Imperial College London

Covariance mapping spectroscopy of molecular decompositions

Taran Dexter Cyrus Driver

Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Science of the University of London and the Diploma of Imperial College.

The Blackett Laboratory Laser Consortium
QOLS Group
Department of Physics
Imperial College London
London SW7 2AZ

October 2018

Abstract

This thesis describes the development of a new type of two-dimensional mass spectrometry: partial covariance-based two dimensional mass spectrometry, or pC-2DMS. Using a commercially available linear ion trap mass spectrometer and the statistical technique of partial covariance mapping, the ability to identify fragment ions originating from the same or consecutive decomposition pathways of the same molecule is demonstrated. As will be shown, this information is beneficial to analytical and mechanistic studies of trapped biopolymers. After initial development on peptide molecules, the generality of the pC-2DMS technique has been confirmed by extension to whole protein molecules as well as DNA and RNA oligonucleotide sequences.

This work represents the first time partial covariance mapping has been used for structural analysis, the first application of a covariance mapping technique to trapped ions, the first time a covariance mapping technique has been applied to molecular species as large as molecules on the order of kilodaltons and the first demonstration of the correlation of secondary consecutive fragmentation products by a covariance mapping technique.

Within pC-2DMS, the total ion count is used as a single fluctuating parameter in the partial covariance formula to suppress spurious correlations due to scan-to-scan fluctuations in a number of different experimental parameters, which cause quasi-uniform fluctuations across all of the tandem mass spectrum.

The applications demonstrated in this thesis include the resolution of mixtures of different combinatorially modified structural isomers (shown here for structural isomers of histone H4 fragment 4–17), an orders-of-magnitude reduction in the false positive rate for matching fragment ion signals in database search algorithms, resolution of charge state for highly charged fragment ions without the need for high mass accuracy resolution of the isotopic envelope and the *in silico* separation of strongly overlapping fragment ion spectra from different co-isolated and co-fragmented molecular species.

Acknowledgements

I am immensely grateful to Leszek, Marina, Vitali and Jon for providing me with the incredible opportunity to work on this project, and their continued involvement throughout. Thank you to you all.

Declaration

I hereby certify that the material of this thesis is entirely my own work, unless otherwise cited or acknowledged within the body of the text.

Taran Dexter Cyrus Driver

Copyright Statement

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the license terms of this work.

Contents

1 Introduction **9**

1.1 Background 9

1.2 Layout of thesis 12

1.3 Author's contributions 15

2 Covariance mapping spectroscopy **16**

3 Partial covariance mapping spectroscopy **23**

4 Formal theory of covariance for noisy Poisson processes **29**

5 Mass spectrometry **39**

5.1 Electrospray ionisation 40

5.2 Linear ion trap mass spectrometry 46

5.3 Collision-induced dissociation 58

5.4 Experimental parameters for pC-2DMS measurements 66

6 Foundations of pC-2DMS **69**

6.1 Total ion count partial covariance 69

6.2 pC-2DMS correlation score 77

7 Properties of pC-2DMS **82**

7.1 Absolute value of pC-2DMS correlation score 83

7.2 Event rate 86

7.3 Number of scans 89

7.4 Abundance of correlated fragment ions 91
7.5 Generality of technique 92

8 Formal theory of total ion count partial covariance for noisy Poisson processes 95

9 Resolution of mixtures of combinatorial isomers 109

10 Reduction in false positive rate for signal identification 120

11 Development of automated database search engine 128

12 Identification of fragmentation pathways 143

13 pC-2DMS determination of charge state 148

14 Extension to larger species 153

14.1 Restricted Hough transform 156
14.2 *In silico* deconvolution of mixtures of intact proteins by pC-2DMS . . 166

15 Two-dimensional diagnostic for fragment spectra from multiple parent ions 171

16 Extension to peptide anions 180

17 Extension to oligonucleotides 186

18 Current shortcomings of pC-2DMS 190

19 Future work 192

19.1 Femtosecond laser-induced ionisation / dissociation (fs-LID) 192
19.2 Computer vision for identification of true correlation signals 195
19.3 Transient ionic species 198

20 Synchrotron and XFEL studies of molecules 201

20.1 Synchrotron SOLEIL beamtime Barillot-20160424, Nov 2016. Measurement of isopropanol photoelectrons from a monochromatic source. 204
20.2 LCLS beamtime LR25, Nov-Dec 2017. Measurement of ultrafast hole migration in photoexcited cationic isopropanol. 208

Bibliography **219**

Appendices **246**

A Selected annotated partial covariance maps for species in Table 7.1 **247**

B Selected MS^3 scans **255**

C Permissions to reproduce figures **258**

Lay summary

Mass spectrometry is the method of choice for the structural analysis of biomolecules. It can be used to deduce information on the 'primary structure' of a biomolecule; i.e. the type, order and modification state of its chemical building blocks (which are amino acids for proteins, nucleotides for DNA, etc.). Typically, this is done by fragmenting an ensemble of biomolecular ions in the gas phase, measuring the mass-to-charge ratio of their resultant fragments[1], and using this fragment information to piece the original structure back together. The biomolecules in the ensemble can dissociate *via* different fragmentation pathways, producing different types of fragment ions to be measured and so providing complementary information on the molecule's primary structure, all of which is stitched together to give information on the molecule's make-up as a whole. This method has experienced incredible success over the past few decades but is inevitably being constantly improved. Current research efforts to improve the technique are focused down two main avenues: a) new methods to fragment the biomolecules, producing new types of biomolecular fragments and so new information, and b) improving the accuracy and precision of the mass-to-charge (m/z) measurement of the fragments, to enable more reliable identification of the fragments from their measured m/z values. The work in this thesis pursues a different avenue. By building on a statistical technique called covariance mapping, which was first developed for a very different purpose (analysing how small molecules behave when they are irradiated with short, intense laser pulses), different pairs of fragment ions which are born in exactly the same fragmentation pathway of exactly the same type of biomolecule can be identified. It is shown below that this new information (specifically the identified 'connection' between each such pair of fragments) can be used to give different, more precise, more detailed information about a biomolecule's structure (and fragmentation behaviour) than is available through any other means.

[1]Different types and modification states of the building blocks are usually distinguishable on the basis of different masses

Note on Terminology

This thesis requires some clarification on terminology, in order to remain consistent with its use of words having divergent meaning across the different disciplines spanned by the work. Herein, a mass spectrometric 'scan' will refer to a single iteration of the trapping, fragmentation and subsequent fragment ion measurement of an ensemble of parent ions. In this way, a scan is the direct equivalent of a 'shot' in an experiment using a pulsed laser, and in some cases will be used interchangeably with this term. An experimental run consisting of the collection of a multiple number of these scans, which is sometimes called a 'scan' itself, will be referred to as a 'measurement'.

The word 'deconvolve' has subtly different specific uses across the field of mass spectrometry. In some works it is used to describe the reduction of fragment ion mass spectra from a list of mass-to-charge (m/z) ratios to a list of fragment masses, whilst in others it is used to describe the separation of different fragment ion m/z values originating from different parent ions, and assignment to the relevant parent. In this thesis, the latter definition will be adopted.

Chapter 1

Introduction

1.1 Background

Mass spectrometry (MS) is the method of choice for the structural analysis of biomolecules such as proteins [1–6], nucleic acids [7–10], lipids [11] and oligosaccharides [12]. The primary aim of protein MS is to characterise a protein by establishing the sequence of its biomolecular building blocks (i.e. amino acids) and its post-translational modifications (PTMs) [13]. In order to do this, the large biomolecules under analysis are typically first cut into smaller pieces, e.g. using enzymes to obtain peptides from large proteins, and these are subsequently sent to a mass spectrometer as gas-phase molecular ions *via* soft ionisation techniques (electrospray ionisation [14] or matrix-assisted laser desorption-ionisation [15]). In the prevalent technique of tandem mass spectrometry (MS/MS), the biomolecular ions are next caused to fragment by one or more of a range of different physical mechanisms (the most common fragmentation method being collision-induced dissociation [16]), producing fragment ions whose mass-to-charge (m/z) ratios and abundances are then measured. This spectral information can be used to piece together the primary structure of the original biomolecule.

The crucial step in the protein MS workflow is the deduction of a protein's amino acid sequence and PTMs from the tandem mass spectra. This task can be accomplished using a range of approaches which rely on either matching the experimental spectra to 'theoretical' ones [17] derived from protein databases and a set of gener-

alised peptide fragmentation rules [2, 18], matching the measured spectra to experimentally generated spectral libraries [19], or performing a first-principles structural reconstruction from the measured spectra without direct reference to any database and using the known masses of different amino acids and modifications and fragmentation rules only (so-called *de novo* analysis) [20]. Whichever method is used for analysis of the acquired MS/MS spectra, typically fewer than 60% of the spectra are successfully interpreted [21–23]. Failures in spectrum-to-structure assignment are caused by a series of factors, many of which are a result of the the strong variability of peptide decomposition patterns as a function of the particular amino acid sequence, presence, type and location of PTMs, peptide length, charge state etc. [24–26]. The key canonical structurally informative sequence ions (e.g. b- and y-type ions for CID) [27] can often appear suppressed in the fragment ion spectrum at the expense of other, less predictable, fragmentation products. As well as inducing false positive identifications by virtue of being mistaken for the expected sequence ions, the structural information provided by non-canonical fragment ions is strongly limited because the number of potential fragmentation pathways available to a dissociating biopolymer is huge. This means that a) the reliable identification of arbitrary fragment ions using a single measured m/z value is highly challenging, and b) even if successfully identified, such fragments typically bear limited specificity to a particular biomolecular structure. Finally, there are some important cases where even a complete set of the key structurally informative sequence ions is insufficient to identify the correct biomolecular structure, because vital information is lost when the biomolecule under analysis is fragmented [28, 29]. One prominent example of this problem is in the analysis of co-fragmented combinatorially modified peptide isomers [30, 31].

Mainstream one-dimensional MS is striving to overcome challenges in spectral interpretation through two main approaches: (i) the development of new fragmentation methods [32–34] and (ii) improving the accuracy and resolution of the m/z measurement [23, 35]. For example, electron-based activation techniques such as electron transfer (ETD) [32] and electron capture dissociation (ECD) [33], which cleave peptide backbone bonds in a more random manner than CID, typically provide increased

sequence coverage compared to CID and thus more structurally specific MS/MS spectra. The remarkable resolving power and mass accuracy of modern Orbitrap mass analysers allows one to measure ions at sub part-per-million accuracy, considerably improving the reliability of fragment ion assignment using the measured m/z value, while ion cyclotron resonance (ICR) technology provides even greater mass accuracy and resolution [36]. However, even the most advanced forms of conventional one-dimensional (1D) mass spectrometry have inherent limits. Some central problems of biomolecular MS cannot be solved by the standard 1D approach even at complete sequence coverage and theoretical infinite mass precision. One particular example is the formidable analytical challenge of identifying co-fragmented isomeric combinatorially modified peptides which often lack, even theoretically, any isomer-specific fragments by which the presence of particular isomers can be detected [29–31]. pC-2DMS solves this problem directly from the MS/MS spectrum by using the statistical correlations between fragment ion peaks to identify isomer-specific 2D fragment signals, which cannot be obtained by any existing MS approach.

pC-2DMS is based on the physical principle of partial covariance mapping [37]. Covariance mapping spectroscopy was developed by Frasinski *et al.* as a tool for the study of mechanisms of radiation-induced fragmentation of di- and tri-atomic molecules [38]. The technique is based on measuring the covariance, $Cov(X, Y)$, between the intensities of every pair of spectral signals X and Y across a series of ionic time-of-flight spectra,

$$Cov(\mathbf{Y}, \mathbf{X}) = \langle \mathbf{YX} \rangle - \langle \mathbf{Y} \rangle \langle \mathbf{X} \rangle,$$

where angular brackets denote averaging over multiple spectra. If a pair of signals, X and Y, are characterised by positive covariance, their intensities fluctuate synchronously across different spectra and the corresponding fragments X and Y can be inferred to have originated from the same decomposition process, e.g. $Z \rightarrow X + Y$. Covariance mapping has been shown to be effective, for example, in unravelling the mechanisms of multiphoton ionisation in a number of different di- and tri-atomic

species [38–40] or in correlating photoelectron emission with the fragmentation of hydrocarbons in intense infrared laser fields [41]. An extension of covariance mapping, known as 'partial covariance mapping', was developed to single out these informative mechanistic correlations from extrinsic uninteresting ones, which result, for example, from fluctuations in laser intensity that affect all signals simultaneously in a uniform manner [42]. Partial covariance mapping has allowed researchers to study X-ray induced fragmentation at XFELs despite significant intensity fluctuations in the ionising pulse [43], including probing the decomposition mechanisms of 'hollow atoms' – unstable states of matter formed by intense X-ray free electron laser (XFEL) irradiation [44]. All the above studies were performed to elucidate the laser-induced ionisation and/or decomposition mechanisms of atoms or small molecules of known structure.

It should be briefly noted here that there exists a different implementation of 2D MS, known as 2D Fourier transform (FT) ICR [45, 46]. This fundamentally different type of 2D MS (based on the same physical principle as 2D exchange NMR spectroscopy) provides a completely different type of structural information to pC-2DMS: it is able to identify *precursor*-fragment correlations as opposed to the *fragment*-fragment correlations which are identified by pC-2DMS. *Precursor*-fragment correlations can otherwise be obtained by standard 1D MS^n measurements provided high enough precursor isolation efficiency [16] and is inherently unable to perform any of the applications demonstrated in this thesis.

1.2 Layout of thesis

This thesis is structured in the following way. First, the technologies upon which pC-2DMS is based are introduced. In Chapter 2 the technique of covariance mapping spectroscopy is described, and the more recently developed partial covariance mapping spectroscopy is introduced in Chapter 3. Chapter 4 provides a formal mathematical treatment of covariance mapping spectroscopy in the case of ensembles which behave according to noisy Poisson statistics (such as for the pC-2DMS measurements

performed in this thesis, see page 30) and in Chapter 5 the mass spectrometric hardware used to perform the pC-2DMS experiments is described.

Next, the fundamentals of pC-2DMS are introduced. In Chapter 6 the initial development of the technique is presented, where for the first time a spectral observable (here, the total ion count) is employed as a single partial covariance parameter to suppress the spurious correlations arising on a simple covariance map. In the case of pC-2DMS, these spurious correlations are a result of the scan-to-scan fluctuations in a large number of often untrackable experimental parameters. In section 6.2, the pC-2DMS correlation score is introduced. This is a statistical measure which makes use of jackknife resampling to extract low-volume true correlations between low-abundance fragment ions from amongst high volume features due to statistical noise, increasing the sensitivity and reliability of pC-2DMS. Some of the key properties of pC-2DMS are then explored in Chapter 7, including the ability to correlate two fragment ions appearing at extremely low relative abundances to produce high-scoring correlation signals. In Chapter 8, the formal theory of covariance mapping for noisy Poisson processes established in Chapter 4 is extended to formally validate the technique of total ion count partial covariance.

A series of applications of the new technology are then presented. In Chapter 9, pC-2DMS is used to resolve mixtures of combinatorial isomers which standard 1D MS is as a matter of principle unable to resolve. The correct identification of such modification states is a biologically relevant problem, e.g. for cracking the 'histone code'. In Chapter 10, numerical simulations demonstrate a dramatic orders-of-magnitude reduction in the false positive rate for matching pC-2DMS correlation signals to database-generated peptide sequences compared to 1D fragment ion signals. In Chapter 11, this specificity is exploited to produce a prototypical pC-2DMS database search engine which, even as a prototype, is able to identify the correct database sequence when its state-of-the-art 1D equivalent fails. In Chapter 12, mechanistic information is extracted from the unique fragment ion correlations provided by pC-2DMS to reconstruct the fragment ion formation tree for a set of abnormal biomolecular decomposition pathways. In Chapter 13, a novel 2D method to iden-

tify the charge state of both fragment ions and parent ions, as well as the mass of the parent ion producing a pC-2DMS correlation, is introduced. This method does not require the resolution of the isotopic envelope. In Chapter 14, this method is applied to assist in the pC-2DMS analysis of intact proteins. In combination with a restricted version of the Hough transform, it is used to unambiguously identify the charge state of complementary fragment ions up to 11+ measured in a linear ion trap. This workflow is then used to deconvolve the overlapping fragment ion spectra of two co-fragmented intact protein molecules, resulting in two separate correct pC-2DMS database search engine identifications from the same intact protein chimera spectrum. This opens the way for straightforward multiplex measurements, even of full intact protein molecules, in a linear ion trap. In Chapter 15, the phenomenon of 'mass conservation lines' is used to theoretically and experimentally demonstrate a robust method to identify chimera spectra of co-isolated protonated peptide ions under CID. The generality of pC-2DMS is then further demonstrated by extending the technique to deprotonated peptide ions (Chapter 16) and both protonated and deprotonated oligonucleotide ions (Chapter 17). In Chapter 18, a brief assessment of the current shortcomings are presented and finally in Chapter 19 some of the work already done on some additional experiments involving the pC-2DMS technique, whose completion is planned for the near future, is discussed.

Additionally, two large-scale collaborative experiments performed at international facilities which the author took part in are described in Chapter 20. These experiments were both aimed at measuring the behaviour of small photoexcited molecules in short-wavelength sources, and some preliminary results from their ongoing analysis are presented.

The work presented in this thesis has produced the following two patent applications:

- Driver, T., Frasiński, L.J., Klug, D., Marangos, J.P. and Edelson-Averbukh, M., Patent PCT/GB2017/052745 (GB1615818.0) on "Methods and Systems for Analysis of Mass Spectrometry Data". Filed 15th September 2017, published 22nd March 2018

- Driver, T., Averbukh, V., Frasiński, L.J., Ayers, R., Klug, D., Marangos, J.P., and Edelson-Averbukh, M., Patent GB1803940.4 on the application of pC-2DMS data to automated database search engines, titled "Methods and Systems for Analysis". Filed 12th March 2018

1.3 Author's contributions

The experimental work, data analysis and software development for the pC-2DMS experiments was performed by the author under the supervision of Leszek Frasiński, Marina Edelson-Averbukh and Vitali Averbukh, with the following additional contributions. The formal theory of total ion count partial covariance was developed with Bridgette Cooper and Serguei Patchkovskii. The optimised weights for the different fragment ion correlation categories in Chapter 11 were provided by Nikhil Bachhawat. Ruth Ayers contributed to the DNA signal annotation in Fig. 17.1 and the RNA experiment in Fig. 17.2, for which she performed all annotations. Ruth also produced the inset of Fig. 17.1.

The two experiments detailed in Chapter 20 were large international collaborations which the author travelled to and contributed towards, primarily in data processing and analysis. Thomas Barillot was the principal investigator for both experiments.

Chapter 2

Covariance mapping spectroscopy

Covariance mapping spectroscopy is a statistical technique that was first developed by Frasinski et al. [38] as a means to unravel the laser-induced fragmentation mechanisms of small molecules from highly fluctuating experimental data. It extends the principle of covariance from single variables to random discrete functions, such as ion, electron or photon spectra, in order to reveal information that is lost during simple one-dimensional averaging of the spectra. Generally, the technique is able to identify pairs of spectral signals resulting from the same physical process [42]. Covariance mapping was first applied to ion time-of-flight (TOF) spectra to study the dynamics of molecular dissociation upon ionisation by a picosecond pulse [38] and more recently an extension of the covariance mapping technique, known as partial covariance mapping, has found use in experiments at X-ray free electron lasers (XFELs) [43, 44]. This is in part due to the highly fluctuating experimental conditions which are inherent to free election laser emission.

Covariance mapping rests on the calculation of the covariance between each spectral channel across a pair of discretised spectra. The covariance between two random variables is a measure of the degree to which they fluctuate together from event to event. Positive covariance indicates that if one variable increases in value, the other does too, whilst negative covariance indicates that if one variable increases in value, the other decreases. The absolute value of the covariance depends on how consistently the variables fluctuate together and on the mean value of both the random variables

(higher covariance indicates more consistently synchronised fluctuations and greater mean value of either variable). Zero covariance between two variables indicates that either they do not fluctuate, or that their fluctuations are independent.

For two random variables, Y and X, the covariance is calculated as the expectation value of the product of the deviation from the mean of each variable across all events:

$$Cov(Y, X) = \frac{\sum_n (Y_i - \langle Y \rangle)(X_i - \langle X \rangle)}{n}$$

$$= \langle (Y - \langle Y \rangle)(X - \langle X \rangle) \rangle,$$

where $\langle Y \rangle$ ($= \frac{\sum_n Y_i}{n}$) and $\langle X \rangle$ are the expectation values of Y and X respectively, and the covariance is calculated across n events[1]. For ease of calculation, this is typically rewritten as:

$$\langle (Y - \langle Y \rangle)(X - \langle X \rangle) \rangle$$

$$= \langle YX - Y\langle X \rangle - \langle Y \rangle X + \langle Y \rangle \langle X \rangle \rangle$$

$$= \langle YX \rangle - \langle Y \rangle \langle X \rangle - \langle Y \rangle \langle X \rangle + \langle Y \rangle \langle X \rangle$$

$$= \langle YX \rangle - \langle Y \rangle \langle X \rangle, \tag{2.1}$$

where the penultimate step is possible thanks to the linearity of the expectation operator.

Covariance mapping extends the principle of covariance to randomly varying discretised functions, which we represent as vectors $\mathbf{Y} = [Y_1; Y_2; \ldots; Y_m]$ (a column vector) and $\mathbf{X} = [X_1, X_2, \ldots, X_l]$ (a row vector), where each element of each vector varies at each sample. Calculation of the covariance between these two vectors yields a covariance matrix:

$$Cov(\mathbf{Y}, \mathbf{X}) = \langle \mathbf{YX} \rangle - \langle \mathbf{Y} \rangle \langle \mathbf{X} \rangle \tag{2.2}$$

[1] In practice $\langle Y \rangle$ and $\langle X \rangle$ are always calculated sample means, in which case Bessel's correction [47] causes the denominator above to become $n - 1$.

$$= \begin{bmatrix} Cov(Y_1, X_1) & Cov(Y_1, X_2) & Cov(Y_1, X_3) & \dots & Cov(Y_1, X_l) \\ Cov(Y_2, X_1) & Cov(Y_2, X_2) & Cov(Y_2, X_3) & \dots & Cov(Y_2, X_l) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Cov(Y_m, X_1) & Cov(Y_m, X_2) & Cov(Y_m, X_3) & \dots & Cov(Y_m, X_l) \end{bmatrix},$$

where the element at (i, j) represents the covariance between element i of vector $\mathbf{Y}$ and element j of vector $\mathbf{X}$.

In covariance mapping spectroscopy, the randomly varying vectors $\mathbf{Y}$ and $\mathbf{X}$ are experimental spectra such as ion or electron time-of-flight spectra, with hundreds to several thousands of bins each. As such, the covariance matrix is more easily visualised as a colour map such as that shown in figure 2.1, reproduced with permission from [38]. The left hand panel shows the covariance map between the ion time-of-flight spectra of the Coulomb explosion of CO, induced by a picosecond laser pulse. The covariance map here has been calculated using the same spectrum for $\mathbf{Y}$ and $\mathbf{X}$, such that $\mathbf{Y} = \mathbf{X}^T$. Such autocovariance maps are symmetric about the line $x = y$.

Panels **A** and **B** of Fig. 2.1 show, as a colour map, the two matrices $\langle \mathbf{YX} \rangle$ (the 'correlated product') and $\langle \mathbf{Y} \rangle \langle \mathbf{X} \rangle$ (the 'uncorrelated product') respectively, and it is the difference between these two matrices that produces the covariance map shown on the left hand side. The covariance map was calculated across 26 000 recorded spectra generated by repeated laser pulses. The 'f' and 'b' subscripts on the one dimensional averaged spectra $\langle \mathbf{X} \rangle$ and $\langle \mathbf{Y} \rangle$ shown along the side of the covariance map, denote ionic fragments ejected 'forwards' (which take a shorter time to reach the detector) and 'backwards' (longer time to reach the detector) from the interaction region, as seen looking towards the detection plate of the TOF spectrometer. The polarisation of the laser field lies along the length of the drift tube, and parent ions are more easily ionised when their axis lies along this electric field, [48], resulting in the splitting of spectral peaks into 'forward' and 'backward' components.

Positive islands on the covariance map show correlations between different regions of the ion time-of-flight spectra, revealing new information which is entirely lost in the standard averaging procedure applied to the averaged spectra $\langle \mathbf{X} \rangle$ and $\langle \mathbf{Y} \rangle$. Where

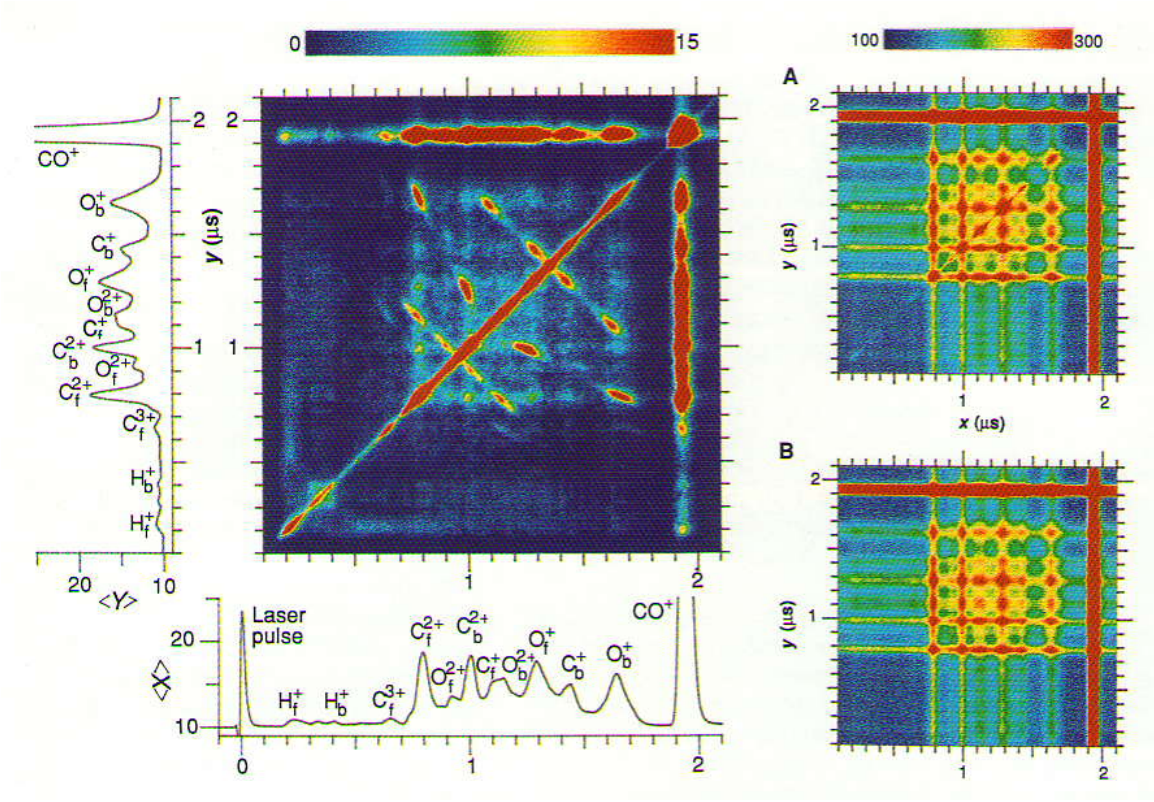


Figure 2.1: Covariance map of the Coulomb explosion of CO, reproduced with permission from reference [38]. The covariance map is the difference between the correlated product (**A**) and the uncorrelated product (**B**).

two regions in the time-of-flight spectrum are correlated by a positive island on the covariance map, the two ions measured at these two regions of the spectrum are produced by the same physical process. This can be understood in the following manner. The number of molecules undergoing a particular fragmentation mechanism at each shot naturally fluctuates. For each different fragmentation mechanism, the shot-to-shot fluctuations of the abundance of each of the ionic products produced by the particular process will follow the fluctuations of the number of molecules that undergo that fragmentation mechanism. This means that the covariance between the regions of the time-of-flight spectrum at which each of the products created in the same fragmentation process is detected will be positive. Zero covariance between spectral signals indicates that the creation of the measured fragment ions is independent.

