution must be taken to verify that the reaction is extremely specific N-terminal labelling. Otherwise, as in the reductive alkylation approach, significant by-products will be generated, complicating the analysis. Since the selective labelling approach proposed by Chen, was not proven to be sufficiently selective, we decided to adapt the imidazolidinone formation defined by the MacDonald. A 5-Bromo pyridinecarboxaldehyde MD label was successfully conjugated into Ubiq, Myo and RNase A. The use of the MD tag for de novo protein sequencing was demonstrated with an averagine-scaled mass defect plot for the Ins B-chain, which allowed for the precise and easy distinction of N- and C-terminal fragments. After MD analysis Ubiq, Myo and RNase A’s primary sequence could be sequenced to 100%, 78% and 79%, respectively. However, our study demonstrates that it is crucial to carefully control fragmentation conditions to overcome fragment inhibition by tag since unwanted cleavage of tag during analysis leads to complicating series of z ions, making ion series recognition more challenging.
These results demonstrate the potential of applying a mass defect tagging approach for de novo sequencing. However, there is abundant room for further progress in determining the optimum conditions to achieve high conversion and selectivity when the process is applied to larger molecular weight systems and, eventually, to middle-down preparations of antibodies. Furthermore, for the system to be valuable in the characterisation of PTMs by top-down MS, one would also need to study the effect of the reductive alkylation reaction on PTMs. Lastly, in future investigations, it might be possible to use automated software to analyse the data.
4. Conclusions

In recent years, top-down and middle-down MS of proteins has gained momentum as a potential tool for studying protein structure. In this thesis, we have attempted to comprehensively investigate how the particular top-down technique of ECD can be used in this manner. ECD was chosen as it is convenient for FT-ICR MS and has been studied previously due to its unique capacity to highly exclusively fragment the covalent peptide backbone while leaving noncovalent interactions intact.

Firstly, an efficient workflow for both intra- and inter-disulfide bridge (-S-S-) assignment was developed, utilising a combination of pepsin and trypsin proteolysis and the usage of dual fragmentation of ECD and CID without the need of a chromatographic system. For this research, BSA as a target was chosen due to its well-known structure. The protein was digested, by trypsin and pepsin, under nonreducing conditions to generate several bridged peptides. These bridged peptides were then reduced and injected into the FT ICR MS to be separated based on their synchrotron frequency. Afterwards, the collected data were run against a database to identify the sequence. Afterwards, the same process was repeated without reducing the disulfide bonds before injection to the mass spectrometer analyzer. Separated ions are online subjected to CID and ECD to generate extensive sequence information and locate the disulfide bonds spectra of each species. We have obtained comprehensive results demonstrating that it is possible to identify the sequence and assign the disulfide connectivity by employing our approach.

The second project was to increase fragment ion assignment efficiency in top-down mass spectra and, therefore, maximise the protein sequence coverage obtained by utilising N-terminal selective mass derivatisation tags into proteins.\textsuperscript{258,259} Such a tag aids classification between \textit{N}- and \textit{C}-terminal fragments from top-down MS data, thus simplifying fragment assignment. For this to be achieved, considerable attention must be paid to confirm that the reaction is extremely specific towards the \textit{N}-terminus. Otherwise, as in the reductive alkylation proposed by Chen\textsuperscript{62}, a lot of by-products will be generated, which will intricate the data analysis. Since the selective labelling approach proposed by Chen\textsuperscript{62}, was not proven to be sufficiently selective, we decided to modify the imidazolidinone formation outlined by MacDonald\textsuperscript{259}. A 5-Bromo pyridinecarboxaldehyde MD label was effectively incorporated into Ubiq, Myo and RNase A. The use of the MD tag for \textit{de novo} protein sequencing was demonstrated with an averagine-scaled mass defect plot for the Ins B-chain, which allowed for the precise and easy distinction of \textit{N}- and \textit{C}-terminal fragments. After MD analysis Ubiq,
Myo and RNase A's primary sequence could be sequenced to 100%, 78% and 79%, respectively. However, it is crucial to carefully control the fragmentation conditions to overcome fragment inhibition by tag since unwanted cleavage of tag during analysis leads to complicating series of $z$ ions, making ion series recognition more difficult.
Bibliography


M. Field and H. Ionization, .


A. G. Marshall and F. R. Verdun, *Fourier transforms in NMR, optical, and mass*


108 Y. Qi and D. A. Volmer, DOI:10.1002/mas.21482.


112 L. Sleno and D. A. Volmer, DOI:10.1002/jms.703.


H. Jiang, S.-L. Wu, B. L. Karger and W. S. Hancock, DOI:10.1021/bp.72.


S. Lu, S.-B. Fan, B. Yang, Y.-X. Li, J.-M. Meng, L. Wu, P. Li, K. Zhang, M.-J. Zhang, Y. Fu, J. Luo, R.-X. Sun, S.-M. He and M.-Q. Dong, Nat. Methods, DOI:10.1038/nmeth.3283.


15506–15516.


266 H. Han, Y. Xia and S. A. McLuckey, *J. Proteome Res.*, 2007, **6**, 3062–3069.


Appendix

The available data can be found in the following URL: https://hdl.handle.net/10283/4084