As such, the experimental results shown in figure 2.1 enabled elucidation of the dif-

ferent fragmentation processes taking place in the Coulomb explosion of the molecule. For example, an O^+ ion observed in the one-dimensional spectrum could be produced by any of the processes [$CO^+ \rightarrow C + O^+$], [$CO^{2+} \rightarrow C^+ + O^+$], and in general [$CO^{n+} \rightarrow C^{(n-1)+} + O^+$] for higher charged precursor ions. It is not possible to tell which process produced the O^+ ion from the spectral signal in the one-dimensional mass spectrum. On the covariance map however, the island of positive covariance observed between the $O_b{}^+$ and $C_f{}^{2+}$ ions unambiguously identifies the [$CO^{3+} \rightarrow C^{2+} + O^+$] fragmentation pathway as one of the dissociation mechanisms involved.

The strong line running through the map along $x = y$ is a result of the fact that the map is an autocovariance map, with $\mathbf{Y} = \mathbf{X}^T$. The increase or decrease in a spectral signal from scan-to-scan is trivially matched by the increase or decrease of that same signal, meaning each spectral signal shows positive covariance with itself. The value of the covariance map at $x = y$ is the variance of the one-dimensional spectrum at x ($= y$), and the width of this so-called autocorrelation line is a function of the resolution of the detector. The wider a spectral signal, the wider its associated feature on the autocorrelation line.

The interpretation of negative covariance on a covariance map is sometimes taken to show competing processes. For example, for the above molecular system, it might be expected that because a triply charged CO^{3+} molecule can decompose into either C^{2+} and O^+ or C^+ and O^{2+}, the covariance between O^+ and O^{2+} would be negative because when one is produced the other is not. However, this is not the case. For a sufficiently dilute sample, there is no interaction between individual parent ions. This means that whether O^+ is produced by the decomposition of one CO^{3+} parent ion has no effect whatsoever on whether O^{2+} is produced by another. This is in contrast to the direct effect that the abundance of C^+ has on the abundance of O^{2+} – if an instance of the former is created then an instance of the latter is also created by the same parent ion. As such, the abundances of the two ions O^{2+} and O^+ cannot 'communicate' with each other, and the covariance between the two is expected, and found, to be zero [49].

It is however possible to find negative regions on covariance maps. Such regions

are physically interpretable, and can be understood through a mediating external fluctuating parameter. Some experimental parameters can be of the nature that their change in one direction will increase the yield of one product whilst decreasing the yield of another. For example, non-negligible fluctuations in laser pulse energy from shot-to-shot in the experimental regime considered above could indeed induce negative correlations between O^{2+} and O^{+}. If higher pulse energy has a tendency to produce more highly charged fragments at the expense of the lower charged fragments, then every time the abundance of O^{2+} increases due to higher pulse energy at a particular shot, the abundance of O^{+} will decrease at the same shot. This will cause the two ions to show negative covariance. The two signal channels O^{+} and O^{2+} are therefore caused to anti-correlate by the mediating fluctuating parameter of the laser pulse, which allows for some 'communication' between the non-interacting parent ions insofar as they all experience the same laser intensity. The fairly stable experimental conditions for the experiment in Fig. 2.1 precluded the measurement of negative covariance, which even in the case of more unstable conditions would likely have been masked by other effects which will be discussed below.

Since its first demonstration on the ion time-of-flight spectra from the photoinduced Coulomb explosion of CO, N_2O and SO_2 [38], covariance mapping has been used, among other applications, to unravel the dynamics of the photoinduced fragmentation of N_2 [39] and I_2 [40], as well as to demonstrate that fragmentation of the diatomic molecules N_2 and CO under illumination with light at 305 nm is predominantly charge-symmetric (meaning each fragment ion produced from a dissociation process carries the same charge) [50]. Despite these successes, a restrictive experimental requirement of the technique is that in order to validate the interpretation of positive covariance as indicating products born in the same decomposition reaction, all experimental parameters need to be kept stable in order to suppress uninteresting common-mode correlations positively correlating every spectral signal with every other due to a mediating external fluctuating parameter (e.g. sample density). A recently realised extension of covariance mapping spectroscopy, known as partial covariance mapping spectroscopy, has enabled this significant hurdle to be overcome

mathematically. This has, for example, allowed covariance mapping techniques to be used to analyse data produced under the inherently unstable experimental conditions at XFEL facilities.

Chapter 3

Partial covariance mapping spectroscopy

The interpretation that positive covariance between two spectral signals indicates the signals are produced by the same physical process relies on the often incorrect assumption that there are no fluctuating experimental parameters. If there is a parameter that fluctuates from shot-to-shot which causes the abundance of all species to grow and shrink in unison (such as the target sample density), this will induce positive covariance between all spectral signals on a covariance map (so-called common-mode correlations). The regions of positive covariance that are due to spectral signals sharing common origin are buried under these spurious correlations. The problem of fluctuating experimental parameters is pertinent to X-ray free electron laser research, where the inherently unstable operation of the laser causes significant shot-to-shot fluctuations in pulse energy [51].

The solution to this problem was first proposed in reference [37] in 1996, but its first successful implementation was only published in 2013 [43, 44]. Provided the increase in the number of fragmentation products grows linearly with the increase in the fluctuating parameter, by recording the value of the offending experimental parameter I for each of the n recorded shots it is possible to mathematically remove the effect of the fluctuating parameters by calculating the partial covariance, $pCov$. Formally, the partial covariance is the covariance between the residuals $\mathbf{e}X$ and $\mathbf{e}Y$

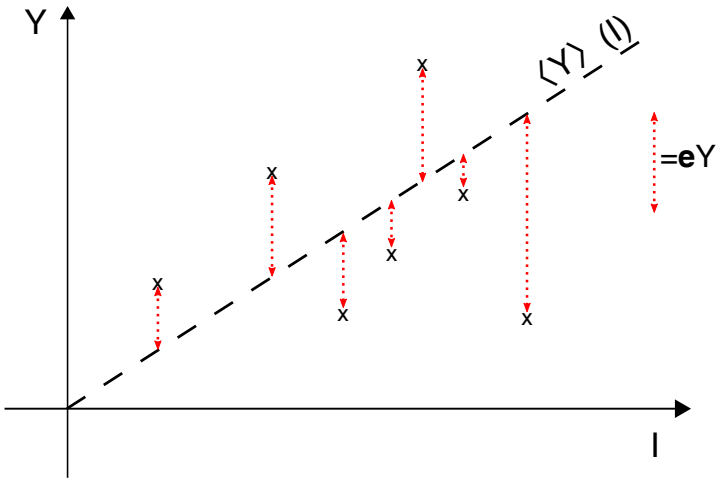


Figure 3.1: Visualisation of the residual $\mathbf{e}Y$ from the linear regression of Y on I.

from the linear regression of X and Y respectively on I. A visualisation of the residual $\mathbf{e}Y$ is shown in Fig. 3.1, where the crosses denote randomly selected shots. The measured value of I at each shot is plotted on the x-axis, and the measured value of Y on the y-axis. The residual $\mathbf{e}Y$ (red dashed line) is the difference between the measured value of Y at each shot and the expected value of Y, given the measured value of I, according to a linear regression of Y on I across all shots (black dashed line).

The formal derivation of the formula for partial covariance (which can be found in Chapter 14.4 of [52]) is fairly involved, but the final result is quite simple:

$$pCov(\mathbf{Y},\mathbf{X};I) = Cov(\mathbf{Y},\mathbf{X}) - \frac{(\langle \mathbf{Y}I\rangle - \langle \mathbf{Y}\rangle\langle I\rangle)(\langle I\mathbf{X}\rangle - \langle I\rangle\langle \mathbf{X}\rangle)}{\langle I^2\rangle - \langle I\rangle^2}. \tag{3.1}$$

In this expression, I is the value of the fluctuating external parameter, the column vector $(\langle \mathbf{Y}I\rangle - \langle \mathbf{Y}\rangle\langle I\rangle)$ is $Cov(\mathbf{Y},I)$, $(\langle I\mathbf{X}\rangle - \langle I\rangle\langle \mathbf{X}\rangle)$ (a row vector) is $Cov(I,\mathbf{X})$ and $\langle I^2\rangle - \langle I\rangle^2$ is the variance of I, $var(I)$. This equation is straightforwardly extended to cases with an arbitrary number of fluctuating parameters causing common-mode correlations [42]:

$$pCov(\mathbf{Y},\mathbf{X};\mathbf{I}) = Cov(\mathbf{Y},\mathbf{X}) - Cov(\mathbf{Y},\mathbf{I})Cov(\mathbf{I}^T,\mathbf{I})^{-1}Cov(\mathbf{I}^T,\mathbf{X}), \tag{3.2}$$

where now $Cov(\mathbf{I}^T, \mathbf{I})^{-1}$ is the inverse of the dispersion matrix of the fluctuating parameters $\mathbf{I}$. $\mathbf{I}$ is now a row vector constructed at each shot, $\mathbf{I} = [I_1, I_2, \ldots, I_n]$, in which element I_i is the value of the i^{th} fluctuating external parameter at the given shot. Because the units of $\mathbf{I}$ cancel out in Eq. 3.2 there is no restriction on the units of each I_i, and the formula is insensitive as to whether the individual elements of $\mathbf{I}$ bear any relationship to each other (e.g. $I_1 = \alpha I_2$, α is a constant), or not.

Partial covariance mapping has proved particularly useful for analysing data from experiments at XFEL facilities, where significant shot-to-shot fluctuations in pulse energy are inherent to the operation of the laser [51]. In a simple picture, shot-to-shot fluctuations in the laser pulse energy induce common-mode correlations between all spectral signals on the covariance map because shots with a greater number of photons will create more fragmentation products (ions or electrons) in every spectral channel. Following this simple argument through, the number of fragmentation products produced will grow linearly with the pulse energy (which in turn grows linearly with the number of photons at each shot), as required for the application of Eq. 3.1. The advanced diagnostics available at XFEL facilities such as the Linac Coherent Light Source (LCLS) at the SLAC National Accelerator Laboratory, U.S.A., and the Free-Electron LASer in Hamburg (FLASH) allow for monitoring of the pulse energy from shot-to-shot.

Partial covariance mapping of the electron spectra from neon irradiated by intense 1062 eV X-rays at the LCLS successfully elucidated a large number of different inner-shell processes following core ionisation [44], and partial covariance mapping of the ion time-of-flight spectra from the X-ray induced Coulomb explosion of I_2 and N_2 measured at FLASH, successfully disentangled a large number of competing Coulomb explosion pathways [43]. Figure 3.2, taken from reference [43], demonstrates the effect of partial covariance mapping. The maps presented are calculated from 378 219 single-shot ion time-of-flight spectra, from the Coulomb explosion of N_2 radiated with $\sim$100 fs pulses of 91 eV photons at average pulse energy 48 µJ.

Panels (a) and (b) show the correlated and uncorrelated product, $\langle \mathbf{YX} \rangle$ and $\langle \mathbf{Y} \rangle \langle \mathbf{X} \rangle$, respectively (equivalent to panels (A) and (B) in Fig. 2.1). $\langle \mathbf{Y} \rangle \langle \mathbf{X} \rangle$ is

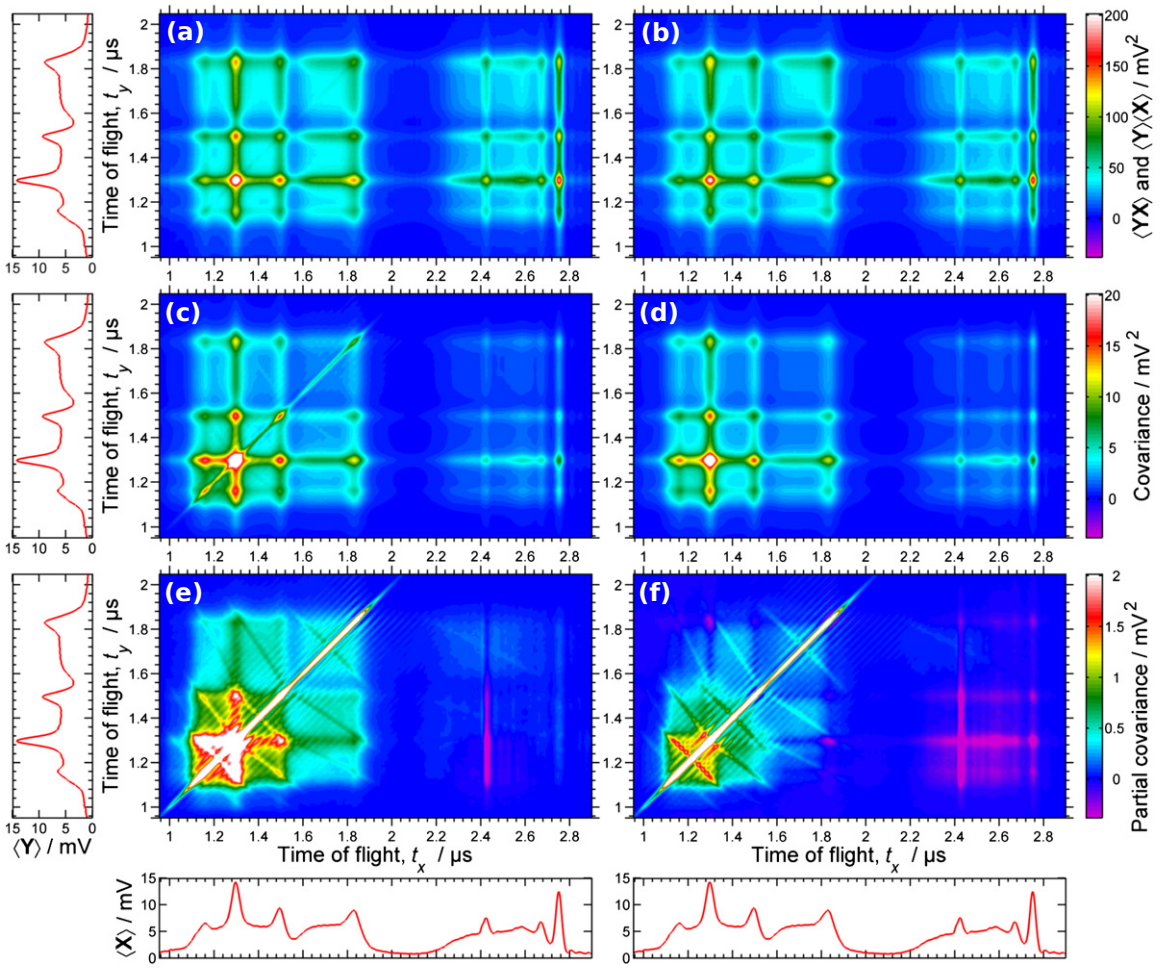


Figure 3.2: The stages of partial covariance mapping, reproduced with permission from reference [43]. The panels show: (**a**) the correlated product, (**b**) the uncorrelated product, (**c**) the simple covariance map, (**d**) the correction term, (**e**) the partial covariance map and (**f**) the over-corrected partial covariance map, with 110% of the correction term removed.

subtracted from $\langle \mathbf{YX} \rangle$ to give the covariance map $Cov(\mathbf{X}, \mathbf{Y})$ in panel (c). In this map, positive covariance islands are be seen between all spectral signals in the one-dimensional averaged spectrum and the map is dominated by vertical and horizontal lines. This gives a 'chequered' pattern which is characteristic of covariance maps featuring common-mode correlations. A vertical (horizontal) line at $x(y)= a$ on the covariance map means that the spectral signal at a on the one-dimensional spectrum correlates with every other spectral signal, usually the result of an external fluctuating parameter causing the abundance of every spectral signal to rise and fall in synchronisation from shot-to-shot. In the case of Fig. 2.1, the external fluctuating parameter is the laser intensity.

To address this issue, the correction term $Cov(\mathbf{Y}, I)Cov(I, \mathbf{X})/var(I)$ has been calculated, and this is plotted on the colour map in panel (d). Close reading of the colourbar for panels (c) and (d) shows that the heavily dominant contribution to the map in panel (c) comes from these common-mode correlations. Finally, the correction term plotted in panel (d) has been taken from the covariance map in panel (c) to give the partial covariance map shown in Fig. (e). The vertical horizontal lines of positive covariance which dominated the map in panel (c) have been supressed, revealing a series of diagonal lines. These diagonal lines are a manifestation of the conservation of momentum. For a molecule that Coulomb explodes into two ionic fragments, if one of the fragments is ejected towards the detector it will arrive at the detector sooner than if it was created with zero initial momentum, at the expense of the other ejected fragment which will be ejected away from the detector to conserve momentum. Depending on the magnitude of the momenta of the ejected particles along the drift tube axis, the additional time taken by the backwards-ejected fragment to reach the detector is (approximately) linearly compensated by the reduction in time-of-flight of the forwards-ejected fragment, meaning the correlation island produced by the Coulomb explosion process in question will trace a straight line across the partial covariance map. In panel (f), 110% of the correction term has been removed to suppress further residual correlations due to the fluctuation of other experimental parameters (such as the density of N_2 molecules in the interaction region).

Partial covariance mapping gives the possibility to extract interesting mechanistic correlations from extremely noisy data which are affected by any number of externally fluctuating parameters. However, the successful implementation of the current partial covariance mapping procedure requires the value of each of the offending fluctuating external parameters to be measured from shot-to-shot, which in many cases presents an insurmountable experimental challenge.

Chapter 4

Formal theory of covariance for noisy Poisson processes

Mikosch and Patchkovskii [53] formulated an analytical theory for covariance measurements under the assumption that the event rate ν (which in the case of the experiments carried out in this thesis corresponds to the number of ions trapped at each scan) follows what they term an 'augmented Poisson distribution':

$$P_\nu(n) = \frac{1}{n!}\nu^n e^{-\nu}, \tag{4.1}$$

where ν is itself distributed from scan-to-scan according to $\nu = \nu_0\gamma$, with ν_0 the average event rate for a given experiment and γ following a normal distribution centred around $\gamma = 1$ and characterised by standard deviation σ:

$$P_\sigma(\gamma) = \frac{1}{(2\pi)^{1/2}\,\sigma}\exp\left(-\frac{(\gamma-1)^2}{2\sigma^2}\right). \tag{4.2}$$

Figure 4.1 shows the distribution of the number of ions trapped from scan-to-scan in our typical nanoelectrospray (NSI, see section 5.1) experiments, for automatic gain control[1] (AGC, see page 70) targets 100 (panel **A**) and 300 (panel **B**). These plots were made by taking long ($\geq 10,000$ scans) MS/MS measurements[2] at normalised

[1] This is an experimental parameter which effectively controls the central event rate ν_0.

[2] see Chapter 5

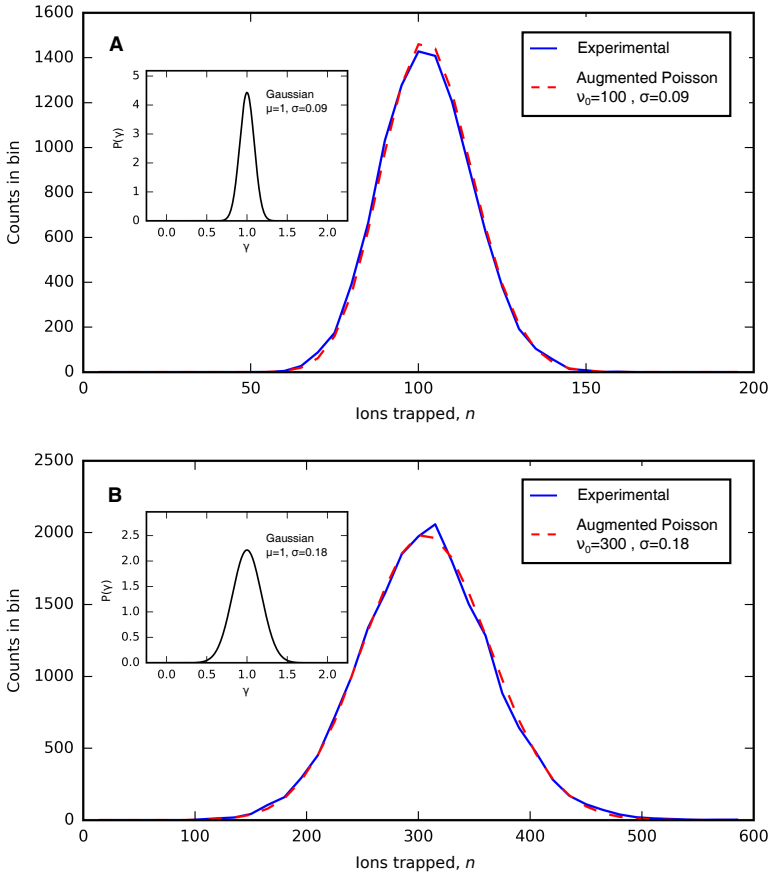


Figure 4.1: Distribution of number of trapped ions at AGC targets (**A**) 100 and (**B**) 300. The sample peptide ions used were $[\text{SDGRG}+2\text{H}]^{2+}$ and $[\text{GKGGK}_{\text{Ac}}\text{GLGK}_{\text{Ac}}\text{GGAKR}+3\text{H}]^{3+}$, respectively.

collision energy 0, which simply accumulates ions in the ion trap during an injection time calculated for the specified AGC target and then scans them out for measurement. The signal response provided by the detector is in arbitrary units, so it is normalised according to the fact that the AGC target corresponds to the approximate number of ions trapped at each scan [54]. The experimental distribution of the number of ions trapped at each scan is plotted in blue, and a fit to theoretical distributions determined numerically from Eqs. 4.1 and 4.2, with ν_0 equal to the AGC target value and σ as the fitting parameter, is plotted in red. The experimental distributions of the number of trapped ions from scan-to-scan show remarkable agreement with this augmented Poisson distribution, demonstrating the validity of such experimental measurements for analysis under this theory. The insets show the distribution described by the Gaussian curve of Eq. 4.2 for the experimentally determined values of σ for each of the two measured distributions. The area of each of these probability density functions falling below $\gamma = 0$ (marked by red line) is negligible, which validates the use of the integration limits used in Eq. 4.3 below.

This chapter introduces the formal theory of covariance for noisy Poisson processes laid out in [53], with close reference to the original publication throughout. The statistics are analysed by considering m possible outcomes of interest for each elementary event, where each outcome occurs with an associated probability P_i. The sum of these probabilities $P_s = \sum_{i=1}^{m} P_i$, and the total probability for all possible outcomes of an elementary event is $P_s + P_0 = 1$, where P_0 is the sum of the probabilities of all other events, which are not of interest.

Mikosch and Patchkovskii consider the overall outcome of an experimental scan as a composite event described by a vector $\mathbb{N}$ of length m, where each element n_i corresponds to the number of times a different one of the m possible elementary outcomes of interest i (e.g. the detection of two complementary ions, the detection of one ion only) has taken place at that scan. The total number of elementary events of interest taking place at each scan is ν, which is the sum of the vector $\mathbb{N}$ ($= \sum_{i=1}^{m} n_i = n_s$) plus n_0, the number of times an outcome that is not of interest occurs. Elementary events are assumed to be independent, which is a valid assumption

for the low sample densities typical of both photoionisation and MS/MS experiments[3] in the gas phase.

Under this model, the authors derive the general probability for the compound outcome $\mathbb{N}$ to occur from a noisy Poisson process given an average event rate ν_0 and standard deviation σ for the distribution of γ:

$$\begin{aligned} P(\mathbb{N}) &= \int_{-\infty}^{\infty} P(\mathbb{N} \mid \nu_0\gamma) P_\sigma(\gamma) d\gamma \\ &= \frac{\nu_0^{n_s}}{(2\pi)^{1/2}\,\sigma} \int_{-\infty}^{\infty} \gamma^{n_s} \exp\left(-\nu_0 P_s \gamma - \frac{(\gamma-1)^2}{2\sigma^2}\right) d\gamma \times \prod_{i=1}^{m} \frac{P_i^{n_i}}{n_i!} \end{aligned} \tag{4.3}$$

Here, integration is extended from $\gamma = -\infty$ to $\gamma = \infty$, which covers unphysical negative event rates (when $\gamma < 0$, $\nu = \nu_0\gamma < 0$). By assumption, however, σ is small enough such that the area of the probability density function of γ that falls below zero is negligible, and so despite these integration limits the theory is not compromised by the inclusion of unphysical events. This assumption has been experimentally justified for the pC-2DMS experiments described in this thesis, see the insets of Fig. 4.1 which show that at the values of σ determined for such experiments this is indeed the case.

The integral in 4.3 reduces to tabulated integrals, and following lengthy manipulation the final expression for the probability of the compound event $P(\mathbb{N})$ is:

$$\begin{aligned} P(\mathbb{N}) =& \nu_0^{n_s} \exp\left[-\nu_0 P_s + \frac{1}{2}\left(\nu_0 P_s \sigma\right)^2\right] \\ & \times f(n_s, \nu_0 P_s, \sigma) \prod_{i=1}^{m} \frac{P_i^{n_i}}{n_i!}. \end{aligned} \tag{4.4}$$

In this expression, f is a polynomial in σ with order $2n_s$:

$$f(n_s, \nu_0 P_s, \sigma) = \sum_{k=0}^{n_s} \sigma^{2k} g\left(k - \frac{n_s}{2}, n_s, \nu_0 P_s\right), \tag{4.5}$$

[3]Sample number densities drop to $\sim 5\times10^{-4}$ cm^3 for linear ion trap mass spectrometers operating just below the space-charge regime [55].

where g is a coefficient given by:

$$g(u, n_s, \nu_0 P_s) = \frac{(-\nu_0 P_s)^{u+|u|}\, n_s!}{(2|u|)!\,(n_s - 2|u|)!!} \tag{4.6}$$

$$\times\ {}_1F_1\left(|u| - \frac{n_s}{2}; 2|u| + 1; 2\nu_0 P_s\right). \tag{4.7}$$

Here, $x!!$ is the odd factorial of x and ${}_1F_1$ is Kummer's confluent hypergeometric function, which avoids being a infinite polynomial when, as is the case for Eq. 4.7, the first argument is a non-positive integer:

$${}_1F_1(a; b; z) = \sum_{l=0}^{|a|+1} \frac{a^{(l)}}{b^{(l)}} \frac{z^l}{l!}, \tag{4.8}$$

with $x^{(l)}$ denoting the rising factorial:

$$x^{(l)} = x\,(x+1)\,(x+2)\ldots(x+l-1)\,.$$

For the consideration of event covariances we will be interested in the moments of the distribution of $P(\mathbb{N})$. The multivariate moment $M(\mathbb{K})$, where $\mathbb{K} = [k_1, k_2, \ldots, k_m]$, denotes the expectation value $\mathbb{E}\left[\prod_{i=1}^{m} n_i^{k_i}\right]$ so that the value of element k_i selects the desired order of the moment for the i^{th} elementary outcome. An explicit example of how these multivariate moments are used to calculate covariances is presented at the end of this chapter.

For a discrete number of possible elementary outcomes n, such as those considered for the experiments analysed here, the moments $M(\mathbb{K})$ are generally defined by:

$$M(\mathbb{K}) = \sum_{n_1=0}^{\infty} \ldots \sum_{n_m=0}^{\infty} P\,(\mathbb{N}) \prod_{i=1}^{m} n_i^{k_i}, \tag{4.9}$$

so for the distribution $P(\mathbb{N})$ these will become:

$$\begin{aligned} M(\mathbb{K}) = & \exp\left[-\nu_0 P_s + \frac{1}{2}\left(\nu_0 P_s \sigma\right)^2\right] \\ & \times \sum_{n_s=0}^{\infty} \nu_0^{n_s} f\left(n_s, \nu_0 P_s, \sigma\right) \\ & \times \sum_{n_1} \cdots \sum_{n_m} \prod_{i=1}^{m} n_i^{k_i} \prod_{j=1}^{m} \frac{P_j^{n_j}}{n_j!}. \end{aligned} \tag{4.10}$$

It can be verified that the final line of this expression can be written as:

$$\sum_{n_1} \cdots \sum_{n_m} \prod_{i=1}^{m} n_i^{k_i} \prod_{j=1}^{m} \frac{P_j^{n_j}}{n_j!} = \frac{1}{n_s!} \left(\prod_{i=1}^{m} \hat{n}_i^{k_i} \right) P_s^{n_s},$$

with the operator $\hat{n}_i$ given by:

$$\hat{n}_i = P_i \frac{\partial}{\partial P_i}.$$

Operating with $\hat{n}_i$ requires evaluation of the partial derivative of the $P_s^{n_s}$ with respect to the individual probabilities P_i. From Eq. 4.4, this requires the partial derivative of $f(n_s, \nu_0 P_s, \sigma)$ with respect to $\nu_0 P_s$, which in turn requires finding the partial derivative of $g(u, n_s, \nu_0 P_s)$ with respect to the same variable. From Eq. 4.7, the partial derivative of $g(u, n_s, \nu_0 P_s)$ with respect to $\nu_0 P_s$ is:

$$\frac{\partial}{\partial \nu_0 P_s} g(u, n_s, \nu_0 P_s) = -n_s g\left(u - \frac{1}{2}, n_s - 1, \nu_0 P_s\right), \tag{4.11}$$

so that:

$$\frac{\partial}{\partial \nu_0 P_s} f(n_s, \nu_0 P_s, \sigma) = -n_s \sigma^2 f(n_s - 1, \nu_0 P_s, \sigma). \tag{4.12}$$

Using the definition of $\hat{n}_i$, it is possible to obtain the following expression for the

compound operation of the operator:

$$\begin{aligned}\hat{n}_i^{k_i}\hat{n}_j^{k_j}\left(P_s^{n_s+1}\right) =& P_s\left[\hat{n}_j^{k_j}\hat{n}_i^{k_i}P_s^{n_s}\right] \\ &+ P_i\sum_{b_i=0}^{k_i-1}\binom{k_i}{b_i}\left[\hat{n}_j^{k_j}\hat{n}_i^{b_i}P_s^{n_s}\right] \\ &+ P_j\sum_{b_j=0}^{k_j-1}\binom{k_j}{b_j}\left[\hat{n}_j^{b_j}\hat{n}_i^{k_i}P_s^{n_s}\right]. \end{aligned} \tag{4.13}$$

Finally, applying operator $\hat{n}_j$ to Eq. 4.10 using Eq. 4.13 and the partial derivative from Eq. 4.12 renders the following recursive definition for the moments of $P(\mathbb{N})$:

$$\begin{aligned}M(\mathbb{K}+\mathbb{I}_j) =& P_j\nu_0 M(\mathbb{K}) + P_j\frac{\partial}{\partial P_j}M(\mathbb{K}) \\ &+ \nu_0^2\sigma^2 P_j\sum_{l=1}^{m}P_l\sum_{b_l=0}^{k_l-1}\binom{k_l}{b_l} \\ &\times M\left(\mathbb{K}+\mathbb{I}_l\left(b_l-k_l\right)\right), \end{aligned} \tag{4.14}$$

where $\mathbb{I}_j$ is a vector of length m with 1 at position j and zeroes elsewhere, meaning the recursion is performed in steps of one order for each individual outcome. The recursion can be initiated by using the fact that the zeroth moment of a probability distribution is equal to 1, so that $M(\mathbb{K}) = 1$ when $\mathbb{K}$ is a vector of zeroes of any length.

We can now examine a simple experimental case to illustrate how this theory is applied to calculate the covariance between two signal channels. Let us consider a doubly charged parent ion which, following photoionisation, always fragments into two ions. These two ions are measured by a detector to which we shall begin by attributing unit detection efficiency. The parent ion can fragment into many different pairs of fragment ions, of which one possible pair is A^+ and B^+. The intensity in the signal channels in which these two ions are measured is denoted A and B, respectively, and the detector response is calibrated so that one unit is recorded for every detected ion. We want to know what the measured covariance between the ions A^+ and B^+, which

is calculated as $\langle AB\rangle - \langle A\rangle\langle B\rangle$, will be, and we know the average event rate ν_0, the noise term σ, and the fragmentation probabilities required to calculate the necessary multivariate moments according to the above analysis. From here, calculating the covariance requires careful consideration of all the possible elementary events that can lead to detection of a signal in either of the channels of interest. In this case, these elementary events are:

1. Both A^+ and B^+, created in the same fragmentation event, reach the detector. This takes place with probability P_1.

2. A^+ and another ion Y^+ which is not B^+ reach the detector. This could be the result of a secondary fragmentation of B^+, and happens with probability P_2.

3. B^+ and another ion Y^+ which is not A^+ reach the detector. This could be the result of a secondary fragmentation of A^+, and happens with probability P_3.

4. Two ions Y^+ and Z^+, neither of which is A^+ or B^+, reach the detector. This has probability P_4.

Let us now see what the result of $\langle AB\rangle - \langle A\rangle\langle B\rangle$ will be. Remember that the average denoted by the triangular brackets is over all scans, and that many different elementary events take place at each scan. The second term is the more simple of the two to analyse so we shall start there. This is the product of the mean of the signal measured in channel A and the mean of the signal measured in channel B. There is a contribution to signal channel A when both A^+ and B^+ are measured together (elementary event 1), and when A^+ is measured by itself (elementary event 2). So the measured signal in channel A will be $n_1 + n_2$, and the expectation value for A across all scans will be $\langle n_1 + n_2\rangle$, which can be rewritten as $\langle n_1\rangle + \langle n_2\rangle$ thanks to the linearity of the expectation operator. According to the definition of the multivariate moment $M(\mathbb{K})$ given on page 33, the mean value for the number of times e.g. elementary event 2 happens is $M(0,1,0,0)$, with the number 1 at element 2 selecting the first moment (i.e. mean) of the second elementary event. This allows us to rewrite $\langle n_1\rangle + \langle n_2\rangle$ in terms of the multivariate moments $M(\mathbb{K})$ as $M(1,0,0,0) + M(0,1,0,0)$. By the

same reasoning as above, the measured signal in channel B has a contribution from elementary events 1 and 3, so will be $n_1 + n_3$, with expectation value $\langle n_1 \rangle + \langle n_3 \rangle$. This term itself can written as $M(1,0,0,0) + M(0,0,1,0)$, and so the full expression for the second term of the covariance formula $\langle A \rangle \langle B \rangle$ in terms of the multivariate moments $M(\mathbb{K})$ will be $[M(1,0,0,0) + M(0,1,0,0)]\,[M(1,0,0,0) + M(0,0,1,0)]$.

Returning to the covariance formula, the first term $\langle AB \rangle$ is the mean value of the scan-to-scan product of the measured signals in channels A and B. The signal measured in channel A is $n_1 + n_2$ and the signal measured in channel B is $n_1 + n_3$, so the product AB becomes $(n_1+n_2)(n_1+n_3) = n_1^2+n_1n_3+n_2n_1+n_2n_3$. The expectation value of this expression can again be written as the sum of the expectation values for each individual term thanks to the linearity of the expectation operator, becoming $\langle n_1^2 \rangle + \langle n_1 n_3 \rangle + \langle n_2 n_1 \rangle + \langle n_2 n_3 \rangle$. Again referring to the definition of the multivariate moments $M(\mathbb{K})$ on page 33, this sum of expectation values can be expressed in terms of the $M(\mathbb{K})$ as $M(2,0,0,0) + M(1,0,1,0) + M(1,1,0,0) + M(0,1,1,0)$. Pulling all these terms together, we arrive at the full expression for the covariance for this simple case in terms of the multivariate moments $M(\mathbb{K})$:

$$\begin{aligned}\langle AB \rangle - \langle A \rangle \langle B \rangle \\ &= M(2,0,0,0) + M(1,0,1,0) + M(1,1,0,0) + M(0,1,1,0) - \\ &\quad [(M(1,0,0,0) + M(0,1,0,0)]\,[M(1,0,0,0) + M(0,0,1,0))]\,. \end{aligned} \tag{4.15}$$

Calculation of these moments according to Eq. 4.9 is a long and uninstructive exercise, but for the simple case above with an average event rate ν_0, noise term σ and associated event probabilities P_i, the final result is:

$$Cov(A,B) = P_1\nu_0 + \sigma^2\nu_0^2(P_1 + P_2)(P_1 + P_3). \tag{4.16}$$

This expression tells us that in the case of a perfectly stable event rate ($\sigma = 0$) the covariance between A and B is directly proportional to the overall central event

rate and the outcome probability of elementary event 1, P_1, which in this case[4] is the same as the branching ratio of the fragmentation process which produces the products A^+ and B^+ together. The introduction of a non-zero noise term σ, meaning the event rate fluctuates from scan-to-scan, introduces a 'false' additional term which is proportional to the square of both the overall central event rate and the standard deviation of the event rate's distribution. It also depends on the probabilities of two elementary events where A^+ and B^+ are detected but do not originate from the same fragmentation event, which in this simplified case assuming unit detection efficiency are again the same as the branching ratios of the two underlying proceses. Even in the case where $P_1 = 0$, this additional term causes the covariance between A and B to be greater than zero, and as such it is determined as 'false' covariance because it is not a result of the two products A^+ and B^+ being produced in the same fragmentation reaction. In such cases the measurement of positive covariance is minimally informative, and this false additional term precludes the use of covariance measurements for the identification of products of the same physical origin.

When all the fluctuating experimental parameters contributing to false covariances are measurable from shot-to-shot, the standard procedure for the removal of spurious correlations described in Chapter 3 can be employed. However, this is not possible in the case of the experiments performed in this thesis, requiring an extension to the existing procedure. Chapter 8 presents a formal treatment of the mathematical procedure developed in this thesis to remove this deleterious false covariance for pC-2DMS measurements.

[4]Where we have assumed 100% detection efficiency, meaning the detection probability is equal to the branching ratio.

Chapter 5

Mass spectrometry

Mass spectrometry (MS) is the analysis of charged particles based on their mass-to-charge ratio (m/z). Perhaps the first instance of a modern mass spectrum was recorded by J. J. Thomson in 1912 [56], using a setup designed by F. W. Aston which measured the deflection of a beam of positive ions by a magnetic field of slowly incremented magnitude[1]. Since then, a near-continual stream of remarkable technical advances have enabled the technique of mass spectrometry to characterise and identify a vast number of different and often biologically important molecules, earning its current position as the method of choice for the structural analysis of biomolecules. It has been applied to measure the structure of species as diverse as peptides and proteins [1–6], nucleic acids [7–10], oligosaccharides [12] and lipids [11].

The work reported in this thesis is, in particular, a development of the prevalent technique of tandem mass spectrometry. In tandem mass spectrometry, different stages of mass analysis, usually separated by fragmentation of the molecules under analysis, are performed in tandem by the mass spectrometer. The most common implementation of tandem MS is known as MS/MS or MS^2, in which the original ions under analysis (often referred to as parent or precursor ions) are trapped and fragmented in the gas phase and their detailed primary structure is pieced together from the measured m/z of the resultant fragments. It is possible to perform multi-

[1] An important feature of the design was the use of a Faraday cup to measure the deflected ions instead of a photographic plate, which ensured the measurement was quantitative.

ple iterations of this process, where at each stage some of the fragments produced in the previous stage becomes the new parent ions to be trapped and fragmented. These experiments are generally denoted as MS^n, where there are $n-1$ iterations of trapping and fragmentation before the final fragment ion spectrum is measured. Although there now exist an impressive number of different techniques, each with their own merits, for a) the introduction of biomolecular ions into the gas phase, b) the fragmentation of these biomolecular ions in tandem MS and c) the measurement of the m/z's of both the intact precursor ions and their fragments, the vast majority of the experimental work presented in this thesis has been performed using one of the most widely available compound experimental setups for the mass spectrometric exploration of biomolecular primary structure. This consists of sample introduction by electrospray ionisation, followed by fragmentation resulting from slow heating *via* multiple collisions with a small pressure of an inert gas, so-called collision-induced dissociation (CID, also known as collision-activated dissociation, CAD), and mass measurement using a linear ion trap[2]. These three cornerstones of the technology will now be introduced in turn.

5.1 Electrospray ionisation

Electrospray ionisation provides a way to put large (up to megadaltons [57, 58] or even hundreds of megadaltons [59]) molecular ions from solution into the gas phase. The phenomenon of electrospray was first studied in 1917 by John Zeleny and his assistant W. B. Lang [60], who observed the emission of jets from a highly charged droplet of ethanol. The Nobel-prizewinning work by John B. Fenn, Masamichi Yamashita and co-workers in the late 1980s [14][3], building on previous work from Malcolm Dole and co-workers first reported in 1968 [62], showed that large intact biopolymeric ions could be produced by the application of a high (few kV) voltage to a needle through which

[2]This remarkably flexible piece of hardware also traps, isolates and fragments the molecules under analysis, as described in section 5.2.

[3]This work was concurrent with the independent demonstration of similar results from Russian researchers, e.g. reference [61].

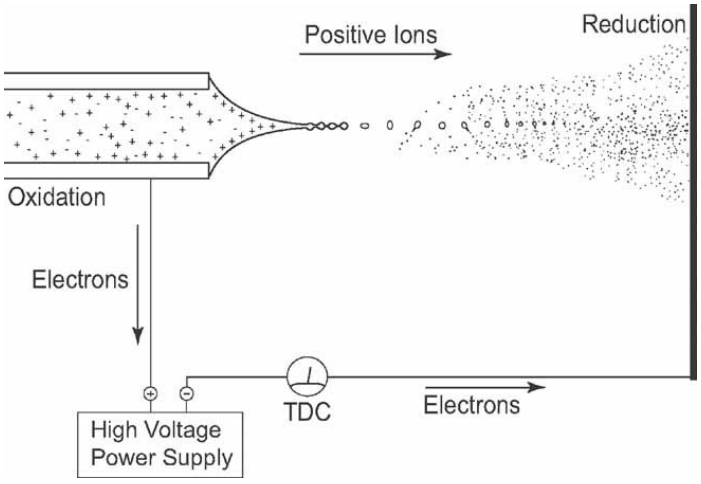


Figure 5.1: Simple schematic of an ESI setup in positive ion mode, reproduced with permission from [59]. 'TDC' stands for 'total droplet current'.

a flow of solvated ions is running. Along with the development of other so-called 'soft ionisation' techniques such as matrix-assisted laser desorption/ionization (MALDI) [15], this paved the way for mass spectrometry to become the preferred method for the analysis of biomolecular structure.

The general setup for electrospray in positive ion mode is shown in Fig. 5.1. The creation of negative molecular ions is possible by simply reversing the polarity of the voltages [14]. Biomolecular ions in solution (for example, the peptide cations measured in this thesis were prepared in a solution of 50% acetonitrile[4] and 2% formic acid in water) are sent through a needle which is charged to a few kV, at atmospheric pressure. The addition of formic acid facilitates the creation of biomolecular cations in solution, through protonation on the basic sites of the molecules. The application of the high voltage to the needle causes a build up of positive ions at the meniscus, which destabilises as a result of Coulombic repulsion. This destabilisation produces a cone, known as the Taylor cone [64], from the apex of which is emitted a stream of small, highly charged droplets.

The droplets in the jet, which are typically ejected from the needle at micrometre radii, undergo solvent evaporation and so reduce in volume. As this happens

[4]Molecular dynamics simulations in [63] found that the use of methanol destabilises an aqueous droplet by interrupting the hydrogen bonding between its water molecules which, given the discussion below, justifies the prevalent use of a partially organic solvent in ESI-MS.

the droplets approach the Rayleigh limit [65], which is the point at which Coulombic repulsion between the charged species in a droplet overcomes the surface tension holding it together[5]. The limit is reached first at the sharpest points on an electrosprayed droplet's inhomogenous surface, where the electric field is highest, and at these points a new Taylor cone is formed from which a stream of smaller, charged droplets (usually around 10% of the diameter of the parent droplet [66]) is emitted [67–69]. Multiple iterations of this process, which is typically aided by applying a flow of a clean, non-reactive gas, produces charged droplets of nanometre size, from which single ions can be emitted [70].

To increase the transfer efficiency of ions from the solvent to the gas phase, a variation of ESI known as nanoelectrospray ionisation (also called nanospray ionisation or NSI) was developed [67, 71]. The fundamental idea behind nanospray ionisation is the creation of such nanometre droplets as soon as possible after leaving the needle to avoid the lossy process of repeated solvent evaporation and jet fission. This is achieved by using a needle opening of only a few µm as opposed to ∼100 µm as typically used for ESI, which in turn requires a much lower sample flow rate (down to less than 10 nL/min [72]) and so uses much smaller quantities of analyte [66, 71]. This also means the nanospray source is less likely to require a flow of auxiliary gas, and that the needle can be placed closer to the entrance of the mass spectrometer without risking contamination by excess solvent, further improving transfer efficiency from the solvent to detectable gas phase ions. The majority of the experiments described in this thesis are performed with a nanoelectrospray ionisation source (Nanospray II Ion Source, Thermo Fisher Scientific).

The gaseous analyte ions detected by MS are directly produced from the highly charged nanodroplets whose production is described above [70, 72]. The precise mechanism for the creation of single gaseous ions from the nanodroplets has proven to be impressively elusive and remains an area of active research [72, 73], but the general mechanics of ion formation are now understood to be different depending on

[5]For a spherical droplet this point is described by $Q = 8\pi \left(\epsilon_0 \gamma r^3\right)^{1/2}$, where Q is the charge on the droplet of radius r, γ is its surface tension, and ϵ_0 is the permittivity of free space.

the characteristics of the analyte. Thus, three general models have emerged which describe different regimes: the 'ion evaporation model' [74, 75], in which Coulomb repulsion forces an ion to the surface of a charged nanodroplet, where the electric field is high enough to facilitate the ejection of a single ion, the 'charged residue model' [62, 76], in which solvent evaporation simply continues until the only the charged analyte remains, and the more recently proposed 'chain ejection model' [70, 77], in which it is the hydrophobicity of an unfolded peptide or protein molecule that forces it to the droplet's surface. The ion evaporation model (IEM) is widely thought to dominate for small, non-polymeric, molecules [59, 72, 78], and the charged residue model (CRM) is believed to be the mechanism by which large, globular, folded protein ions are put into the gaseous phase [79]. The chain ejection model (CEM), which bears many mechanistic similarities to the IEM, was developed to describe the behaviour of unfolded, disordered polypeptide chains. Such unfolding is triggered by acidic (or basic) solutions, which pertains to the sample solutions used for the experiments on peptide, protein and oligonucleotide molecules described in this thesis[6]. Considering the above, the CRM is not expected to play a significant role for the production of gas phase biomolecular ions for the experiments described in this thesis, but both the CEM and, in some cases[7], the IEM are expected to contribute.

As noted above, the CEM shares many similarities with the IEM. In both cases, a single ion is ejected from a highly charged nanodroplet approaching the Rayleigh limit. Within the IEM, electrostatic repulsion inside the droplet causes the even distribution of charge carriers on the surface of the droplet, where the repulsion between a single ion and the electric field on the droplet's surface can become high enough to overcome surface tension and eject a single ion into the gas phase. In the CEM, the hydrophobic effect acting between the aqueous component of the sample solution and the hydrophobic side chains[8] of an unfolded polypeptide is the main cause of the migration of the analyte ion to the droplet's surface instead. Particularly

[6]As above, this ensured sufficient ionic concentration of the analyte through protonation or deprotonation.

[7]Particularly for the shorter peptide molecules measured.

[8]These are alanine (A), phenylalanine (F), glycine (G), isoleucine (I), leucine (L), methionine (M), proline (P), valine (V) and tryptophan (W)

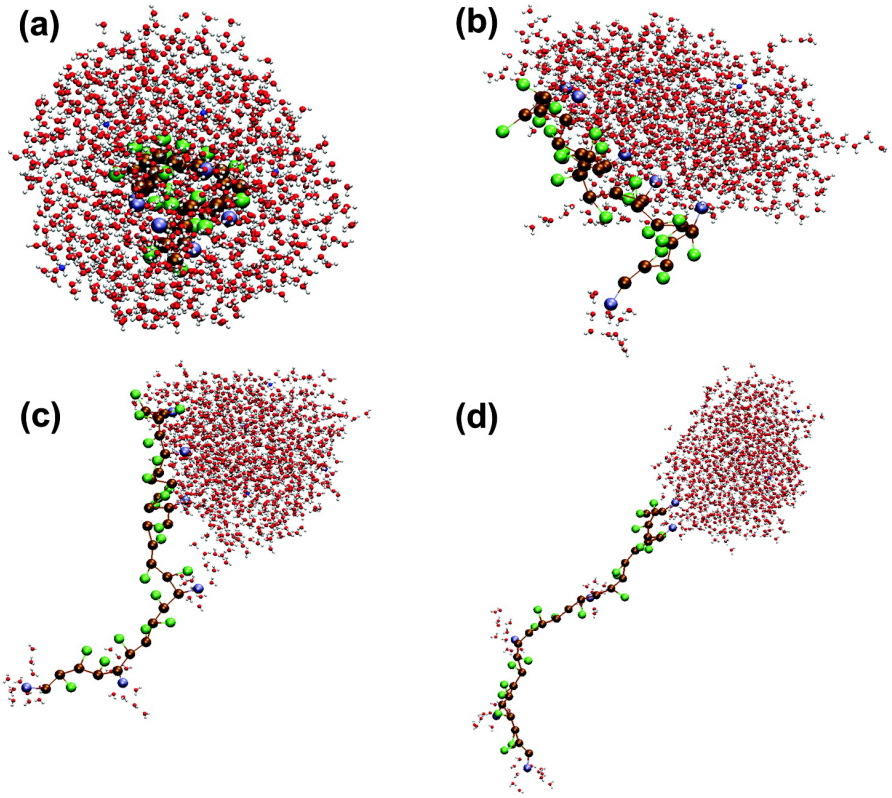


Figure 5.2: The chain ejection model (CEM). Molecular dynamics simulation of the ejection of a 27 residue polypeptide chain from a droplet of 1000 water molecules. Reproduced with permission from [77]. Only the first four panels (a–d) of the original six-panel figure are shown. The different panels show snapshots taken at **(a)** −80 ps, **(b)** 67 ps, **(c)** 302 ps and **(d)** 1147 ps.

for short peptide chains, both mediating forces can be expected to play a role in propulsion of the analyte ion to the surface of a droplet. Within the CEM, once at the surface, one of the chain termini is expelled from the droplet, followed by sequential ejection of the remainder of the polypeptide chain [72]. As for the IEM analyte ejection is primarily mediated by the electrostatic force, this time through a repulsive interaction between charged side chains and the electric field on the droplet's surface, but it is also assisted by the hydrophobic effect [77]. Molecular dynamics simulations performed on a naïve model of a 27 residue polypeptide in water and ammonium ions has demonstrated the CEM. The result of one such simulation, as performed in [77], is shown in Fig. 5.2. The simple 27 residue long polypeptide chain has been modelled as a string of beads. The 27 beads constituting the backbone are plotted in brown, to each of which is connected a positive side chain (purple) or neutral side chain (green). The hydrophobic sequence shown consists primarily of neutral side chains, with six positive side chains. Half of the simulations launched returned the behaviour displayed in this figure, where the unfolded peptide chain is ejected from the droplet. The excess solvent on the ejected ion is lost through further desolvation (e.g. in the heated ion transfer capillary tube, see Fig. 5.3) and collisions with other gaseous molecules (e.g. the buffer gas during cooling in the linear ion trap). Molecular dynamics simulations were also performed modelling the IEM for smaller analytes [63]. For the ejection of ammonium ions from nanodroplets of methanol and water, it was found that a transient 'solvent bridge', consisting of a chain of hydrogen-bonded solvent molecules, can be formed between the charged droplet and the departing ion. This affects the dynamic character of the barrier seen by the departing analyte[9], and this complex interplay between solvent and ejected analyte ion can be expected to play a role in the CEM as well.

Given the stochastic nature of each process which eventually produces gaseous molecular ions in electrospray ionisation, the instantaneous flow of ions produced by an electrospray source is inherently unstable [80]. The nanospray ionisation source

[9]Originally thought to be solely dependent on the interplay between Coulombic repulsion and surface tension before the ion leaves, followed by an induced dipole in the solvent once the ion has left.

used for the majority of experiments in this thesis instrument provides a measured $\sim 10 - 15\%$ relative standard deviation (RSD) in the ionic flow rate measured from scan-to-scan.

5.2 Linear ion trap mass spectrometry

The remarkable versatility of the linear ion trap for tandem mass spectrometric measurements [55] means this piece of hardware can now be found in many modern mass spectrometers. A linear ion trap mass spectrometer (Thermo Fisher Scientific LTQ XL [81]) was chosen for the initial development of pC-2DMS for a number of reasons: the high sensitivity should translate to a high fragmentation-to-measurement detection efficiency as required for successful covariance measurements [42], the fast scan rate should enable the time-efficient acquisition of the repeat spectra required for a covariance measurement, and the ubiquity of linear ion traps should allow for the immediate application of any techniques developed in a very large number of different laboratories. The high detection efficiency is partly due to the fact that in the linear ion trap mass spectrometer the trapping, isolation, fragmentation and measurement of ions all happens in the same place, eliminating potential losses during the shuttling of ions between different pieces of hardware. Fragment ion loss is also minimised in linear ion trap CID because the RF voltages applied (see below) continually focus the ions towards the centre of the trap [82].

In the LTQ XL, the molecular ions created under atmospheric pressure by electrospray or nanospray ionisation are transferred to the linear ion trap through the ion source interface and by a set of ion optics. A schematic of these ion transfer components, reprinted from [83], is shown in Fig. 5.3. The ion source interface is enclosed in a vacuum manifold which is pumped by two rotary vane pumps to a pressure of around 1.1 Torr ($\sim$1.5 mbar), and the linear ion trap itself is in a different vacuum manifold pumped by a turbo pump to $\lessapprox 1 \times 10^{-5}$ Torr (1.3×10^{-5} mbar). The ions are first drawn by a pressure gradient into and through an ion transfer capillary of length $\sim$10.2 cm, held at a constant temperature (200°C for the measurements in this

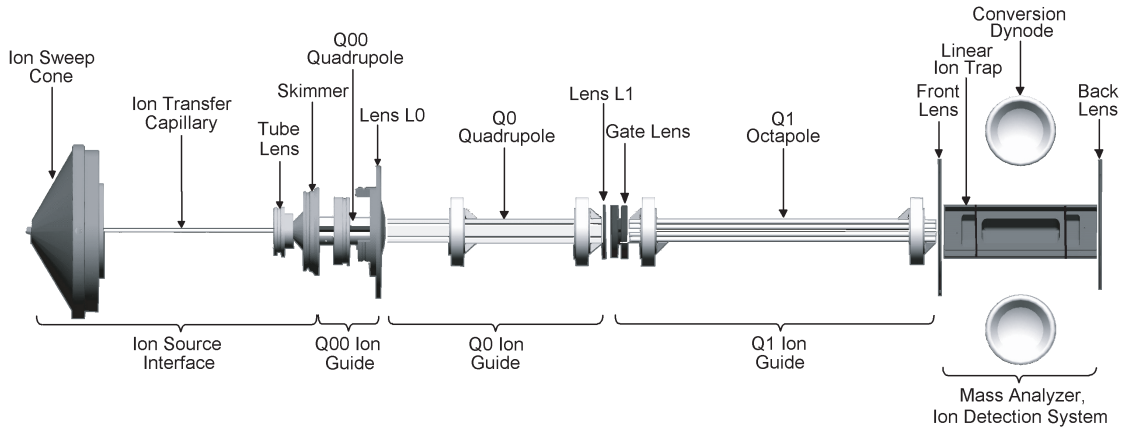


Figure 5.3: Schematic of the ion optics in the Thermo Fisher Scientific LTQ XL, reproduced from [83].

thesis) to encourage the desolvation of any excess solvent. A small (few volts) voltage of the same polarity as the ions is applied to this ion transfer capillary to encourage the transfer of ions down the pressure gradient and along the capillary, at the end of which they meet an electrostatic lens known as the tube lens. This focuses the ions onto a skimmer, which helps to form a collimated beam of ions and also acts as a vacuum baffle between the ion source interface and the ion optics further downstream. These optics are at a lower pressure of around 50 mTorr ($\sim$0.067 mbar) thanks to pumping from the third inlet in the Holweck backing stage of the turbo pump which pumps the linear ion trap itself. From the skimmer, the ions pass through a radio frequency (RF) lens, the Q00 RF quadrupole lens, which consists of a square array of square rods. As described in detail below[10], applying an RF voltage to these rods produces a dynamic electric field which focuses the ions along the axis between the electrospray source and the ion trap, and a small direct current (DC) voltage is also applied (~ -4 V for positive ions) to accelerate the ions towards the trap. Lens L0 is another electrostatic lens, which acts as a vacuum baffle between the ion source interface and lower pressure manifolds further downstream, and a small DC offset (~ -5 V for positive ions) is also applied to this to carry the ions towards the trap. The Q0 quadrupole and lens L1 act in much the same way as the Q00 RF lens and

[10]For hyperbolic rods; here the square shape of the rods results in contributions from higher-order multipoles whose effects are essentially inconsequential for the singular task of focusing the ion beam.

lens L0, respectively. The Q0 quadrupole lens is again a square array of square rods, which are longer than the rods found in the Q00 RF lens but which serve the same purpose of guiding the ions along the axis which separates the ion source interface and the ion trap. The DC offset voltage applied to the Q0 quadrupole accelerates the ions towards the trap, and the L1 lens is an electrostatic lens of the same design as the L0 lens (a metal disk with a small hole in the centre) to which an electrostatic voltage of magnitude up to 150 V can be applied to either accelerate or decelerate the ions as they approach the linear ion trap. Lens L1 also acts as a vacuum baffle. Finally, the Q1 octupole is an RF ion guide consisting of an octagonal array of circular rods which produce an approximate octupole field [55], further focusing the ion beam into the ion trap. The gate lens, which controls when the trap is 'open' to be filled with ions or not and so controls the Automatic Gain Control (AGC) function whose use is described in Chapter 6, directly precedes the octupole ion guide.

When the gate lens is open, and following transmission through the final Q1 octupole ion guide, the ions are transmitted to the linear ion trap. The basic geometry of the linear ion trap found in the LTQ XL [81] used for the vast majority of the measurements presented in this thesis is shown in Fig. 5.4. The trap consists of four hyberbolic[11] rods arranged in a square array. Each rod is cut into three sections, of 12, 37 and 12 mm in length, separated by insulating quartz spacers. Axial trapping is achieved through the creation of a potential well in the z-direction, by applying a voltage of -9 V, -12 V and -7 V (for positive ions) to the front, central and back sections respectively during ion storage. During mass analysis, the well is deepened by applying 0 V instead to both the front and back sections. The mass analyser, which is held at relatively high vacuum as described above, is filled with helium at a partial pressure of $\sim 1 \times 10^{-3}$ Torr (1.33×10^{-3} mbar). This acts as a buffer gas as the trap is filled and the ions entering the trap are slowed by multiple collisions with the helium atoms.

Radial trapping is achieved by a 2D quadrupole RF electric field. Neglecting

[11] The reason for this electrode geometry is explained below.

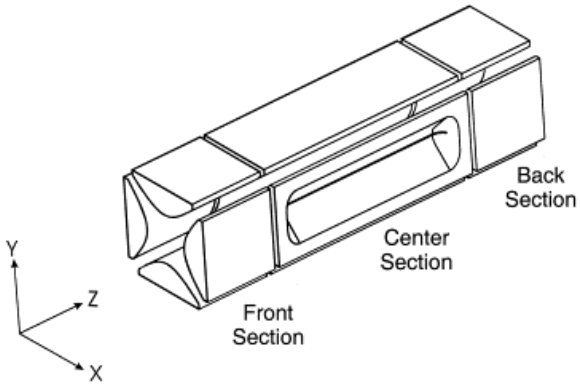


Figure 5.4: Geometry of the linear ion trap found in the Thermo Fisher Scientific LTQ XL. Reproduced with permission from [81].

the terms describing a constant potential and the potential from a line of charge[12], the multipolar expansion of an arbitrary 2D potential can be written in Cartesian coordinates as [84]:

$$\Phi(x,y) = \sum_{N=0}^{\infty} A_N \phi_N(x,y), \tag{5.1}$$

where A_N is the amplitude of the multipole and the terms $\phi_N(x,y)$ describe the shape of the potential and correspond to:

$$\phi_N(x,y) = \mathrm{Re}\left[(x+iy)^N\right]. \tag{5.2}$$

When $N = 0$, $\phi_0(x,y) = 1$ describes the shape of a constant potential throughout all space, and $\phi_1(x,y) = x$ is the shape of a potential between two parallel planes of opposite charge. The quadrupole potential which we are interested in here is described by $N = 2$, for which $\phi_N(x,y)$ can be written:

$$\phi_2(x,y) = \frac{x^2 - y^2}{r_0^2}, \tag{5.3}$$

where r_0 is a useful normalisation constant whose physical significance will become clear shortly. The general form of such a quadrupole potential is $x^2 - y^2 = a^2$, where a^2 is a constant. This condition is completely traced out in the 2D plane by

[12]The superposition principle for the electric field means these can be straightforwardly reintroduced later if required.

the set of hyperbolic curves parametrised by $(a\cosh t, a\sinh t)$, $(a\sinh t, -a\cosh t)$, $(-a\cosh t, a\sinh t)$ and $(a\sinh t, a\cosh t)$. A pure quadrupole potential $\Phi_2(x,y)$ can therefore be created by applying a voltage of amplitude A_2 to a set of four parallel hyperbolic electrodes with apices equally spaced a distance r_0 from the central axis. This is the reason for the hyperbolic shape of the electrodes in the trap of the LTQ XL, illustrated in Fig. 5.4. Thanks to the remarkable machining precision of the equipment used to fabricate this trap and others like it, higher-order multipole contributions due to manufacturing imperfections can be neglected in their theoretical treatment.

To radially trap the ions, the two pairs of opposing electrodes in the linear ion trap are connected to a split RF voltage source. The RF voltages applied to all four electrodes have the same frequency and amplitude, but one of the electrode pairs is set to be 180° out of phase with the other. If we write the time-varying potential $\Psi(t)$ applied to alternate pairs as $\Psi(t) = \pm(U - V_{RF}\cos\Omega t)$, where U is a constant DC offset and V_{RF} and Ω are the amplitude and frequency of the applied RF voltage respectively, the time-varying quadrupole becomes [55]:

$$\begin{aligned}\Phi_2(x,y,t) &= \phi_2\Psi(t) \\ &= \frac{x^2-y^2}{r_0^2}(U - V_{RF}\cos\Omega t).\end{aligned} \tag{5.4}$$

The force felt by an ion of mass m and charge z in this time-varying quadrupole field is:

$$q'\vec{E} = -ze\nabla\Phi_n(x,y,t), \tag{5.5}$$

where e is the elementary unit of charge and $q'\vec{E}$ is the electric Lorentz force felt by a charge q' in an electric field $\vec{E}$. The quadrupole potential described by Eq. 5.3 has the very useful property that it is separable in x and y[13], so the equation of motion in x, $m\frac{d^2\vec{x}}{dt^2} = \vec{F}$, for such an ion becomes:

[13]This is quite the opposite for all higher order multipole fields, for which motion in x and y is strongly coupled, and one of the reasons that contributions from higher order multipoles are keenly avoided by the precision machining of hyperbolic rods.

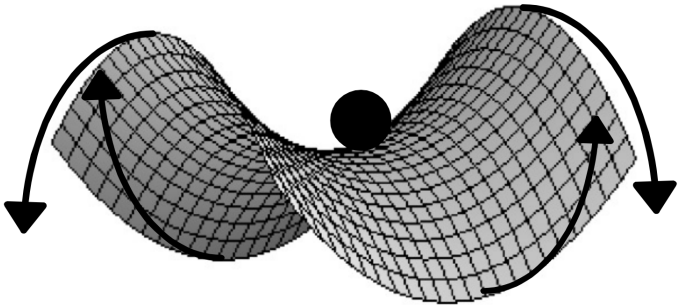


Figure 5.5: The shape of the potential seen by an ion in the xy plane is essentially a saddle whose sides bend upwards and downwards periodically. Reproduced with permission from [85].

$$\begin{aligned} m\frac{d^2\vec{x}}{dt^2} &= -ze\frac{\partial\Phi_n(x,y,t)}{\partial x} \\ &= -ze\frac{2\vec{x}}{r_0^2}\left(U - V_{RF}\cos\Omega t\right) \end{aligned} \tag{5.6}$$

and equally:

$$\begin{aligned} m\frac{d^2\vec{y}}{dt^2} &= -ze\frac{\partial\Phi_n(x,y,t)}{\partial y} \\ &= ze\frac{2\vec{y}}{r_0^2}\left(U - V_{RF}\cos\Omega t\right) \end{aligned} \tag{5.7}$$

for motion in y. The dynamic potential that an ion sees in the xy plane has the effective shape of the saddle shown in Fig 5.5, which is plotted close to a maximum of $\cos\Omega t$. As illustrated by the arrows, opposite sides of the saddle bend upwards and downwards in synchronisation, so that the every half period of the RF voltage the shape of the saddle effectively rotates by 90°.

To analyse the equations of motion defined above we can make a useful change of variables. Introducing:

$$a_x = -a_y = \frac{8zeU}{m\Omega^2 r_0^2};\ q_x = -q_y = -\frac{4zeV_{RF}}{m\Omega^2 r_0^2};\ \zeta = \frac{\Omega t}{2}, \tag{5.8}$$

Eqs. 5.6 and 5.7 can be rewritten as:

$$\frac{d^2x}{d\zeta^2} + (a_x - 2q_x \cos 2\zeta)\, x = 0 \tag{5.9}$$

$$\frac{d^2y}{d\zeta^2} + (a_y - 2q_y \cos 2\zeta)\, y = 0. \tag{5.10}$$

These are the Mathieu equations, first introduced by Émile Mathieu to describe the vibrations of an elliptical drumhead [86], and for which the solutions are well known. They are given by Floquet's theorem [87], which tells us that since we have a linear homogenous differential equation with a coefficient that is a periodic function in ζ [88], there exist solutions of the form:

$$x(\zeta) = \exp(i\gamma_x \zeta) P(\zeta) \tag{5.11}$$

and

$$y(\zeta) = \exp(i\gamma_y \zeta) P(\zeta), \tag{5.12}$$

where $P(\zeta)$ is a periodic function with the same period as $\cos 2\zeta$ (i.e. π) and γ is complex number which is a function of a and q. Provided γ_x is not an integer, $x(-\zeta)$ is also an independent solution (if it is an integer, one solution is simply proportional to the other), and *mutatis mutandis* for $y(-\zeta)$ [88]. Depending on γ, and therefore on a and q[14], these solutions can either grow unbounded or repeat periodically, with frequencies given by:

$$\omega_n = (2n + \beta)\frac{\Omega}{2};\ 0 \leq \beta \leq 1;\ n = 0, \pm 1, \pm 2, \ldots, \tag{5.13}$$

where β is again a function of a and q [55]. At low values of q, the largest amplitude motion is for the case $n = 0$, whilst greater q increases the contribution from higher n.

The fact that the solutions to the equations of motion can be either unstable (i.e.

[14] An example of the fairly involved calculation of the stability boundaries for these solutions, demonstrated for the case of solving the Schrödinger equation for weak coupling to a cosinusoidal potential, is found in [89].

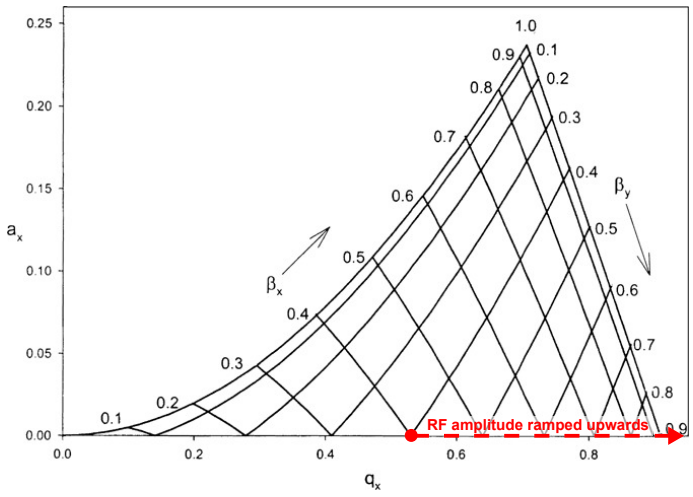


Figure 5.6: Top half of the first stability region ($a_x \geq 0$) for Eqs. 5.11 and 5.12. Reproduced with permission from [55]. The trajectory of the stability conditions during the ion ejection stage (dot and arrow in red, see below) has been added by the author.

grow unbounded) or stable results in so-called stability regions in aq space, where the range of values of a_x,a_y and q_x,q_y which give stable solutions for the 2D motion of an ion of a given m/z can be plotted. The stability regions for motion in x and y overlap in four areas on such stability diagrams, of which the so-called first of these (the one at smallest q) represents the usual operating regime for an RF linear quadrupole. The shape of the top half ($a_x \geq 0$) of this region is plotted in Fig. 5.6, reproduced with permission from [55]. If the position of an ion on the aq diagram is fixed, as its m/z increases the stability region maintains its shape but its area increases, as the apex is drawn upwards and to the right. As can be seen from Eqs. 5.8, increasing a corresponds to an increased DC offset U and increasing q corresponds to increasing amplitude V_{RF} of the RF trapping voltage. When operating as a trap and mass analyser, as for the experiments presented in this thesis, there is typically no DC offset applied and so a is equal to zero. The normalised values of β for motion in the x and y directions (from Eq. 5.13) at a given point in aq space are also shown.

The change in the dimensions of the stability region for ions of different m/z, which is also exploited by the quadrupole mass filter [90], lies at the heart of the successful use of linear ion traps as mass analysers. During the trapping phase in the operation of the LTQ XL, the amplitude of the RF voltage (which is held at a

constant frequency of 1.2 MHz) is maintained at a value so that all ions above a certain m/z ratio (chosen to be above the m/z values associated with abundant contaminant species from air and solvents) have a stable trajectory in the trap. Collisions with the buffer gas in the trap reduce the axial kinetic energy of the ions to aid trapping by the DC voltages on the end electrodes and, provided they are not close to the stability boundaries, they also lose radial kinetic energy so that they move closer to the centre of the trap where the potential is lower (as can be seen in Fig. 5.5). The focusing of the ions to the centre of the trap by the buffer gas greatly increases the sensitivity and mass resolution of the linear ion trap. Once the ions under analysis are trapped, they are either scanned out to measure their m/z values and relative abundances (Full MS scan), or isolated, fragmented and either have their fragment ions measured or submitted to further rounds[15] of isolation and fragmentation (MS^n scans).

To measure the m/z of the ions in the trap, the amplitude of the RF voltage applied to the quadrupole is ramped upwards. As this happens, the trajectories of ions of successively increasing m/z become unstable in the trap as the applied value of q slips over the right hand boundary of the stability region. This is illustrated in red in Fig. 5.6: the red dot denotes the value of a and q in the ion trap at a given time for an ion of a given m/z value, which moves to the right and eventually outside of the stability region as the amplitude of the RF voltage is ramped over time. As the trajectories of the ions become unstable, the amplitude of their motion exceeds the spatial limits of the trap, and many of the ions leave the trap by the 0.25 mm $\times$ 30 mm slits cut into the horizontal rods (so-called exit rods) of the centre section of the trap shown in Fig. 5.4. The efficient ejection of these ions through the two exit rods is aided by the application of an additional alternating voltage to the two exit rods during the ramping of the main RF voltage, with opposing exit rods at 180° out of phase with each other. Like the main RF voltage, this additional so-called resonance ejection voltage is applied at increasing amplitude and constant frequency, which is calculated so that only ions which are just about to be ejected from the trap fall into

[15]Up to eight further rounds are possible with the LTQ XL.

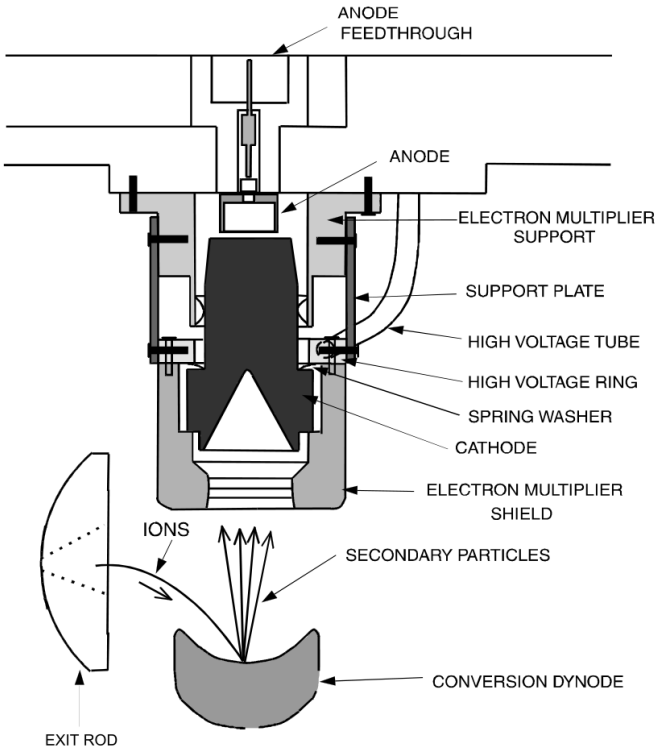


Figure 5.7: Illustration of the ion detection system in the LTQ XL, reproduced from [83]. Following ejection from the linear ion trap *via* the exit rods, the ions are attracted towards the conversion dynode and strike it, creating a stream of secondary particles. These secondary particles initiate electron cascades in the conical cathode which eventually produces a detectable current, collected by the anode and of magnitude proportional to the number of secondary particles emitted by the conversion dynode.

resonance with it.

As illustrated in Fig. 5.7, the ions which leave the trap through the slits in the exit rods are drawn onto two metallic, concave surfaces known as conversion dynodes, facing 90° to the initial direction of ejection of the ions. For the detection of positive (negative) ions, these conversion dynodes are held at −15 kV (15 kV). The off-axis configuration reduces the noise from neutral species leaving the trap. The ejected ions are attracted to the conversion dynodes, and upon striking the dynode surface initiate the release of either negative ions and electrons for the measurement of positive ions, or positive ions for measurements in negative ion mode. These secondary products are repelled and focused by the curved surface of the conversion dynode towards the cathode of the electron multiplier detection system which is stationed directly

opposite. This cathode consists of a lead oxide cavity with walls tapered inwards, held at a negative voltage of up to -2.5 kV to create an attractive potential gradient across from the conversion dynode for measurements in either polarity. The secondary particles produced from the conversion dynode strike the tapered walls of the cathode, initiating a cascade of electrons which are eventually collected by an anode, held near to ground potential, at the end of the multiplier. The current detected at the anode, which is proportional to the number of secondary particles reaching the cathode, gives the measurement of the abundance of ejected ions at different times (and therefore different m/z values). The cascade effect is produced by the repeated creation of secondary electron bunches as the released electrons strike the continually encroaching walls of the cathode. Overall, the electron multiplier is set to a gain of roughly 10^5, so that every secondary particle striking the cathode produces on average $\sim 10^5$ electrons. The function of the conversion dynodes is to redress the insensitivity of the electron multiplier to both negative ions and high mass ions [91].

The implementation of MS^n scans begins with the isolation of the parent ion of interest. This is typically desired because it provides more sensitive and straightforward identification of the isolated parent ion from the fragment spectrum, without contamination from the fragment ions of other dissociated parent ions. To achieve parent ion isolation in the LTQ XL, additional alternating voltages to the main RF trapping voltage are applied across the exit rods, at 180° out of phase with each other. This voltage is a sweep of sinusoidal RF waveforms with a Fourier spectrum spanning between 5 and 600 kHz with a small notch cut into it, such that it contains all the resonant frequencies according to Eq. 5.13[16] for all the trapped ion m/z's except the ions to be isolated. At this stage the net effect of the collisions with the buffer gas can be modelled as a drag force which is proportional to the ratio between the ion speed and thermal speed of the gas: linearly at low ions speeds, and quadratically at higher ion speeds [92]. This means that the resultant averaged radial motion of all the ions falling into resonance with this additional applied voltage is similar to the behaviour of a damped driven harmonic oscillator; the amplitude of motion initially

[16] Although not explicitly specified by the manufacturer for the LTQ XL, typically for $n = 0$ [55].

increases linearly before eventually reaching a constant value [55]. The amplitude of the resonant ejection voltages is sufficient to cause the amplitude of motion of the ions to be ejected to increase beyond the confines of the trap, and they are lost, primarily by ejection through the slits in the exit rods.

The remaining isolated ions are subsequently subjected to CID. To perform this, an alternating voltage is again applied across the exit rods, which are again 180° out of phase with each other, but this time the voltage is at the resonant frequency of the motion for the isolated ions. An advantage of this so-called resonant excitation is that fragment ions produced from the isolated parent ions do not fall into resonance with this excitation voltage, and are therefore likely to remain intact until measurement. This reduces the complexity of the fragment spectra and preserves the structurally informative primary[17] fragment ions, but can in some cases be problematic if the parent ion suffers an efficient neutral loss and so falls out of resonance with the excitation voltage before it has accumulated sufficient internal energy to produce structurally informative fragment ions[18]. The operating value of q during this step determines the position of the deleterious so-called 'low mass cut-off' (LMCO); at a given value of q the operating regime falls outside the stability region for ions below a certain m/z so that any fragment ion produced below this m/z value is lost and cannot be measured [81]. The typical operating parameters for a linear ion trap mass spectrometer give the low mass cut-off at around $\frac{(m/z)_{Par}}{3}$, where $(m/z)_{Par}$ is the m/z value of the resonantly excited parent ion [93]. The application of the resonant excitation voltage causes the amplitude of the ionic oscillations to initially linearly increase in the radial x-direction until it reaches a steady state, but this time the amplitude of the voltage is not sufficient to eject the resonantly excited ions from the trap. As the amplitude of the oscillations and so radial kinetic energy of the ions increases, they experience a greater number of collisions with the buffer gas at higher collision energies. After a number (typically hundreds) of inelastic collisions, during which the transfer of centre-of-mass frame kinetic energy, E_{CoM}, to internal energy

[17] i.e. directly resulting from the parent ion decomposition, without secondary fragmentation

[18] This is the case for the CID measurements in e.g. Figs. 6.4 and A.5.

of the ion can approach 100% [92, 94], the internal energy of the ion is sufficient to induce fragmentation.

5.3 Collision-induced dissociation

There exist a number of different methods for so-called 'ion activation', i.e. inducing the fragmentation of biopolymers for tandem MS analysis [95]. These include electron transfer dissociation (ETD) [32], electron capture dissociation (ECD) [33] and numerous different forms of photo-induced fragmentation [34]. Because the technique of pC-2DMS described in this thesis is based on the principle of partial covariance mapping, the successful application of which relies almost exclusively on the instrumental detection efficiency, once high enough detection efficiency is assured [42] it should be applicable to any ion activation method. However, the initial development described in this thesis has all taken place using what undoubtedly remains the most popular form of ion activation: collision-induced (or -activated) dissociation, CID (or CAD) [16, 96], whose implementation in a linear ion trap is described above. This technique involves collisions between the biopolymeric ion under analysis and an inert gas, resulting in the transfer of some of the kinetic energy of the colliding species into internal energy of the ion under analysis and eventual fragmentation. There are, moreover, a number of different ways to implement CID, providing different collisional energies and different observed fragmentation chemistry. For example, the beam-type collisional activation employed in higher-energy collisional dissociation (HCD)[19] [97] avoids one of the main drawbacks of conventional CID (namely, the low mass cut-off) and produces different fragmentation patterns, providing complementary sequence information to the standard lower energy technique implemented in an ion trap. The work presented in this thesis was performed using linear ion trap CID.

The implementation of CID in a linear ion trap such as that found in the LTQ XL is discussed in detail in the previous section. Briefly, the kinetic energy of the ions trapped by the RF quadrupole field is further increased by the application of

[19]Originally christened higher-energy C-trap dissociation.

the resonance excitation AC voltage to the exit rods (the rods which sit horizontally opposite each other), and after many collisions with atoms of the small partial pressure of helium gas in the trap ($\sim 10^{-3}$ Torr) the internal energy of the ions is sufficiently high to induce fragmentation. Numerous studies of the kinetics of ion trap CID have demonstrated a strong dependence of CID behaviour on analyte, mass of the colliding gas, pressure of the colliding gas, and many other experimental parameters [82, 98–102]. Typical parameters for the ion trap CID described in this thesis consist of a laboratory-frame energy, E_{Lab}, in the range of a few (∼1–20) electron volts for each collision, with efficient (many tens of percent) energy transfer from the centre-of-mass collisional energy[20] to internal ion energy meaning dissociation occurs after hundreds of collisions and on a timescale of several tens of milliseconds.

After a sufficient number of collisions, a protonated peptide ion is excited to a high enough internal energy to induce fragmentation. This value is typically several to several tens of eV, with the internal energy of the ensemble of trapped ions typically following a Boltzmann-like distribution [82]. A peptide consists of a chain of amino acids (sometimes referred to as 'residues'), and in much the same way that all words in the English language are constructed from 26 different letters, the overwhelming majority of all naturally occurring peptides are constructed from only ∼20 different amino acids. The distinction between a peptide and a full protein is loosely defined, but is dependent on the number of amino acids in the chain; the shorter peptide typically consists of 10–15 amino acids whilst an intact protein can be formed from many hundreds of residues.

The basic structure of an amino acid itself is illustrated in Fig. 5.8. The type of amino acid is determined by the functional group R^n bonded to the α-carbon, and different amino acids (e.g. glycine) are commonly denoted by a three-letter or single-letter code (e.g. Gly, or G). In a peptide, the chain of amino acids is joined together by HN–CO amide bonds, known as peptide bonds, as illustrated for the tetrapeptide

[20]This represents the maximum amount of translational energy that can be transferred to internal energy of the ion in a collision according to momentum conservation, and for the rough approximation of the collision of an ion of mass m_1 with a stationary neutral atom of the buffer gas with mass m_2 [94] it is given by $E_{CoM} = \frac{m_2}{m_1+m_2} E_{Lab}$.

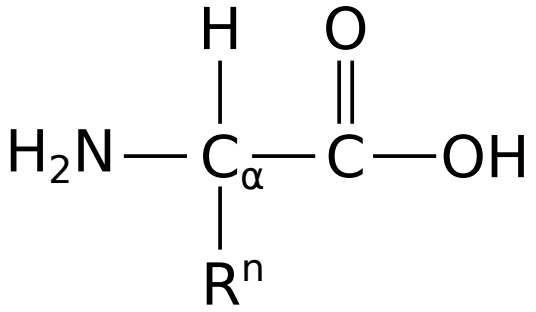


Figure 5.8: An amino acid. R^n is the functional group, which is attached to the C_α and determines the type of amino acid. The smallest amino acid is glycine (G), for which R^n is a hydrogen atom.

in panel **A** of Fig. 5.9. Four different amino acids, differentiated by their functional groups R^n, are joined by the peptide bond between the carboxyl group of amino acid n (which has lost OH from the illustration in Fig. 5.8) and the amino group of amino acid $n+1$ (which has lost H). The second of the two dominant resonance forms of the peptide bond involves a HN=CO double bond, which means the bond has a planar character [103]. This has significant biological implications because it limits the degrees of freedom available upon protein folding. The overwhelmingly favoured conformation of the peptide bond in native folded proteins is the trans isomer [104], i.e. consecutive functional groups R^n appear on opposite sides of their respective α-carbons in the plane of the peptide bond separating them. The set of bonds drawn horizontally in panel **A** are collectively referred to as constituting the peptide 'backbone' and the two termini are known as the N-terminus (left hand side, beginning with NH_2) and the C-terminus (right hand side, ending in COOH) respectively.

The dominant fragmentation channels for ion trap CID of protonated peptides, which describes the majority of the measurements presented in this thesis, produce the fragment ions whose typical structures are illustrated in panel **B** of Fig. 5.9. It should be noted that these conformations are inferred from model studies of simple molecules [105] and although they serve usefully as standard models it is to be expected, and has been observed, that isomeric versions of these fragment ions can also exist in

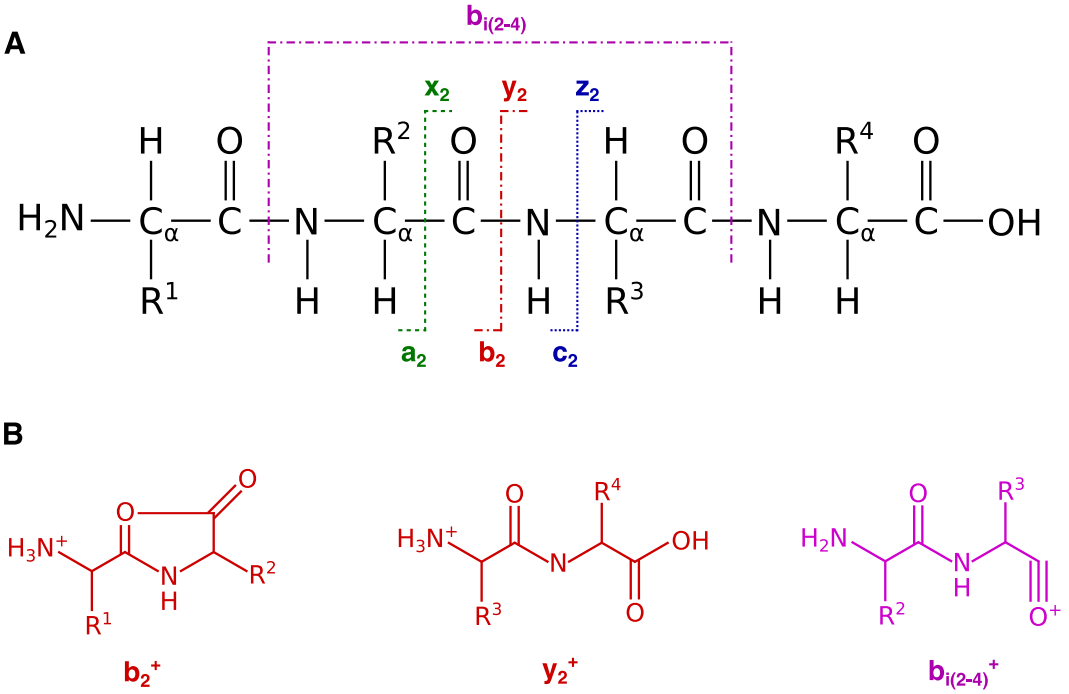


Figure 5.9: **A**, a tetrapeptide and **B**, its characteristic fragment ions produced under ion trap CID.

other structures [2, 25]. Cleavage of the peptide bond produces so-called b- and y-ions, according to the standard nomenclature detailed in [27], where the b-ion includes the N-terminus and the y-ion the C-terminus. Depending on whether the proton(s) remain on the N-terminus, C-terminus, or both, stable and detectable b-ions, y-ions, or both (a 'complementary pair') are produced. Different b-ions are identified by a subscripted number which gives the number of amino acid residues they include, counting from the N-terminus, whilst different y-ions are identified in the same manner but counting from the C-terminus. The additional hydrogen gained by a y-ion has migrated from a nitrogen which was N-terminal to the cleaved amide bond on the intact peptide ion [106]. Secondary cleavage of an additional peptide bond can produce a b-type internal ion[21], denoted $b_{i(n\text{-}m)}$, which includes residues (counting from the N-terminal end) n to m. Along with b- and y-ions and their associated derivatives, the b-internal ion (and its own derivatives) is the most

[21]These internal ions are much less commonly used for peptide and protein sequencing in 1D MS/MS than terminal b- and y-ions because of the high false positive rate they give for fragment ion matching [107], see Chapter 10 for details.

common type of fragment ion observed under ion trap CID [108]. Cleavage of the N–C_α bond, producing so-called c- and z-ion fragments, is rarely observed for CID but is the dominant fragmentation pathway for the electron-based activation techniques ETD and ECD [108]. Cleavage of the C_α–CO bond, producing a- and x-ions, is observed under some fragmentation schemes, such as ultraviolet photodissociation (UVPD) [95]. The a-ion is commonly produced under ion trap CID through loss of a stable CO neutral from the b-ion, with the mechanism proposed in [109] featuring elimination of the $C^{(5)}$=O from the oxazolone ring of the cyclic b-ion (top right of the b_2 ion in Fig. 5.9) following weakening of the O–CO bond after protonation of the $N^{(3)}$ at the bottom of the oxazolone ring. The b-type internal ion can also lose a stable CO neutral under ion trap CID to form an a-type internal ion. All fragment ions can also undergo neutral loss of small stable molecules H_2O and NH_3. Loss of H_2O is most common for fragments involving serine (S), threonine (T), glutamic acid (E) and aspartic acid (D) because these are easily lost from the side-chain[22] (functional group R^n), and the same applies to NH_3 loss[23] for fragments containing arginine (R), lysine (K), asparagine (N) and glutamine (Q). Note that the cyclic nature of the typical b-ion structure means the b_1 ion is not usually observed in standard ion trap CID experiments.

Fragmentation mechanism The so-called 'charge-directed'[24] pathways which predominantly produce the fragment ions illustrated in panel **B** [25] are understood in the general framework of the 'mobile proton model' [24, 110–112]. Within this model the proton(s) on the protonated peptide, which are added in acidic solution to the biomolecule under analysis prior to ESI, are naturally initially localised on the basic sites of the peptide. These are typically the side chains of the basic residues (in order of decreasing basicity) arginine (R), lysine (K) and histidine (H), or the N-terminal

[22]Although peptide fragments without these residues can also undergo loss of water through rearrangement reactions [108].

[23]Which is also commonly observed from y-ions [2].

[24]So-called because the cleavage happens at the site where the charge (proton) is localised on the molecule, in contrast to 'charge-remote' pathways which do not require the charge to be localised to the position of the cleavage [24].

amino group [113]. The mobile proton model asserts that the important effect of the increase in the ion's internal energy during collisional activation is to provide the proton(s) on the gaseous biomolecular ion with the energy required to overcome the potential barriers associated with sampling different protonation sites, and that protonation at the site of a peptide bond initiates its cleavage. Molecular orbital calculations [114, 115] show that in a comparison of the two most favoured protonation sites of the peptide bond, protonation of the amide oxygen, which is energetically favoured over protonation of the amide nitrogen, increases the bond order[25] of the peptide bond. In contrast, it was found that protonation of the amide nitrogen, whilst not energetically favoured, reduces the bond order. Moreover, when the proton is localised here the carbon atom of the protonated amide group becomes vulnerable to nucleophilic attack[26] from electron-rich groups nearby, which can induce cleavage of the peptide bond.

Creation of b- and y-ions, as well as secondary b-type internal ions, can take place through a number of different competing pathways in a fragmenting protonated peptide. The known fragmentation pathways producing b- and y-ions are thoroughly examined in reference [25]. Although there exist mechanisms which proceed through the energetically favoured protonation of the amide oxygen [25], fragmentation *via* the nitrogen-protonated reactant is thought to dominate in low-energy CID, particularly in the case of shorter peptides, because of the high barrier protecting the transient reactive structures along the oxygen-protonated pathway [117]. Nucleophilic attack on the carbon centre of the protonated amide bond, initiating its cleavage, can come from the oxygen of the N-terminal neighbouring amide bond (b_x-y_z pathway) [116] or a nucleophilic group on a side chain [particularly arginine (R), asparagine (N), glutamine (Q), histidine (H) and lysine (K)] neighbouring the protonated amide bond. The b_x-y_z pathway is thought be the dominant pathway for the low-energy CID experiments performed in this thesis, and is the pathway by which the cyclic b-ion in

[25]Bond order is a measure of bond strength which broadly denotes the number of chemical bonds joining two atoms. There exist various different theoretical and experimental methods for its calculation, and it can assume a non-integer value.

[26]The donation of an electron pair to a positively charged atom, which usually forms a new bond and displaces a leaving group.

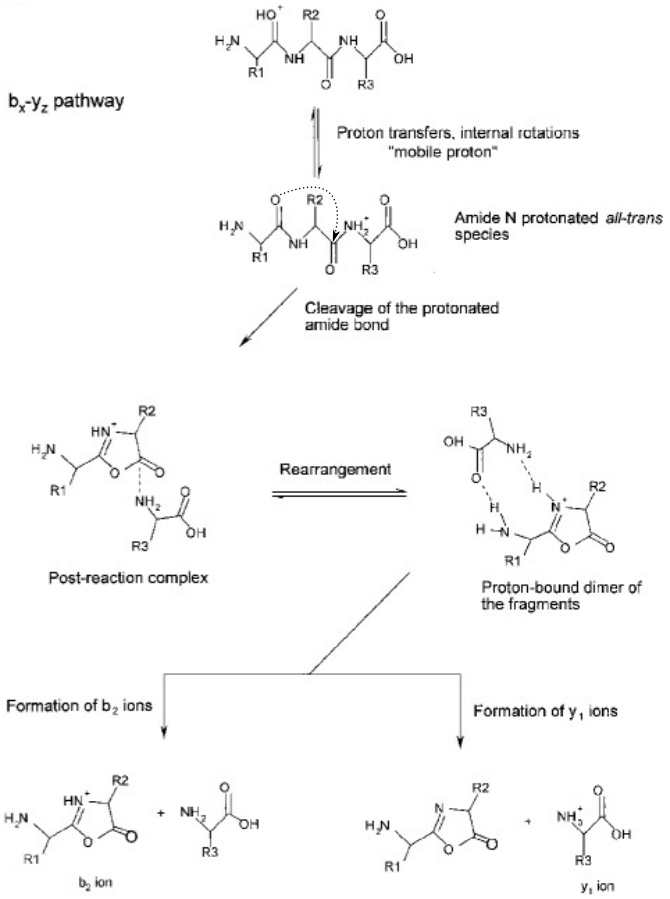


Figure 5.10: The b_x-y_z pathway for the cleavage of a peptide bond, reproduced with permisison from [116]. The dotted arrow on the amide N-protonated species, illustrating nucleophilic attack by the C=O oxygen from the first base on the carbon centre of the N-protonated amide bond, has been added from the original.

panel **B** of Fig. 5.9 is known to be formed [116]. The b-type internal ions, requiring cleavage of at least two different peptide bonds, can be produced through secondary cleavage of both b-ions and y-ions [118].

The general b_x-y_z fragmentation scheme is shown in Fig. 5.10 (reproduced with permission from [116]) for the particular example of the singly protonated sequence GGG dissociating into b_2-y_1. In this figure, the proton begins on the O of the carboxyl group of the N-terminal residue before being 'mobilised' by the internal energy increase thanks to multiple collisions. Nucleophilic attack of the carbon centre of the N-protonated amide bond comes from the oxygen of the C=O from the N-terminal neighbouring peptide bond. This results in cleavage of the protonated peptide bond, producing a loosely-bound complex between the two resultant fragments which has a sufficiently long lifetime to enable it to rearrange to a proton-bound dimer. The thermochemistry of the two proton-bound fragments (i.e. their relative proton affinity) determines the relative abundance of different fragment ions - on dissociation of the proton-bound dimer, the fragment with the highest proton affinity tends to hold on to the proton which initiated the cleavage and mediates the dimer [25, 119]. The existence of more than one proton on the precursor ion can result in the production of a complementary pair of detectable b- and y-ions from the same parent molecule [2], which we would expect to be able to measure in covariance measurements. Once more, it should be noted that this b_x-y_z pathway, whilst expected to be the dominant initial fragmentation pathway for the production of the ions most abundantly observed under the typical experimental conditions employed in this thesis, is one of a large number of competing dissociation pathways for activated ions under CID [25]. A more complete understanding of the fragmentation chemistry of protonated peptides and other ionic biomolecules, including the analytically important task of measuring and predicting the branching ratios for different dissociation pathways, remains an active pursuit. It is hoped the techniques developed in this thesis may be able to play a role in furthering the experimental exploration of these complex mechanisms.

5.4 Experimental parameters for pC-2DMS measurements

The pC-2DMS measurements presented in this thesis were all performed at the following experimental parameters, unless otherwise stated. For the measurements on peptide samples, synthetic peptide crystals were either purchased from suppliers or kindly provided by Rüdiger Pipkorn (German Cancer Research Centre, Heidelberg). Full protein and oligonucleotide samples were purchased from suppliers. The peptide samples were dissolved in 50% acetonitrile/2% formic acid in water at concentrations of ∼1 µM to ∼10 µM for 1D CID MS/MS measurements and 100 fM to ∼1 µM for pC-2DMS measurements, respectively. In most cases, the impurity of the crystals used precluded precise determination of the concentration. For oligonucleotide experiments, the sample crystals were dissolved in 0.1% ammonium acetate in water to produce a concentration of 10 µM for positive ion mode measurements, and in 49% acetonitrile/2% triethylamine in water at a concentration of 20 µM for negative ion mode measurements. For top-down measurements, the intact protein samples were prepared to a concentration of ∼1 µM in a solution of 1% formic acid and 50% acetonitrile in water.

The MS measurements were performed on an LTQ XL (Thermo Fisher Scientific) linear ion trap mass spectrometer. The sample solutions described above were directly infused using a Harvard Apparatus 11 Plus Single Syringe Pump coupled to a 'Nanospray II' Ion Source (Thermo Fisher Scientific) at a set flow rate between 3 and 5 µl/min (1 µl/min for protein samples) and set spray voltage of between 1.8 and 2.2 kV in positive ion mode, and at a flow rate of 1 µl/min and spray voltage of −1.5 kV in negative ion mode. The flow rate, spray voltage and exact position of the ion source emitter were manually adjusted at each measurement for optimal spray stability. No auxiliary desolvation gas was used for the electrospray, and the temperature of the ion transfer capillary was held constant at 200°C. The parent ions of interest were isolated in the linear ion trap and fragmented by CID at normalised collision energies of 35% for oligomers and 70% for intact proteins, at 30 ms

activation time and fragmentation q-value 0.25. The MS/MS scans were performed at scan rates of 16 666 Da/s for 1D measurements and 125 000 Da/s for pC-2DMS measurements (i.e. 'Turbo mode'). The choice of scan mode for the pC-2DMS measurements is discussed in Chapter 18. For 1D measurements the default AGC MS^n target of 10^4 was used, and for pC-2DMS measurements the AGC MS^n target was 10^2 for pure oligomers and 3×10^2 for intact proteins, mixtures of oligomers and mixtures of intact proteins. The AGC MS^n target was reduced to roughly two orders of magnitude below its default target for 1D MS/MS measurements in order to augment the scan-to-scan fluctuations for optimal covariance measurements, see section 7.2. Lower sample concentrations were used for pC-2DMS measurements in order to ensure the injection time determined by the automatic gain control function of the LTQ XL did not fall to the minimum possible injection time of 1 µs, given the lower AGC MS^n target values used. No averaging over microscans was performed in the pC-2DMS experiments. The experimental parameters used for the measurements in this thesis are summarised in Table 5.1 below.

In order to perform the covariance analysis, the '.raw' files produced by the LTQ XL were converted to text files using the Xcalibur File Converter software provided by the instrument manufacturer. A script written in Python then converted these text files to '.npy' NumPy files, which were stored and used for the data analysis.

Sample solution	50% acetonitrile /2% formic acid/48% water (peptides) 50% acetonitrile /1% formic acid/49% water (intact proteins) 0.1% ammonium acetate/99.9% water (oligonucleotides, positive ion mode) 49% acetonitrile/2% triethylamine/49% water (oligonucleotides, negative ion mode)
Sample concentration	~1 µM – ~10 µM (peptides, 1D MS measurements) 100 fM – ~1 µM (peptides, pC-2DMS measurements) ~1 µM (intact proteins) 10 µM (oligonucleotides, positive ion mode) 20 µM (oligonucleotides, negative ion mode)
Direct infusion flow rate	3–5 µL/min (peptides) 1 µL/min (intact proteins & oligonucleotides)
Electrospray source voltage	1.8–2.2 kV (positive ion mode) -1.5 kV (negative ion mode)
Ion transfer capillary temperature	200°C
AGC MS^n target	10,000 (1D MS measurements) 100 (pC-2DMS measurements on single peptides & oligonucleotides) 300 (pC-2DMS measurements on mixtures & intact proteins)
CID normalised collision energy	35% (70% for intact proteins)
CID activation time	30 ms
Scan rate	16,666 Da/s (1D MS measurements) 125,000 Da/s (pC-2DMS measurements)
Average over microscans	1

Table 5.1: Table summarising the experimental parameters used for the measurements presented in this thesis, unless stated otherwise. The sample flow rate and electrospray source voltage were set at precise values in the given ranges, and manually optimised at each measurement for the best spray stability. In most cases, the impurity of the sample crystals used precluded precise determination of concentrations.

Chapter 6

Foundations of pC-2DMS

6.1 Total ion count partial covariance

Initial efforts in the development of pC-2DMS were directed towards the calculation of simple covariance maps across repeated MS/MS spectra on different ensembles of the same parent ion, to identify whether the signals from fragment ions which were expected to be created in the same decomposition reaction shared positive covariance. To begin with, the 1D MS/MS spectra were measured, under standard ion trap CID experimental parameters, for roughly twenty different protonated peptide molecules. The primary function of these experiments was to identify those CID spectra which featured a large number of b- and y-type complementary ions. It is well known and ubiquitously observed that under ion trap CID, peptide molecules of length n protonated by $z \geq 2$ protons can dissociate to form a complementary ion pair of b_a^{i+} & y_b^{j+}, where $a + b = n$ and $i + j = z$ [2]. Because these complementary pairs can be formed in the same fragmentation reaction of the same parent molecule, and because b- and y-type fragment ions are generally the most commonly formed in ion trap CID [2], it was expected that they would correlate to produce observable areas of positive covariance on the corresponding covariance map. To carry out this investigation software was developed, which would later form the basis for a large number of further applications, that accepts as input a peptide sequence (complete with any modifications) and parent peptide ion charge state and produces a table of

the expected m/z values for all possible b- and y-type ions formed.

Of the measured spectra, it was found that the complementary pairs being sought were most readily observed from the triply protonated ions of tryptic peptides (meaning those that had, or at least could, be formed from the tryptic digest of a full protein). This is understood by the fact that the trypsin enzyme cleaves after one of the basic residues arginine (R) or lysine (K), meaning tryptic peptides typically have a strongly basic C-terminal residue. A triply protonated peptide will very likely be protonated at this basic C-terminal residue as well as the basic N-terminal amine group. This leaves one of the protons free to sample different protonation sites across the peptide molecule and induce b/y type cleavage, primarily *via* protonation of one of the amide nitrogens of a backbone peptide bond (see section 5.3). Because both termini are protonated, both the b- and y-type fragments are charged and therefore detected. One such triply charged protonated peptide is the human histone H3.1 fragment [^{118}VTIMPK$_{\mathrm{Ac}}$DIQLAR^{129}], K$_{\mathrm{Ac}}$ = N$^{\varepsilon}$-acetyllysine (peptide sequence **1**), for which the 1D mass spectrum is shown in Fig. 6.1. The 1D mass spectrum shows six abundant complementary b/y pairs, rendering the peptide ion a strong candidate for covariance analysis. The fundamental principles and characteristics of pC-2DMS are explained herein using this ion as an example.

In order for the covariance between two variables to be non-zero their values must grow and shrink in some synchronisation with each other, which requires the values to change from scan-to-scan. In one-dimensional mass spectrometry, scan-to-scan repeatability of a mass spectrum is considered important to give a reliably reproducible method, and to this end the functional design of mass spectrometry experiments minimises the scan-to-scan variations in the abundance of each ion. The primary functionality for minimising scan-to-scan fluctuations in the LTQ XL borrows its name from an analogous process in electronics, known as 'Automatic Gain Control' or AGC. The purpose of AGC in electronic circuits is to maintain a steady amplitude for the output signal despite fluctuating amplitude in the input signal. The AGC function in the LTQ XL attempts to provide a consistent number of trapped parent ions to be fragmented at each scan (this value is the event rate ν in Chapter 4),

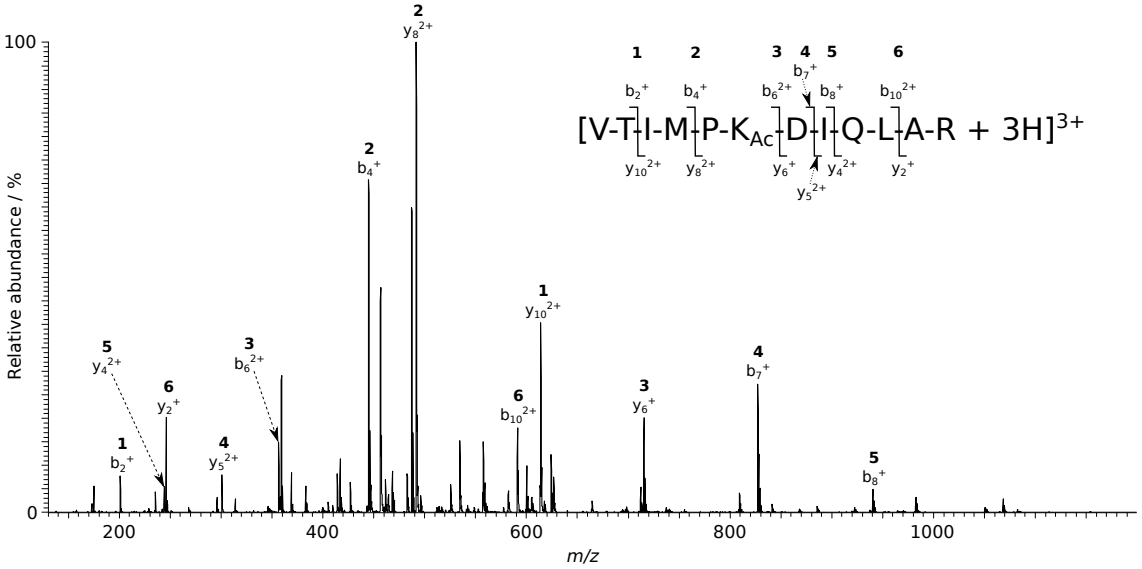


Figure 6.1: 1D tandem mass spectrum, measured in a linear ion trap, of the triply protonated ion of peptide **1** under CID at NCE 35%.

despite the inevitable fluctuations in the ion current[1] arriving to the trap during an MS/MS measurement. Moreover, the AGC function also attempts to trap as many parent ions as possible at each scan before the onset of deleterious space-charge effects in the trap [55], which minimises the scan-to-scan variations in fragment ion relative abundances due to the stochastic outcome of competition between all different fragmentation pathways. In the LTQ XL a target parameter for the AGC is specified by the user, which corresponds to approximate number of ions trapped at each shot [54]. The AGC works by briefly sampling the ion current flowing into the trap at a time just before the injection of ions for analysis, by opening the trap for 50 ms and measuring the signal from the accumulated ions. It approximates that this ion current will be the same as that flowing when the trap is subsequently opened to accumulate ions for analysis, and so performs a simple linear calculation based on the measured ion current to find the time for which the trap should be opened in order to accumulate the desired number of ions for analysis. The fluctuations in ion current are primarily a result of the inherent instability of an electrospray ionisation

[1]This is terminology that is widely used in the mass spectrometric community to denote a number of ions that have been measured, trapped, etc. Each ion is counted once regardless of the charge it is carrying, so the term 'current' can be misleading and should not be taken to be dependent on charge state as a measure of electrical current would be.

source [80]: a typical measured value for the nanospray ionisation source used for the majority of experiments in this thesis is $\sim 10 - 15\%$ relative standard deviation (RSD).

The AGC function was expected to prove useful for covariance measurements by minimising the potential for common-mode correlations due to fluctuating ion density in the trap and to provide a globally repeatable central event rate for covariance measurements. However, the AGC target for MS^n scans ($= 1 \times 10^4$) is set, by default, to trap the maximum possible number of ions at each scan in order to minimise scan-to-scan fluctuations in relative fragment ion abundances. Because significant scan-to-scan fluctuations of abundances in different signal channels is required for high signal-to-noise on the covariance map, it was predicted that for covariance mapping a lower number of ions was required to be trapped at each scan. A series of different CID measurements at different AGC target values was therefore performed and the covariance maps for these spectra were calculated. All calculations were performed in the Python language. The raw files provided as output by the Thermo software were converted into text files, which were then read and saved in the NumPy .npy format.

These simple covariance maps were scanned for peaks of positive covariance in the expected positions, given the known structure of the ions under analysis. The highest volume features on these maps was found to be for scans taken at AGC MS^n target value $\sim$100, so the scans at this target value were chosen for initial further investigation (a later, more thorough study of the optimal event rate for pC-2DMS is detailed in section 7.3). The first simple covariance maps showed a large number of islands of positive covariance. However, they also presented the vertical and horizontal lines (indicating correlation of all spectral signals with each other) characteristic of maps dominated by spurious common-mode fluctuations induced by external fluctuating parameters, such as that in panel (c) of Fig. 3.2.

For these MS/MS CID experiments, this presented a significant problem. The common-mode correlations are induced by the fluctuations of a number of different, non-coupled experimental parameters (e.g. electrospray ionisation source stability, ion trap helium pressure, ion trap voltages), each fluctuating in their own manner from

scan-to-scan. Unlike e.g. a laser pulse intensity, the scan-to-scan monitoring of any of these experimental mass spectral parameters is impractical, precluding the standard procedure for suppression of spurious common-mode correlations of monitoring the fluctuating parameter or parameters and applying Eq. 3.1 or 3.2.

To overcome this problem, a novel and general method for partial covariance mapping was developed that overcomes the requirement of scan-to-scan monitoring of each fluctuating experimental parameter. Instead of monitoring every fluctuating parameter from scan-to-scan, the combined effect of all experimental parameters is approximated to the integral of the entire fragment mass spectrum (often known as the 'total ion current/count', TIC) and this value is applied as a single fluctuating experimental parameter using Eq. 3.1:

$$pCov(\mathbf{Y},\mathbf{X};TIC) = Cov(\mathbf{Y},\mathbf{X}) - \frac{Cov(\mathbf{Y},TIC)Cov(TIC,\mathbf{X})}{var(TIC)}. \qquad (6.1)$$

Fig 6.2 shows the difference between a small region of the simple covariance map (calculated with Eq. 2.2) and the TIC partial covariance map (calculated with Eq. 6.1) of the fragmentation of the triply charged peptide molecule **1** under CID. The simple covariance map in panel **a** shows positive covariance islands between all spectral signals. All annotated signals are correlations of different fragment ions with the y_8^{2+} fragment ion at m/z 491.8, of which all those labelled in blue correlate two fragment ions which cannot both be created in the same fragmentation reaction of the same parent ion. Correlations a), b), and k) correlate two ions from the same peptide terminus, correlations c), d), f) and g) correlate two ions that include one or more of the same residues and correlation e) correlates the full parent ion minus a neutral loss of two water molecules with the entire y_8^{2+} fragment ion. Subtraction of the correction term $Cov(\mathbf{Y},TIC)Cov(TIC,\mathbf{X})/var(TIC)$ yields the TIC partial covariance map in panel **b**, in which all these unphysical correlations have been supressed, leaving only the three islands labelled in red. Each of these three islands correlates the y_8^{2+} ion with a fragment which is expected to be produced in the same decomposition pathway of the parent ion. Each is produced by the cleavage of the

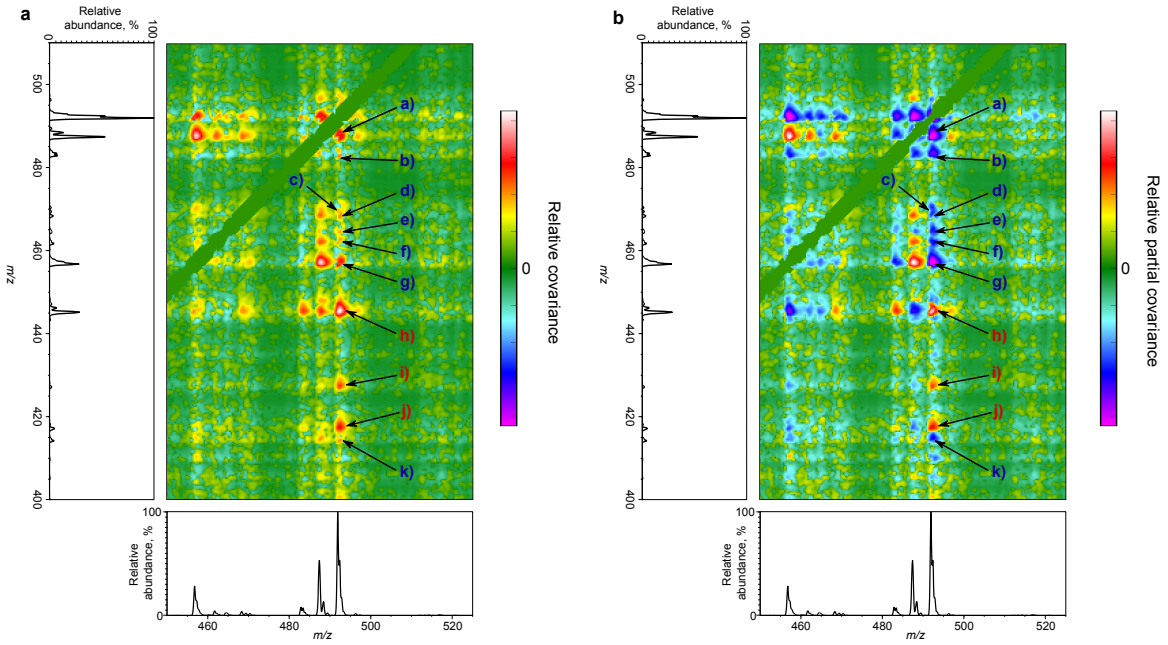


Figure 6.2: **a**, 2D view of a (m/z 450−525×m/z 400−510) region of the 2D simple covariance map of CID on triply protonated peptide **1**, constructed using Eq. 2.2. The corresponding region of the 1D CID spectrum is shown along the x- and y-axes. Annotations denote correlations between the following ions: a) y_8^{2+} & y_4^+, b) y_8^{2+} & $[y_8\text{-}H_2O]^{2+}$, c) y_8^{2+} & b_8^{2+}, d) y_8^{2+} & $a^+_{i(5-8)}$, e) y_8^{2+} & $[M+3H\text{-}2H_2O]^{3+}$, f) y_8^{2+} & $[b_8\text{-}H_2O]^{2+}$, g) y_8^{2+} & a_8^{2+}, h) y_8^{2+} & b_4^+, i) y_8^{2+} & $[b_4\text{-}H_2O]^+$, j) y_8^{2+} & a_4^+, k) y_8^{2+} & $[y_1\text{-}2NH_3]^+$, where M is the neutral peptide. **b**, 2D view of the same region of the TIC partial covariance map of peptide **1**, constructed using Eq. 3.1. The autocorrelation line has been removed from both maps for visual clarity.

8th peptide bond along the backbone (counting from the N-terminal end), followed by either no neutral loss [h)], neutral loss of water from the N-terminal b-ion [i)] or neutral loss of CO from the N-terminal b-ion [j)].

The use of the total ion count as a single fluctuating parameter for the application of Eq. 3.1 has been theoretically investigated and verified in collaboration with Bridgette Cooper (University College London) and Serguei Patchkovskii (Max Born Institute for Nonlinear Optics and Short Pulse Spectroscopy, Berlin), see Chapter 8.

Figure 6.3 shows a surface plot of the full two-dimensional pC-2DMS map of triply protonated peptide ion **1**. The full map shows a multitude of positive covariance island signals, many of which provide structural and dynamic information which is, as a matter of principle, not available from the 1D mass spectrum. The map's vertical axis represents the relative value of the partial covariance and the horizontal axes

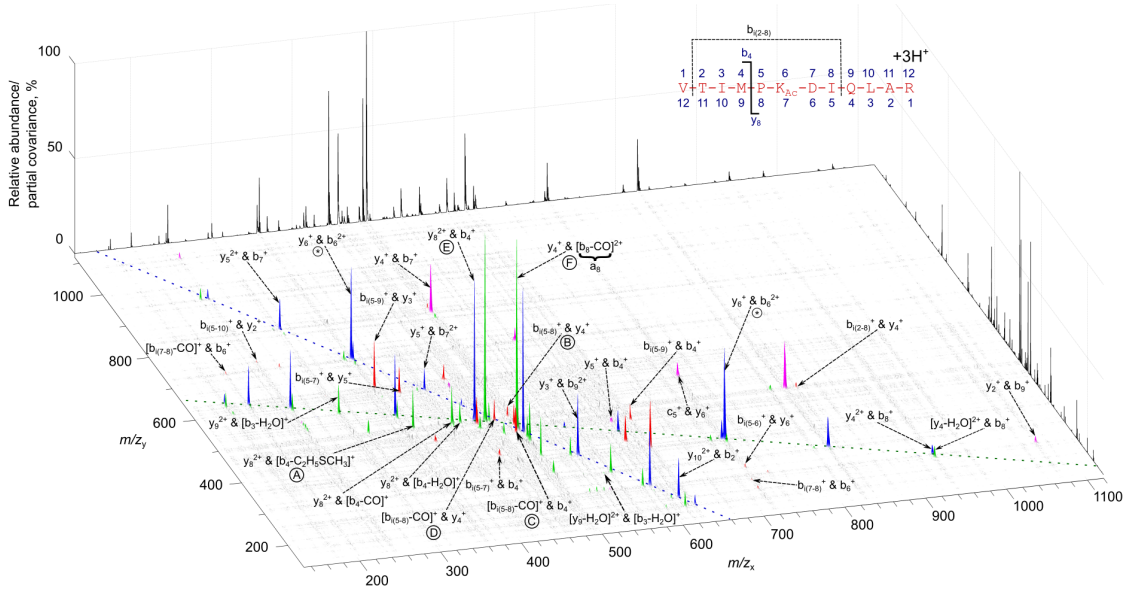


Figure 6.3: 2D partial covariance map, measured in a linear ion trap, of the triply protonated ion of peptide **1** under CID at NCE 35%. The different colours of the peaks indicate different classes of fragment ion correlations, as described in the text. The dashed lines demonstrate the conservation of mass in the fragmentation process.

m/z_x and m/z_y display the m/z ratios of the correlating fragment ions. The black line graph plotted against the back walls of the map is the mean-averaged 1D MS/MS CID spectrum as shown in Fig. 6.1. The autocorrelation line, which trivially correlates every spectral signal to itself, has been removed for clarity and the off-diagonal peaks reveal correlations between pairs of peptide fragments (X, Y) of shared origin, i.e. generated in the same or consecutive molecular decompositions. Such pairs have been christened 'siblings' (Z $\rightarrow$ X+Y), 'niblings' (Z $\rightarrow$ X+Y_1, $Y_1 \rightarrow X_1$+Y), 'cousins' (Z $\rightarrow$ X_1+Y_1, $X_1 \rightarrow$ X+X_2, $Y_1 \rightarrow$ Y_2+Y), etc. To the author's knowledge, this represents the first experimental demonstration that products produced in consecutive decomposition reactions can be identified by covariance analysis. Selected pC-2DMS peaks are annotated and the peptide fragment nomenclature used, as in [27], is given in the inset. Apart from the two y_6^+ & b_6^{2+} peaks (marked by *) , the pC-2DMS peaks are annotated only on one side of the autocorrelation diagonal for clarity. The labels A–F mark the correlations discussed below.

The physical origin of the correlation signals on the pC-2DMS map can be directly inferred from the geometrical position at which they appear on the map. Some of this

information is demonstrated in the colour-coding of the signals shown in Fig. 6.3. Blue signals are the correlation signals of pairs of complementary (sibling) fragments generated by fission of a single particular bond, which are arranged along the so-called mass conservation diagonals. Analogous to the momentum-conserving lines observed on the covariance maps in e.g. Fig. 2.1, correlations between complementary pairs of b/y ions arrange themselves along lines of the form $m_1/z_1 \times z_1 + m_2/z_2 \times z_2 = Mr_{Par}$, where Mr_{Par} is the mass of the parent ion. Intuitively, mass conservation diagonals can be understood by observing that for a parent ion of a particular mass dissociating into two products A and B (which will be correlated because they are produced in the same reaction), the larger in mass (and so m/z) fragment A is, the smaller the m/z value of fragment B will be. The two mass conservation diagonals are shown by the dashed lines on the map – $Mr_{Par} = m/z_x \times 1 + m/z_y \times 2$ (green) and $Mr_{Par} = m/z_x \times 2 + m/z_y \times 1$ (blue). This feature allows the deduction of the charge states of fragment ions within pC-2DMS using only the gradient of the mass conservation diagonals on which their pC-2DMS correlation signals fall, see Chapter 13 for a detailed discussion.

Plotted in green are the correlations of fragments originating from consecutive decompositions through neutral loss (NL) of small molecules (i.e. $Z \rightarrow X+Y_1$, $Y_1 \rightarrow Y+NL$), causing horizontal and vertical peak series relative to the corresponding pC-2DMS signals of the intact ions (e.g. y_8^{2+} & $[b_4\text{-CO}]^+$). Plotted in red are the manifold of scattered signals revealing correlations between terminal b- and y- fragments and terminus-free internal fragments [118] and their derivatives, such as $b^+_{i(2-8)}$ & y_4^+ and $[b_{i(5-8)}\text{-CO}]^+$ & y_4^+. And in magenta are the correlations between non-complementary internal fragments where there has been a loss of one or more internal residues, e.g. y_2^+ & b_9^+.

The pC-2DMS correlations provide direct experimental evidence for the formation history of the fragment ions. For example, the identity of the peptide fragment at m/z 369.2, generated by an uncommon methionine-specific neutral loss of methylthioethane, can be readily determined as $[b_4(\text{VTIM}) - C_2H_5SCH_3]^+$ based on its correlation with the complementary ion y_8^{2+} (PK_{Ac}DIQLAR), see signal A in Fig. 6.3.

Another example is the internal fragment $b^+_{i(5-8)}$ $(PK_{Ac}DI)^+$ at m/z 496.3, which can be unambiguously identified thanks to the correlation of itself and its CO-loss derivative with its corresponding N- and C- terminal fragments (b^+_4 (VTIM) and y^+_4 (QLAR)), see signals B, D and E. The origin of these two terminal fragments can in turn be deduced from their correlations with their own complementary fragments, and derivatives (signals E and F).

6.2 pC-2DMS correlation score

Despite the success of using the total ion count as a single partial covariance parameter to remove spurious extrinsic correlations from the covariance map, it was found that some pC-2DMS maps presented true correlation features between two ions created in the same fragmentation process[2] which were lower and less voluminous than features which could be readily determined as unphysical (e.g. features correlating two m/z values that summed to a value larger than the mass of the parent ion under analysis). Through comparison with prior experience and previous formal analysis [42][3], these false correlations were deemed to be the result of statistical noise. This drove efforts to find a metric that was able to differentiate low-volume, true pC-2DMS correlations from higher volume false ones due to statistical noise.

To this end, the use of statistical resampling was investigated as a method to test the quality of the inference from the processed data to the partial covariance features appearing on the pC-2DMS map, and so identify statistical noise. Resampling is a method for testing the precision of sample statistics, such as covariance, by recalculation of the statistic upon some rearrangement of the original dataset (such as random re-drawing of scans with replacement [bootstrap] or systematic selection of a subset [jackknife]). The distribution of the statistic of interest across different resamples provides an estimate of the sampling distribution in the absence of additional measurements.

[2]These appeared at expected positions according to the known structure of the parent molecule under analysis.

[3]This was later corroborated by the results of the formal analysis in Chapter 8.

In jackknife resampling, a linear approximation of the more rigorous bootstrap method, one or more samples are sequentially eliminated from the dataset and the sample statistic is calculated for each of the resultant subsets. For example, for a set of n different samples (these could be a set of MS/MS scans), the statistic may be calculated for each of the n subsets of $n-1$ scans produced by the elimination of the i^{th} scan with $i = 1, 2, 3, \ldots, n$. In bootstrap resampling, each resample is produced by randomly drawing from the original samples with replacement. For the purpose of estimating the sampling distribution of the partial covariance the jackknife method (so-called because of its general applicability to a vast range of tasks) was chosen, because it is a) less computationally expensive and b) produces fully replicable results because of its systematic nature. Bootstrap resampling was tested to determine whether its use improved performance, but it gave no discernible difference in the quality of results over the jackknife method.

To test the likelihood that a pC-2DMS correlation corresponds to a true intrinsic correlation instead of statistical noise, the pC-2DMS correlation score, $S(x, y)$, was introduced for a feature correlating two m/z values x and y. This score is calculated by determining the standard deviation of the volume of each partial covariance feature upon jackknife resampling, and normalising the volume of the given feature as calculated from the full set of scans to this standard deviation:

$$S(x, y) = \frac{V(x, y)}{\sigma(V(x, y))}, \tag{6.2}$$

where $V(x, y)$ is the volume of the partial covariance feature calculated for the full number of scans and $\sigma(V(x, y))$ is the standard deviation of this volume across the n resamples of $n-1$ scans constructed by the sequential elimination of each scan. It was found, as demonstrated below, that this score, which is highly non-linear with feature volume or height, provides a reliable measure of the likelihood that a feature on the pC-2DMS map corresponds to a true intrinsic correlation feature. Work to provide a formal treatment of this metric remains outstanding; although the quantity expressed in Eq. 6.2 bears resemblance to Student's t-test [120], a full

characterisation has so far not been performed. Nonetheless, the general method of normalising a sample statistic to a estimate of the standard deviation of its sampling distribution[4] is ubiquitously employed and robust, and as illustrated below the pC-2DMS correlation score performs very well in singling out true pC-2DMS correlation islands of small volume from larger volume islands due to statistical noise. The pC-2DMS correlation score has been tested across a large number of different samples (see Table 7.1) and has been found to provide consistent robust performance, over other metrics such as feature volume or height, in singling out low volume true correlations (see Fig. 6.4), reducing the number of scans required to extract reliable signals from the pC-2DMS maps (see section 7.3) and as a measure to weight the contribution of pC-2DMS features to the pC-2DMS search engine score (Chapter 11).

Shown in Fig. 6.4 is an off-diagonal region (m/z $170-600 \times m/z$ $560-1000$) of the pC-2DMS map from the fragmentation under CID of the doubly protonated ion of the sulphated peptide RDsYTGW-Nle-DF-NH_2, sY=sulphotyrosine (peptide **2**). The standard averaged 1D MS/MS spectrum is plotted along the back walls and is seen to be dominated by the $[M\text{-}SO_3H]^{2+}$ peak, due to efficient elimination of the labile sulphur trioxide group from the parent ion during CID. Whilst the majority of the map shows very little statistical noise, the m/z $575-600 \times m/z$ $560-1000$ area marked by the dotted line shows a multitude of scattered high-volume pC-2DMS features representing false correlations (blue). These are primarily a result of statistical noise involving the dominant $[M\text{-}SO_3H]^{2+}$ spectral peak, and are straightforwardly identified as spurious in this test case because the structure of the analysed peptide is known (for example, many of the correlations fall outside of the mass conservation line [dashed dark green line] and so violate the law of mass conservation).

The inset scatter plot demonstrates the successful identification of the true correlation island y_4^+ & b_5^+ from the higher-volume statistical noise surrounding it in the small indicated region (marked out by dashed line), using the pC-2DMS correlation score. The correlation score for all identified features in the marked region is

[4] in other words, its standard error

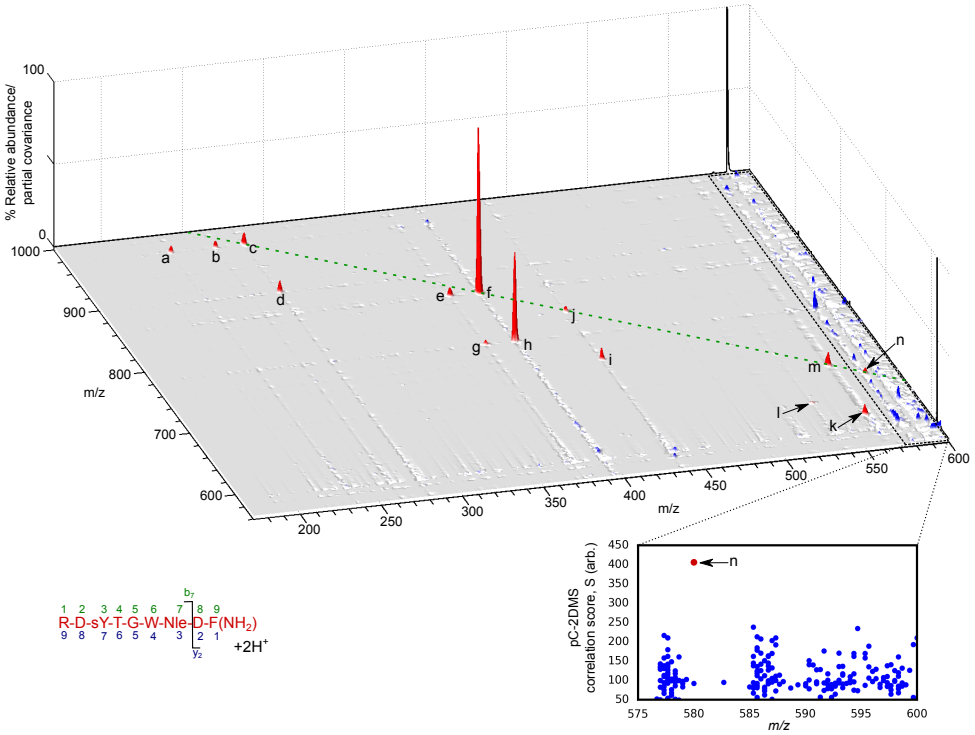


Figure 6.4: m/z 170 − 600 × m/z 560 − 1000 region of the pC-2DMS map from the fragmentation under CID of the sulphated sequence RDsYTGW-Nle-DF-NH_2, sY=sulphotyrosine. Plotted along the back walls is the averaged 1D MS/MS spectrum from the same fragmentation, which is dominated by the $[M+2H\text{-}SO_3]^{2+}$ peak resulting from the facile loss of sulphur trioxide from the parent ion. The horizontal axes are the m/z value of the correlating fragments, the vertical axis of the map is the relative value of the partial covariance, and the vertical axis of the 1D spectra along the back walls is the measured percentage relative abundance of the fragments. The dashed dark green line is the 'mass conservation line' for the doubly charged parent ion (i.e. the line along which correlations between two intact fragments of the parent ion will fall). The inset demonstrates the identification of the true, low volume correlation signal n (in red, y_4^+ & b_5^+) from amongst a large number of high-volume features due to statistical noise (blue), by virtue of its outstanding pC-2DMS correlation score.

Label	Identity
a	$[y_2\text{-}HCONH_2]^+$ & b_7^+
b	$[y_2\text{-}NH_3]^+$ & b_7^+
c	y_2^+ & b_7^+
d	y_2^+ & $[b_7\text{-}SO_3]^+$
e	$[y_3\text{-}H_2O]+$ & b_6^+
f	y_3^+ & b_6^+
g	$[b_6\text{-}SO_3]^+$ & $[y_3\text{-}H_2O]^+$
h	$[b_6\text{-}SO_3]+$ & y_3^+
i	y_6^+ & $[b_3\text{-}SO_3]^+$
j	$[y_3\text{+}HNCNH]^+$ & $[b_6\text{-}HNCNH]^+$
k	z_4^+ & $[c_5\text{-}SO_3]^+$
l	y_5^+ & $[b_4\text{-}SO_3]^+$
m	z_4^+ & c_5^+
n	y_4^+ & b_5^+

Table 6.1: Identity of labelled features in Fig. 6.4

calculated and plotted on the scatter plot, which allows the true correlation to be immediately separated from the statistical noise thanks to the outstanding value of its score. The identities of the labelled features are given in Table 6.1. Note that the sensitivity of pC-2DMS enables the identification of very low abundance intact sulphated fragment ions which have not lost the suphur trioxide group [a), b), c), e), f), j), m), n)], greatly assisting in the localisation of the post-translational modification under CID.

Chapter 7

Properties of pC-2DMS

Having established the fundamentals of pC-2DMS, this chapter describes some of the key properties of the technique.

The problem of identifying structurally informative fragment ion signals from chemical or even electrical noise in a fragment ion spectrum remains a significant challenge for the analysis of the tandem mass spectra of biomolecules. In 1D MS/MS the obvious discriminatory parameter available is the measured relative abundance of the fragment ions. Although this does generally discriminate from electrical noise, selection of signals based on their abundance suffers from two primary drawbacks: a) many sequence-informative fragment ions appear at extremely low abundance and so are discarded as noise by such procedures and b) many uninformative signals, such as small-molecule neutral losses from the parent ion, can appear at high relative abundance and therefore suppress structurally informative, less abundant signals. Moreover, the significant variation in absolute measured abundances between different MS/MS scans, due to efficiency of electrospray ionisation, competition between decomposition pathways, discrepancies and artefacts in detector response, etc. [121], precludes an absolute threshold which would be capable of providing a global identifier of (likely) structurally informative 1D signals across different measurements. Instead, fragment signal selection in 1D MS is performed by varied implementations of an intra-spectrum comparison of abundances.

By requiring the correlation of two detectable fragment ions to produce a spectral

signal, pC-2DMS is able to provide a more specific identifier of structurally informative signal. For example, the technique favours two low abundance complementary ions (such as standard b/y sequence ions in CID) over abundant neutral loss from a parent ion (which produces only one detectable ion). This is well illustrated by comparison of the pC-2DMS map and 1D MS/MS spectrum (plotted along the back walls) in Fig. 6.4[1] (both spectra have been normalised according to the largest spectral signal in the full measured m/z range). Moreover, as described in section 6.2, even low-volume features (low volume corresponds to low-abundance of correlated ions [42]) are able to be identified as true correlation features by calculation of the pC-2DMS correlation score for each feature. In accordance with the standard approach, an intra-spectrum comparison of the pC-2DMS correlation score for different features is currently employed. It has been found that a simple ranking works well, negating the requirement for a more complex ranking scheme requiring e.g. accounting for m/z-dependent detection efficiencies, and a combination of automatic and manual interpretation has shown that typically the top 40 (for 2+ parent ions) and top 50 (for 3+ parent ions) ranked signals are all true correlation signals.

7.1 Absolute value of pC-2DMS correlation score

Because calculation of the pC-2DMS correlation score (see Eq. 6.2) requires the normalisation of the peak volume $V(x, y)$ to its own standard deviation $\sigma(V)$ upon jackknife resampling, it could be reasonably expected that the pC-2DMS correlation score is insensitive to the varying absolute abundance of fragment ions recorded across different measurements. This is because the standard deviation $\sigma(X)$ directly scales with the value of the X, such that $\sigma(aX) = |a|\sigma(X)$ for any constant a. If this is indeed the case, the pC-2DMS correlation score could provide an absolute threshold for true correlation signals across different measurements, which would e.g. be able to account for the varying number of structurally informative fragment ions produced by different fragmenting parent ions. This could also be used as a robust

[1] see also e.g. Fig. A.5

absolute measure to determine which pC-2DMS correlation signals to include for automated sequence identification by database search engines, such as the one described in Chapter 11.

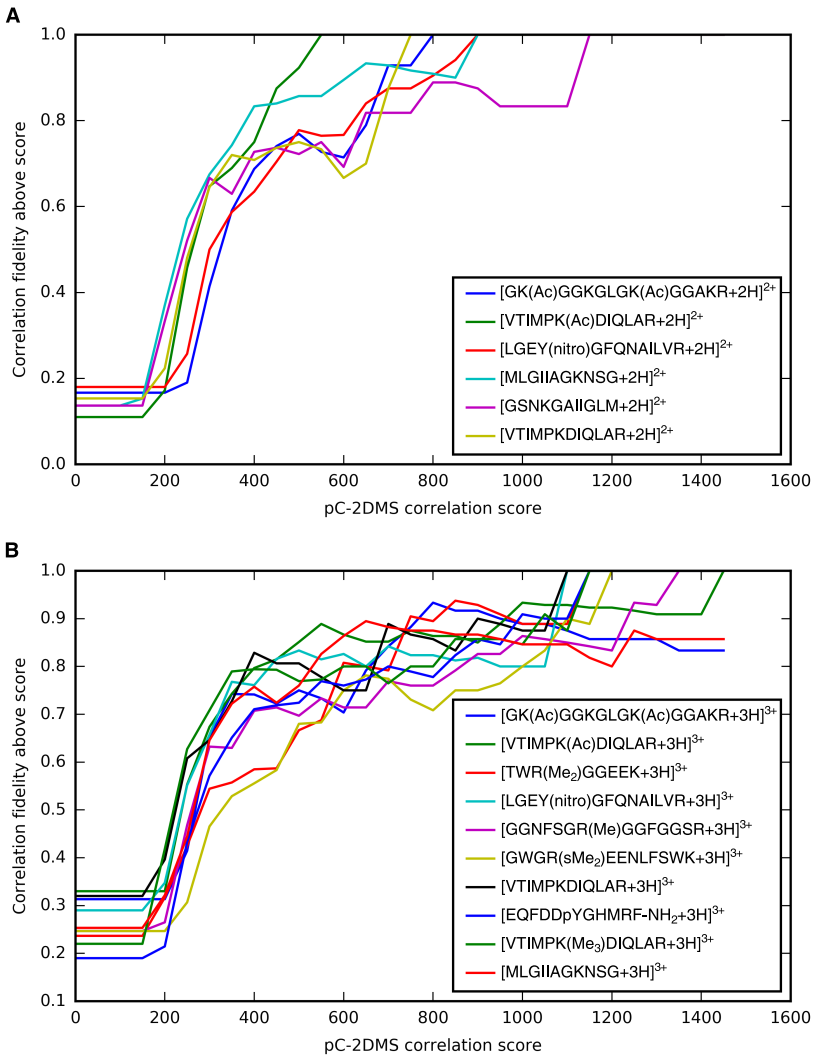


Figure 7.1: The fraction of pC-2DMS correlation features identified as canonical fragment ion correlations by the automated correlation matching procedure ('Correlation fidelity', y-axis) is shown for all correlations scoring above a particular pC-2DMS correlation score (x-axis), for measurements on different 2+ (panel **A**) and 3+ (panel **B**) parent ions. The absolute value of the correlation score at which true correlations emerge is roughly uniform across measurements on different peptide ions.

In Fig. 7.1, the top 300 pC-2DMS correlation scored features from each scan have been selected and subjected to the automated fragment ion matching software which was developed for the pC-2DMS search engine described in Chapter 11. Panel **A** is

for measurements on 2+ parent ions and panel **B** for 3+ parent ions. The absolute value of the correlation score is plotted on the x-axis and plotted on the y-axis is the 'fidelity', as determined by the automatic correlation matching software, of all the correlation signals scoring above this value x. The fidelity was calculated as the fraction of pC-2DMS correlations scoring higher than x which were identified as canonical fragment ion correlations by the automated fragment correlation matching[2]. These plots suggest the emergence of an absolute scale for pC-2DMS correlation fidelity (for example, the measurements all provide 75% fidelity above a pC-2DMS correlation score of ∼500), but further characterisation is evidently needed. The primary requirement for such an investigation will be a large number of additional measurements on different peptide ions. In Fig. 7.2, the two plots in Fig. 7.1 have been overlaid to determine the difference between measurements with a 2+ parent ion (blue) and a 3+ parent ion (red), showing consistent behaviour across the two parent charge states. Identification of an absolute threshold for the pC-2DMS correlation score would doubtless provide a useful feature, but, as demonstrated by the absence of such a threshold for 1D analysis, it does not prevent an obstacle to the successful application of pC-2DMS.

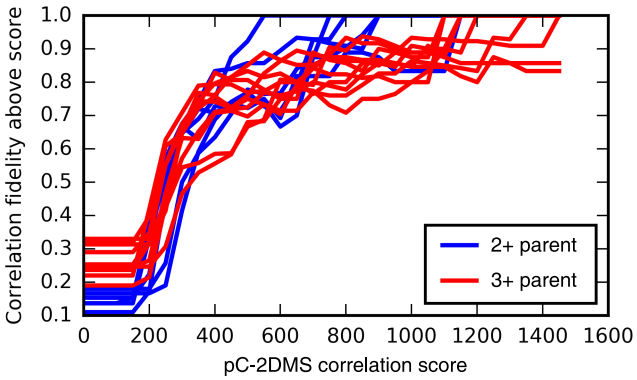


Figure 7.2: Comparison of dependence of correlation fidelity on pC-2DMS correlation score for 2+ parent ions (blue) *vs.* 3+ ions (red), for all measurements on shown in Fig. 7.1. The absolute value of the correlation score at which true correlation signals begin to appear is independent of the different charge states.

[2]Manual analysis would be expected to identify further true correlations between non-canonical fragment ions, as in e.g. Figs. 6.3, A.1, A.3.

7.2 Event rate

Following initial development using the roughly determined AGC MS^n target (= central event rate ν_0) of 100, an investigation of the optimal event rate for pC-2DMS measurements was performed. Figure 7.3 shows the dependence of the number of extracted true pC-2DMS correlation signals on the AGC MS^n target value setting for pC-2DMS measurements on the triply protonated peptide $[VTIMPKDIQLAR+3H]^{3}+$ (peptide **3**) at NCE 12.5%[3] and scan rate 125 000 Da/s. To create this figure, the manual analysis which was originally performed on these scans has been superceded by more thorough automatic analysis using the correlation matching software developed for the pC-2DMS database search engine (see Chapter 11). The general result is the same as that provided by the manual analysis, but automated analysis allows for the plotting of more data points and so is employed here. The number of pC-2DMS correlation signals from the top 50 pC-2DMS correlation-scoring features which are automatically identified at a standard fragment ion tolerance of 0.8 Da is plotted against the AGC target value setting (1D default 10^4) for the measurement from which the features are extracted, for feature lists calculated using different numbers of scans (see section 7.3 for further investigation of the dependence of pC-2DMS signals on number of scans). Consistent with the very rough initial optimisation in Chapter 6, the number of identifiable correlation features peaks at an AGC target of $\sim 10^2$, two orders of magnitude lower than the default setting for the instrument. The behaviour across different AGC target values shows the same general shape (albeit with greater variance at higher numbers of scans) at any number of scans.

The lower number of identifiable correlation signals at low AGC target values can be attributed to the extremely low count in the fragment ion signal channels at low AGC MS^n target values. When detection efficiency is less than 100% (which it always is in practice) very low signals counts cause the correlated product $\langle XY \rangle$ to fall to zero or almost zero, resulting in negligible covariance and partial covariance. The lower number of identifiable correlation signals at high AGC target values is, in

[3]This NCE was manually optimised for complete fragmentation of the parent ion under analysis, and before the general suitability of NCE 35% for pC-2DMS measurements was verified.

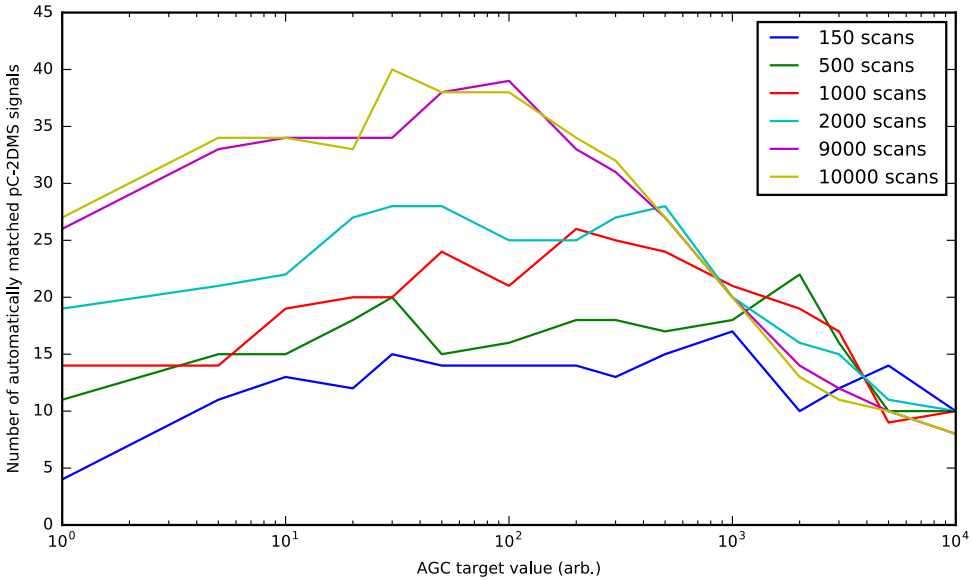


Figure 7.3: Number of automatically identified pC-2DMS correlation features as a function of AGC target value. The number of pC-2DMS correlation signals from the top 50 pC-2DMS correlation-scoring features which are automatically identified at fragment ion tolerance of 0.8 Da is plotted against the AGC MS^n target value setting (default 10^4) for the correlation feature list generated using different numbers of scans. All measurements performed on the triply protonated ion of peptide **3**. The optimum is found to be at a target value of around 100 ions trapped at each scan.

contrast, a result of the very small scan-to-scan fluctuations in the signal intensity in each ion channel, usually a desirable effect in 1D MS and one which is expressly targeted by setting the default AGC target to 1×10^4. An illustration of these two effects is provided in Fig. 7.4, where the histogrammed signal intensity is plotted for the b_9^{2+} and y_3^+ fragment ions of triply protonated peptide **3**. These ions correlate to give the highest pC-2DMS correlation scoring feature for the scan taken at AGC MS^n target 100.

At AGC MS^n target value 1, there is no intensity measured in the signal channel for the overwhelming majority of scans. Although, impressively, this still enables the b_9^{2+} & y_3^+ correlation signal to be the 4^{th} highest-scoring feature in the AGC target value 1 scan, for many other signals this low signal count combined with the finite detection efficiency of the mass spectrometer causes the correlated product $\langle XY \rangle$ to be too low to produce an identifiable pC-2DMS correlation feature. For example,

the scan at AGC MS^n target value 100 produces correlation features at b_6^+ & $b_{i(7-8)}^+$ and b_7^+ & $b_{i(8-9)}^+$ in the top 50 pC-2DMS correlation-scored features, but neither of these features appears in the same list at AGC target value 1. In the case of both $b_{i(7-8)}^+$ and $b_{i(8-9)}^+$, the signal intensity in the respective channels is only measured to be non-zero at 12 different scans of the full set of 10 000.

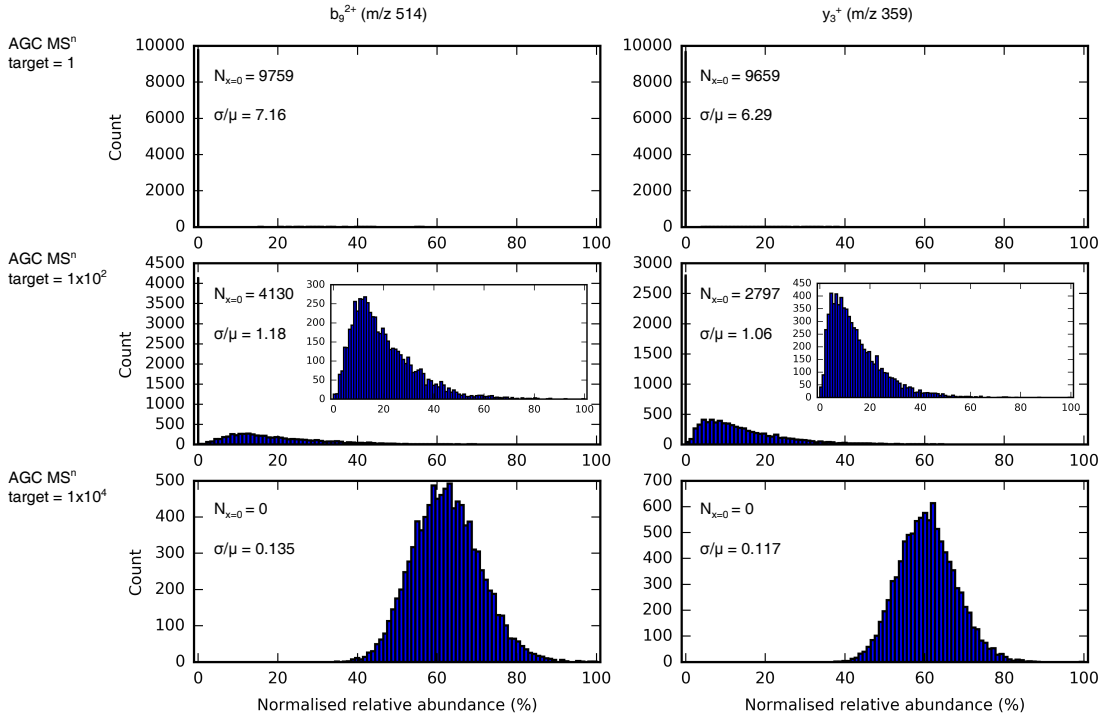


Figure 7.4: Distribution of the normalised signal intensity in b_9^{2+} and y_3^+ signal channels for the pC-2DMS measurements of triply charged peptide **3**. Width of the signal channel is 3.3 Da. $N_{x=0}$ is the number of scans (out of 10 000) for which there is zero measured intensity in the signal channel in question. The width of the first bin is restricted so that it only features counts at zero intensity, and the count in this first bin dominates the histograms at AGC MS^n targets 1 and 1×10^2. The insets for the scans at AGC MS^n target 1×10^2 show the same histograms (same axis labels) without the dominant zero bin, to show the distribution of non-zero signal intensities more clearly. For each signal channel at each AGC target value, the signal intensity has been normalised to the percentage of its maximum single-scan value across all 10 000 scans.

The quantity σ/μ is the ratio of the sample standard deviation to the sample mean. Known as the coefficient of variation, this quantity acts as a standardised measure of the dispersion of values in a dataset for comparison of datasets with different mean values. A greater coefficient of variation means greater variability of the values in a

dataset. At AGC MS^n target 1×10^4, which is the default setting for the instrument for 1D MS/MS scans, the coefficient of variation is an order of magnitude lower than for the scans at target value 1×10^2. This default value has been chosen to minimise scan-to-scan fluctuations before the onset of space-charge effects due to a high density of ions in the trap [55], and the resultant suppression of scan-to-scan fluctuations also suppresses the covariance and partial covariance between signal channels. Although the b_9^{2+} & y_3^+ correlation appears (albeit at a significantly lower ranking of 12^{th}) in the list of top-scoring pC-2DMS features from the AGC MS^n target 1×10^4 scan, it is one of only two correlation features in the top 50 features that survives from the AGC MS^n target 1×10^2 scan (the other being b_{10}^{2+} & y_2^+), and one of only eight automatically identified features overall from the top 50 pC-2DMS correlation-scored features measured at this higher event rate.

Although the default AGC MS^n target value of 1×10^4 for 1D MS/MS scans in the LTQ XL also exists to maximise the number of ions detected at each scan, to improve the ion statistics and therefore mass accuracy and resolution, it is illustrated throughout this thesis that using a value of 1×10^2 for pC-2DMS measurements does not impair the m/z-based determination of fragment ions at the m/z tolerance of ± 0.8 Da generally used for linear ion trap measurements.

7.3 Number of scans

For all the pC-2DMS measurements presented in this work, 10 000 repeat MS/MS scans were acquired. This large number of scans ensures a sufficiently high level of signal-to-noise on the pC-2DMS map [42] for clean visual presentation, but is a gross overestimation of the number of scans that is currently needed to extract structurally informative correlation signals.

Figure 7.5 shows the dependence of pC-2DMS signals on the number of scans used to calculate the pC-2DMS map, for a representative sample of doubly and triply charged peptide ions. For each number of scans, the map is calculated and the pC-2DMS features extracted are ranked according to their pC-2DMS correlation score

calculated using Eq. 6.2. The manually verified top N ($N = 40$ for a doubly charged parent ion and $N = 50$ for a triply charged parent ion) correlation signals obtained at 10 000 scans are used as a reference. The red curves (upper panels) give the number of the top N correlations calculated at each number of scans which are also found in the verified reference correlations (this value is called X), while the blue curve (lower panels) gives the percentage of these manually verified correlations that appear in the X top-scored signals. As such, the value plotted by the blue curve provides a measure of the fidelity of the top X signals. It is clear from these plots that it is possible to extract a very high quality signal list after only $\sim$1000 scans with the use of pC-2DMS correlation scoring to identify true signals.

The application of either bootstrap resampling to calculate the pC-2DMS correlation score (Eq. 6.2), or simply ranking the partial covariance features according to their volume, was investigated to determine whether it increased the number of extracted true pC-2DMS signals at low numbers of scans. At the full number of scans, it was found that calculation of the pC-2DMS score enables the identification of low-volume true correlation signals from high-volume statistical noise (see e.g. Fig. 6.4), and for calculation of this score it was found that the use of jackknife resampling over bootstrap resampling provides greater computational efficiency and perfect repeatability without compromise in performance (section 6.2). However, it was possible that at lower numbers of scans one of either pC-2DMS feature volume or bootstrap resampling could provide improved performance over the pC-2DMS correlation score calculated using jackknife resampling. The number of true signals extracted using each of the three methods (red - pC-2DMS correlation score from jackknife resampling, black - pC-2DMS correlation score from bootstrap resampling, green - volume of pC-2DMS features), up to 3000 scans, is plotted in the insets of the top panels in Fig. 7.5. It can be seen that use of the pC-2DMS correlation score provides a clear advantage in the number of extracted signals at lower numbers of scans over simply ranking by volume. For the calculation of the pC-2DMS correlation score, the less computationally heavy jackknife resampling method does not perform any worse than the more rigorous bootstrap.

7.4 Abundance of correlated fragment ions

The requirements on fragment ion abundance for the production of a true, high scoring pC-2DMS correlation feature were investigated, to determine whether a high-scoring pC-2DMS feature requires the ions which it correlates to appear at high abundance. To do so, pC-2DMS measurements were performed on a representative sample of doubly and triply protonated peptides of different modification states[4]. The pC-2DMS correlation score was calculated for all identified features and used as a ranking parameter for each measurement to identify the true correlation features (taking the top 40 features for 2+ parent ions and the top 50 for 3+ parent ions). Moreover, the plotted features were manually verified. Plotted in Fig. 7.6 is a 2D histogram showing the number of true, identified pC-2DMS correlation signals (z-axis) as a function of the measured relative abundances (RA, in %) in the 1D MS/MS spectrum of the two correlating fragments, $(m/z)_1$ and $(m/z)_2$. $(m/z)_1$ chosen to be the smaller of the two m/z values to avoid double counting. Note the logarithmic bin size of the histogram, which demonstrates that pC-2DMS is able to simultaneously identify true correlation signals of molecular fragments spanning more than three orders of magnitude of measured abundance in the 1D MS/MS spectrum. Ions at very low abundance do not necessarily require the ion with which they correlate to be of high abundance to produce a top-scoring pC-2DMS feature, e.g. correlations between ions of 0.1% RA and 1% RA are very common.

[4]The nine peptide ions used to construct the figure are: $[GGNFSGR(Me)GGFGGSR+3H]^{3+}$ (R(Me) = N^G-monomethylarginine), $[EQFDDpYGHMRF\text{-}NH_2+3H]^{3+}$ (pY = phosphotyrosine), $[LGEY(nitro)GFQNAILVR+2H]^{2+}$ (Y(nitro) = 3-nitrotyrosine), $[GWGR(Me_2)EENLFSWK+3H]^{3+}$ ($R(Me_2)$ = N^G,N^G-dimethylarginine), $[TWpTLCGTVEY+2H]^{2+}$ (pT = phosphothreonine), $[LGEY(nitro)GFQNAILVR+3H]^{3+}$ (Y(nitro) = 3-nitrotyrosine), $[VTIMPKDIQLAR+3H]^{3+}$, $[SApTPEALAFVR+2H]^{2+}$ (pT = phosphothreonine), $[EQFDDpYGHMRF\text{-}NH_2+2H]^{2+}$ (pY = phosphotyrosine).

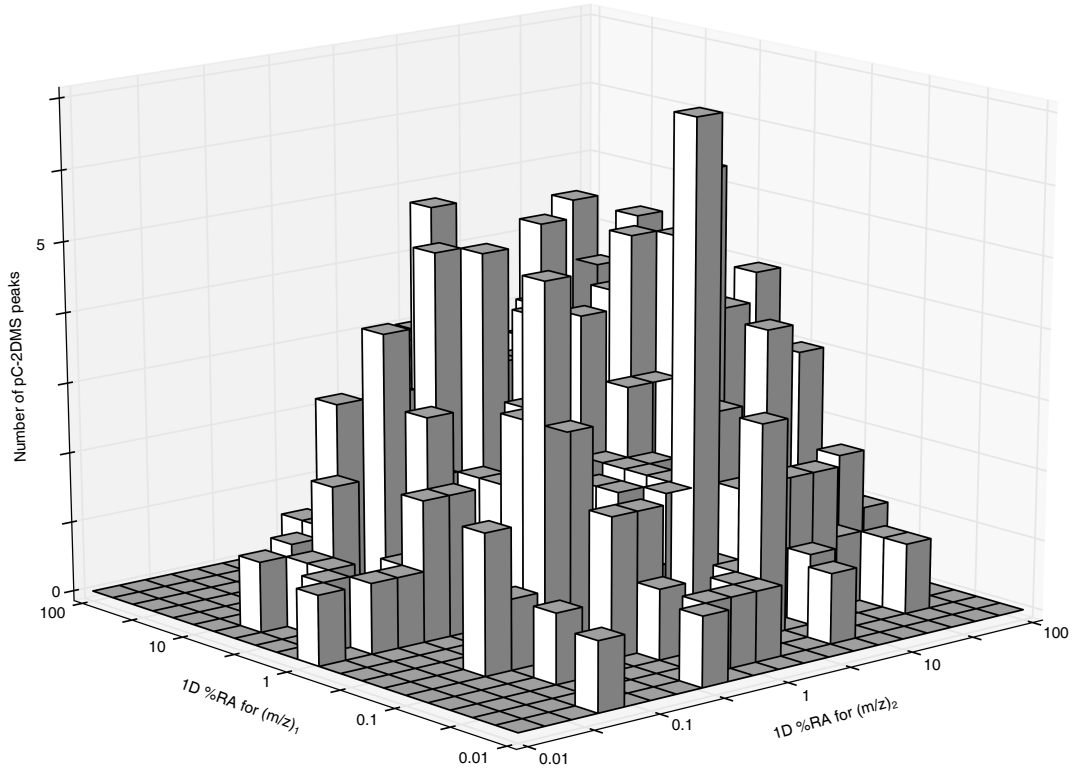


Figure 7.6: Two-dimensional histogram demonstrating that pC-2DMS is able to identify, in the same 2D spectrum, true correlation features with a high pC-2DMS correlation score from fragment ions spanning more than three orders of magnitude of measured 1D MS/MS relative abundance, including where one or both ions is measured at extremely low abundance.

7.5 Generality of technique

The general applicability of pC-2DMS was confirmed through measurements on a variety of modified and unmodified peptide sequences, as well as DNA and RNA oligonucleotide sequences (see Chapter 17), at different charge states and polarity. Table 7.1 shows gives a list of oligomer ions on which pC-2DMS measurements were performed, and the resultant pC-2DMS correlation features were manually determined to correspond to true intrinsic correlations due to production of correlated fragment ions in the same fragmentation reaction of the same molecule. This table demonstrates the generality of the pC-2DMS technique across parent ion charge

Sequence	Modification	Charge states analysed
SApTPEALAFVR	pT=phosphothreonine	2+
TWpTLCGTVEY	pT=phosphothreonine	2+
RDpYTGW-Nle-DF-NH_2	pY=phosphotyrosine	2+, 2−
DsYMGWMDF-NH_2	sY=sulphotyrosine	2+
EQFDDsYGHMRF-NH_2	sY=sulphotyrosine	2+
RDsYTGW-Nle-DF-NH_2	sY=sulphotyrosine	2+
GDFEEIPEEYLQ	None	2−
VTIMPK(Ac)DIQLAR	K(Ac)= N^{ε}-acetyllysine	2+, 3+
GGNFSGR(Me)GGFGGSR	R(Me)=N^{G}-monomethylarginine	2+, 3+
GWGR(Me_2)EENLFSWK	R(Me_2)=N^{G},$N^{G'}$-dimethylarginine	2+, 3+, 2−
TWR(Me_2)GGEEK	R(Me_2)= N^{G},N^{G}-dimethylarginine	2+, 3+
VTIMPK(Me_3)DIQLAR	K(Me_3)= N^{ε},N^{ε},N^{ε}-trimethyllysine	2+, 3+
LGEY(nitro)GFQNAILVR	Y(nitro)=3-nitrotyrosine	2+, 3+
EQFDDpYGHMRF-NH_2	pY=phosphotyrosine	2+, 3+
GSNKGAIIGLM	None	2+, 3+
MLGIIAGKNSG	None	2+, 3+
VTIMPKDIQLAR	None	2+, 3+
YGGFLRRIRPKLK	None	2+, 3+
GK(Ac)GGKGLGK(Ac)GGAKR	K(Ac)= N^{ε}-acetyllysine	2+, 3+
GKGGK(Ac)GLGKGGAK(Ac)R	K(Ac)= N^{ε}-acetyllysine	2+, 3+
GK(Ac)GGKGLGKGGAK(Ac)R	K(Ac)= N^{ε}-acetyllysine	2+, 3+
GKGGK(Ac)GLGK(Ac)GGAKR	K(Ac)= N^{ε}-acetyllysine	2+, 3+
GDFEEIPEEpYLQ	pY=phosphotyrosine	2−, 3−
d(5'–GATCGT–3')	None	2+
r(5'–GAUCGU–3')	None	2−, 3−, 4−

Table 7.1: Table listing all oligomer ions for which the initial validation of pC-2DMS was performed.

states, modification states and polarities, as well as different classes of biopolymer. Nonetheless it is not fully exhaustive, and for example does not include the intact protein molecules on which the top-down experiments in Chapter 14 were performed.

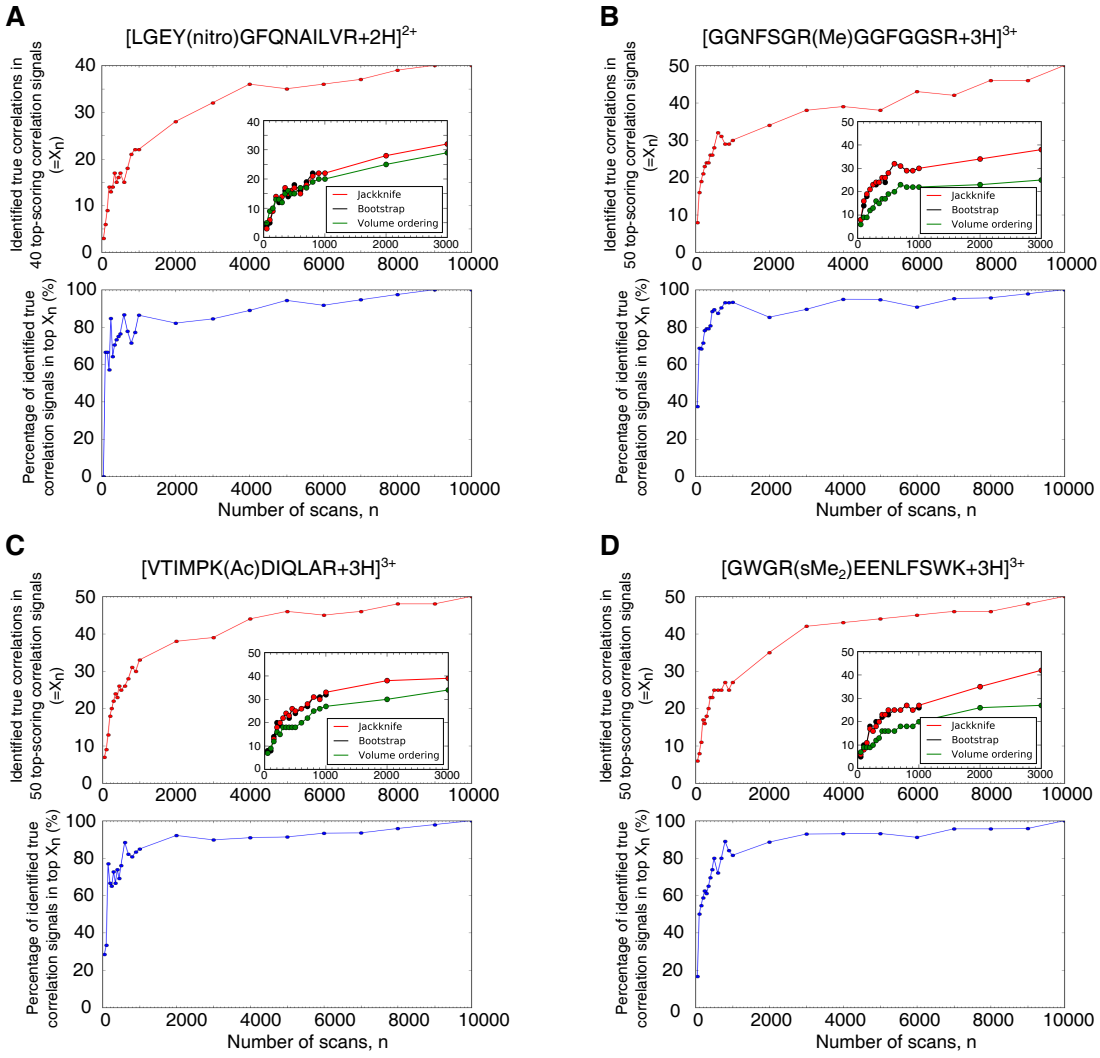


Figure 7.5: Dependence of extracted signals on number of scans used to calculate the pC-2DMS map, showing that it is possible to extract a very high fidelity signal list after only ∼1000 scans using pC-2DMS correlation scoring. The results from the more computationally expensive bootstrap resampling method (plotted in black) are mostly hidden behind identical results from the more efficient jackknife resampling method (plotted in red).

Chapter 8

Formal theory of total ion count partial covariance for noisy Poisson processes

The derivation of the formulae for partial covariance mapping with one and multiple fluctuating parameters (Eqs. 3.1 and 3.2), and thus the implementation of all partial covariance mapping experiments up until now, relies on the individual scan-to-scan monitoring of all fluctuating parameters which induce common mode correlations. Moreover, it requires that the measured abundances of the correlated products grow and shrink linearly with the fluctuating parameter(s). Without the ability to monitor all fluctuating parameters individually, in pC-2DMS the effect of all fluctuating parameters is instead approximated by the total ion count (TIC) at each shot. This clearly deviates from the required specifications for the application of the partial covariance formula, so the validity of this new method requires formal mathematical investigation.

Working with Bridgette Cooper (University College London) and Serguei Patchkovskii (Max Born Institute for Nonlinear Optics and Short Pulse Spectroscopy, Berlin) the formal theory of covariance for noisy Poisson processes described in Chapter 4 and reference [53] has been extended to pC-2DMS and total ion count (TIC) partial covariance mapping. The applicability of this formal theory to pC-2DMS as implemented

in this work has been verified by the augmented Poisson distribution measured for the event rate (i.e. number of ions trapped at each shot) in our typical ESI experiments, see Fig. 4.1 on page 30.

Recall that, as described in Chapter 4, an analytical term for the covariance measured in an experiment with known central event rate ν_0, shot-to-shot standard deviation of event rate σ, and relevant branching ratios for different fragmentation events, is derived by considering the multivariate moments of all elementary outcomes (e.g. detection of a particular ion or pair of ions) of interest. Following the derivation of a general recursive expression for the multivariate moments $M(\mathbb{K})$, where $\mathbb{K}$ is a vector in which each element selects the desired order of the moment of a different elementary event, careful consideration of all the elementary events which can produce signals in the two correlated channels reduces the formula for the covariance to an expression consisting exclusively of different multivariate moments, as demonstrated for a simple case in Eq. 4.15 on page 37. The same analysis as for this simple example can be performed to find an expression for the TIC partial covariance calculated from Eq. 3.1.

Model The analytical model for pC-2DMS is a development of the simple example presented at the end of Chapter 4. We will begin by continuing with this simple case of a doubly charged parent ion which dissociates to two fragment ions, both always detected at unit detection efficiency. Instead of fragmentation induced by photoionisation, in our case this fragmentation is induced by the slow heating of the parent molecule under low energy ion trap CID. We have the expression for the covariance measured between the fragment ions of interest A^+ and B^+ in terms of both the signal measured in the relevant signal channels, A and B, and the multivariate moments of the events of interest (replicated from page 37):

$$\begin{aligned}\langle AB\rangle - \langle A\rangle\langle B\rangle \\ =&M(2,0,0,0) + M(1,1,0,0) + M(1,0,1,0) + M(0,1,1,0) - \\ &[(M(1,0,0,0) + M(0,1,0,0)]\,[M(1,0,0,0) + M(0,0,1,0))]\,, \quad (4.15)\end{aligned}$$

which following lengthy manipulation gives the useful expression for the covariance between A and B:

$$Cov(A,B) = P_1\nu_0 + \sigma^2\nu_0^2(P_1+P_2)(P_1+P_3). \tag{4.16}$$

In order to extend this expression to the TIC partial covariance, we will need to consider subtraction of the second 'correction' term from Eq. 6.1, again replicated here from page 73:

$$pCov(\mathbf{Y},\mathbf{X};TIC) = Cov(\mathbf{Y},\mathbf{X}) - \frac{Cov(\mathbf{Y},TIC)Cov(TIC,\mathbf{X})}{var(TIC)}. \tag{6.1}$$

This requires us to be able to express terms involving the TIC in using the multivariate moments $M(\mathbb{K})$. Happily, this is possible to do. In each of the four elementary events considered on page 36 (which cover all possible outcomes where a signal is measured), a known number of ions are detected. The moments of the TIC can therefore be written as an appropriate sum of the moments of all these four elementary events, for example in the simple case where every elementary event $n_{1,2,3,4}$ produces two detected ions:

$$\begin{aligned}\langle TIC\rangle &= 2\langle(n_1+n_2+n_3+n_4)\rangle \\ &= 2M(1,0,0,0) + 2M(0,1,0,0) + 2M(0,0,1,0) + 2M(0,0,0,1),\end{aligned} \tag{8.1}$$

and:

$$\begin{aligned}\langle TIC^2\rangle =& \langle[2\,(n_1+n_2+n_3+n_4)]^2\rangle \\ =& \langle\left(4n_1^2+4n_2^2+4n_3^2+4n_4^2+2\,(4n_1n_2+4n_1n_3+4n_1n_4+4n_2n_3+4n_2n_4+4n_3n_4)\right)\rangle \\ =& 4M(2,0,0,0)+4M(0,2,0,0)+4M(0,0,2,0)+4M(0,0,0,2)+ \\ & 2[4M(1,1,0,0)+4M(1,0,1,0)+4M(1,0,0,1)+ \\ & \quad 4M(0,1,1,0)+4M(0,1,0,1)+4M(0,0,1,1)].\end{aligned} \tag{8.2}$$

Following the lengthy calculation of all relevant moments and application to Eq.

6.1, the final result for the TIC partial covariance measured between ions A^+ and B^+ for the simple case of the fragmentation of a doubly charged parent ion into two fragment ions measured with 100% detection efficiency (elementary outcomes of interest listed on page 36) is:

$$pCov(A, B; TIC) = \nu_0 P_1 - \frac{\nu_0(P_1 + P_2)(P_1 + P_3)}{P_1 + P_2 + P_3 + P_4}. \tag{8.3}$$

This shows us that the TIC partial covariance has the same 'true covariance' term $\nu_0 P_1$, linear in central event rate and branching ratio of the process in question, as the analytical expression derived for the simple covariance (Eq. 4.16). However, the 'false' covariance term introduced by non-zero σ in Eq. 4.16 is suppressed by the subtraction of the additional 'correction' term, and replaced by the expression $-\frac{\nu_0(P_1+P_2)(P_1+P_3)}{P_1+P_2+P_3+P_4}$ which is always negative because all probabilities and the central event rate are always positive. Note that because both terms are linear in the event rate, changing the event rate will not change the relative volume of different such features on the pC-2DMS map.

This model can be improved to account for more realistic experimental conditions. In a real CID experiment, not every parent ion produces two fragment ions which are detected. This can be a result of finite detection efficiency, or the production of only one final detectable product (such as the intact parent ion, or the parent ion minus a neutral loss). Incorporating these effects into our model increases the number of different elementary events to consider. We are still considering the fragmentation of a doubly charged parent ion to different fragment products (now including the possibility for no fragmentation, i.e. parent ion stays intact) and we are interested in the TIC partial covariance measured between A^+ and B^+. Let us attribute a finite single particle detection efficiency, β, to our detector and denote fragment ion branching ratios by α_i: α_1 is the branching ratio of the reaction producing A^+ and B^+ together, α_2 is the branching ratio of the reaction producing A^+ and any other ion Y^+ which is not B^+, α_3 for a reaction producing B^+ and any other ion Y^+ which is not A^+, α_4 for a reaction producing Y^+ and Z^+ where neither ion is A^+ or B^+ and

α_5 for a reaction producing a single detectable ion Y^+ which is neither A^+ nor B^+, such as the parent ion staying intact or suffering a neutral loss only. This results in the following elementary events of interest:

1. Both A^+ and B^+, produced in the same fragmentation reaction, are detected. The probability of this happening is:

$$P_1 = \alpha_1 \beta^2$$

2. Only A^+ reaches the detector - either B^+ or Y^+ is also produced but does not reach the detector because of the finite detection efficiency. These two physically different events can be grouped together because they produce the same observable (detection of A^+).

$$P_2 = (\alpha_1 + \alpha_2)\beta(1 - \beta)$$

3. Analogous to the above, only B^+ reaches the detector.

$$P_3 = (\alpha_1 + \alpha_3)\beta(1 - \beta)$$

4. Both A^+ and Y^+ are detected.

$$P_4 = \alpha_2 \beta^2$$

5. Both B^+ and Y^+ are detected.

$$P_5 = \alpha_3 \beta^2$$

6. Y^+ is detected alone. This could be because Y^+ is produced with either A^+, or B^+, or Z^+, but the second ion does not reach the detector, or this could be the result of a process which only produces one detectable ion such as a neutral

loss from the parent ion, or no fragmentation at all (intact parent ion detected). Again, all these physically different events produce the same observable so can be grouped together.

$$P_6 = \beta(\alpha_5 + (1-\beta)(\alpha_2 + \alpha_3 + 2\alpha_4))$$

7. Any two ions which are not A^+ or B^+ reach the detector.

$$P_7 = \alpha_4\beta^2$$

Collecting all the terms which contribute to the correlated product $\langle AB\rangle$ and uncorrelated product $\langle A\rangle\langle B\rangle$ in an identical method to that demonstrated for Eq. 4.15 on page 37, and calculating the required moments according to Eq. 4.9, returns the the covariance between A and B, $Cov(A,B)$:

$$Cov(A,B) = \nu_0 P_1 + \nu_0^2\sigma^2(P_1+P_2+P_4)(P_1+P_3+P_5),$$

which can be rewritten substituting the probabilities for the outcomes of individual elementary events P_i with the branching ratios α_i and the detection efficiency β:

$$Cov(A,B) = \beta^2\nu_0\alpha_1 + \beta^2\nu_0^2\sigma^2(\alpha_1+\alpha_2)(\alpha_1+\alpha_3). \tag{8.4}$$

This is in perfect alignment with the expression 4.16 found for the simple case considered as an example in Chapter 4, where the assumed unit detection efficiency rendered the detection probabilities identical to the branching ratios and so both quantities were depicted by P_i. It can be seen that when the detection efficiency β is reduced below 100%, the value for the covariance is reduced through multiplication by a factor of β^2. As seen on the simple covariance maps for our experiments (e.g. panel **a** of Fig. 6.2), there is an undesirable noise term $\beta^2\nu_0^2\sigma^2(\alpha_1+\alpha_2)(\alpha_1+\alpha_3)$ which produces 'false' islands of positive covariance that are not a result of two fragment ions originating from the same decomposition process of the same parent molecule. In Chapter 6 it was

discovered experimentally that it is possible to suppress these unwanted correlations by subtraction of the correction term $Cov(A, TIC)Cov(TIC, B)/var(TIC)$ from the simple covariance. The formal effect of removing this term can be investigated by again rewriting it in terms of the multivariate moments $M(\mathbb{K})$ as demonstrated in Eqs. 4.15, 8.1 & 8.2[1], calculating these moments using Eq. 4.14 and subtracting the result from Eq. 8.4. This lengthy process results in the final analytical expression for the TIC partial covariance between A and B:

$$\begin{aligned} pCov(A, B; TIC) =& \beta^2 \nu_0 \alpha_1 - \frac{\beta(1+\beta)\nu_0(\alpha_1+\alpha_2)(\alpha_1+\alpha_3)}{2\alpha_s} \\ &+ \frac{\nu_0(\alpha_1+\alpha_2)(\alpha_1+\alpha_3)\alpha_5\beta(1+\beta+\nu_0\sigma^2\beta(2\alpha_s+\alpha_5(1+\beta)))}{2\alpha_s(2\alpha_s+\alpha_5+(2\alpha_s+\nu_0\sigma^2(2\alpha_s+\alpha_5)^2)\beta)}, \end{aligned} \tag{8.5}$$

where $\alpha_s = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4$.

The third, final term of this expression is a small unwanted positive expression which is non-zero even when the branching ratio of the process which produces A^+ and B^+ together is zero ($\alpha_1 = 0$). This term is unwanted because it opens the possibility that positive partial covariance can still be observed between A and B even when A^+ and B^+ are never produced in the same reaction.

It will be instructive to consider the numerical behaviour of this complex multivariate expression in some different example cases. Below we consider three different cases, designed to answer the following three questions: i) if I am probing an experiment where A^+ and B^+ are created in the same reaction, will I observe positive TIC partial covariance between A and B?, ii) does this remain the case when the two signal channels A and B are congested by additional fragment ions A^+ and B^+ which are produced by fragmentation mechanisms other than the one producing A^+ and B^+ together? and iii) if we are probing a reaction where A^+ and B^+ are not produced in the same reaction but I measure signal in channels A and B from other decomposition reactions, will I observe false positive TIC partial covariance due to

[1]In this more complex example, where only one ion is detected for elementary events 2, 3 and 6, the measured TIC becomes $2n_1 + n_2 + n_3 + 2n_4 + 2n_5 + n_6 + 2n_7$.

the third positive term?

For this numerical analysis, we use the experimentally measured values for the central event rate ν_0 and the noise term σ for our typical ESI experiments in panel **A** of Fig. 4.1. These are the measured values for an AGC MS^n target value of 100, which represents the standard experimental value used for a pC-2DMS measurement on a single peptide ion. The parameters in panel **B** measured for AGC MS^n target 300 ($\nu_0 = 300, \sigma = 0.18$) were also investigated and found to give the same behaviour as demonstrated below except with higher absolute values for the TIC partial covariance. The sensitivity of TIC partial covariance measurements on instrumental detection efficiency β is investigated by evaluating the expression at three different values of β; $\beta = 0.8$, $\beta = 0.5$ and $\beta = 0.3$.

For the first scenario, scenario I, we will consider the case where all measured signal in the signal channels A and B comes from the fragmentation reaction which produces A^+ and B^+ together and which has branching ratio α_1. Because there are no other fragmentation pathways which produce either A^+ or B^+, the branching ratios α_2 and α_3 are both zero. The two-dimensional colour plots scan α_1 and α_5 (which is the branching ratio of the fragmentation reaction producing only one detectable product which is neither A^+ nor B^+, e.g. the intact parent ion), with the 'remaining' branching ratio attributed to α_4 (branching ratio of any fragmentation reaction which produces neither A^+ nor B^+), $\alpha_4 = 1 - (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_5)$. Each plot is truncated by removal of the upper right triangle, which avoids the unphysical situation where the $\alpha_4 < 0$, i.e. the sum of the branching ratios $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_5$ exceeds 1.

The values of the three different terms of Eq. 8.5 and the final result $pCov(A, B; TIC)$ for scenario I are plotted on the colour maps in Fig. 8.1. It can be seen that in this scenario, the value of $pCov(A, B; TIC)$ grows with increasing α_1 at small α_1. At larger values of α_1, the second negative term dominates the TIC partial covariance and causes it to become negative beyond a certain threshold value of α_1. This threshold value shrinks as the branching ratio α_5 increases. This negative value can be understood as the use of the TIC as a single partial covariance parameter causing an overestimation of the second, 'correction' term in Eq. 3.1. The value at which the

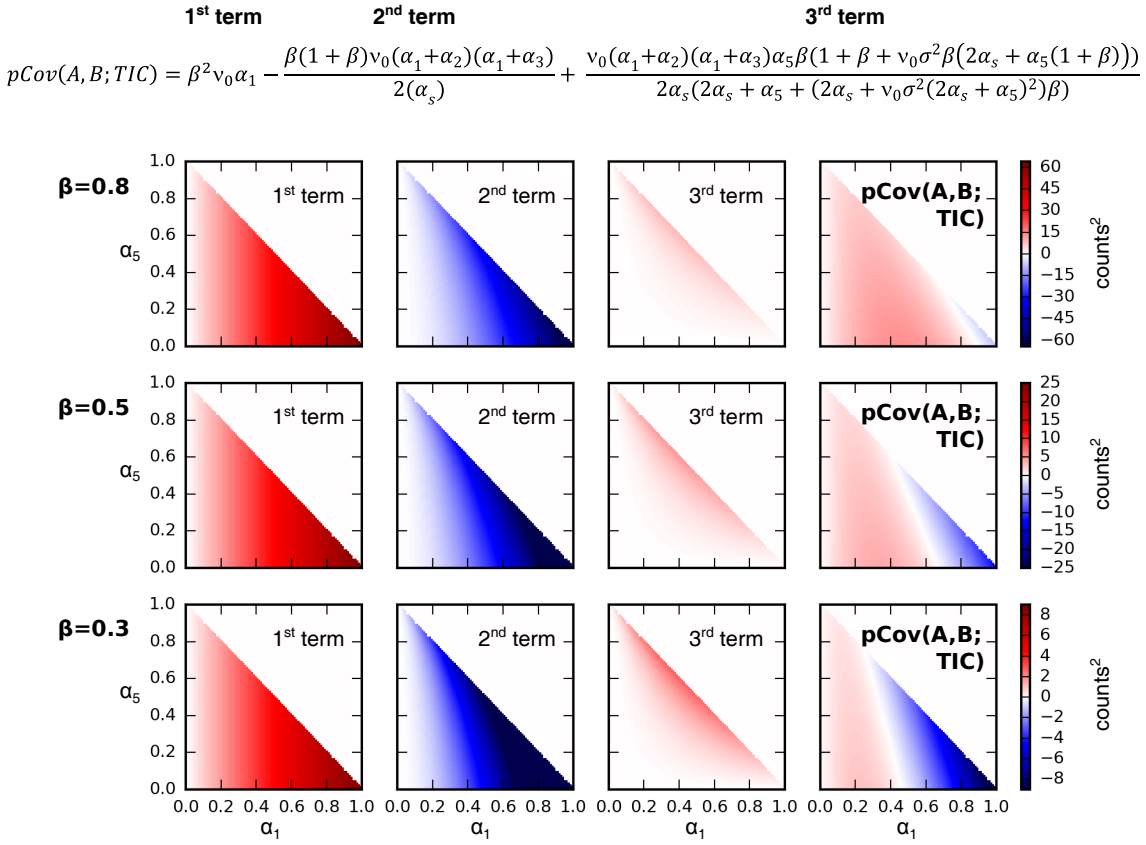


Figure 8.1: Behaviour of Eq. 8.5 in scenario I, for the TIC partial covariance between A and B when A^+ and B^+ are produced in the same reaction and the two signal channels are uncongested. α_1 and α_5 are scanned, $\alpha_2 = \alpha_3 = 0$, $\alpha_4 = 1 - (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_5)$. $\alpha_s = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4$.

TIC partial covariance turns negative is also dependent on the detection efficiency β; at larger values of β the partial covariance remains positive until greater values of α_1. However, even at $\beta = 0.3$ and $\alpha_5 > 0.7$ (such a large value of α_5 can be observed for e.g. loss of a highly labile phosphate group from a phosphorylated serine (S), threonine (T) or tyrosine (Y) residue [122]), the partial covariance remains positive until $\alpha_1 \sim 0.25$. For a fragmenting biopolymer, which typically has a large number of fragmentation pathways available to it, the value of α_1 can be expected to always remain much lower than 0.25. The dominant positive contribution from the partial covariance comes from the first term $\beta^2\nu_0\alpha_1$, especially at the low values of α_1 which are expected for the fragmentation of biopolymeric molecules.

We now investigate what happens to the measured TIC partial covariance if the

signal measured in A and B has a contribution from other fragmentation pathways which do not produce A^+ and B^+ in the same reaction. In other words, the signal channels A and B now become congested so that the averaged 1D spectrum will have contributions in these channels due to fragmentation pathways that are different from the pathway we are trying to probe with our measurement of the partial covariance between A^+ and B^+. This is scenario II, and is described mathematically in the same way as scenario I except that the branching ratios of the reactions which produce only A^+ or B^+, α_2 and α_3 respectively, are now non-zero.

$$pCov(A,B;TIC) = \underbrace{\beta^2 \nu_0 \alpha_1}_{1^{st}\text{ term}} - \underbrace{\frac{\beta(1+\beta)\nu_0(\alpha_1+\alpha_2)(\alpha_1+\alpha_3)}{2(\alpha_s)}}_{2^{nd}\text{ term}} + \underbrace{\frac{\nu_0(\alpha_1+\alpha_2)(\alpha_1+\alpha_3)\alpha_5\beta(1+\beta+\nu_0\sigma^2\beta(2\alpha_s+\alpha_5(1+\beta)))}{2\alpha_s(2\alpha_s+\alpha_5+(2\alpha_s+\nu_0\sigma^2(2\alpha_s+\alpha_5)^2)\beta)}}_{3^{rd}\text{ term}}$$

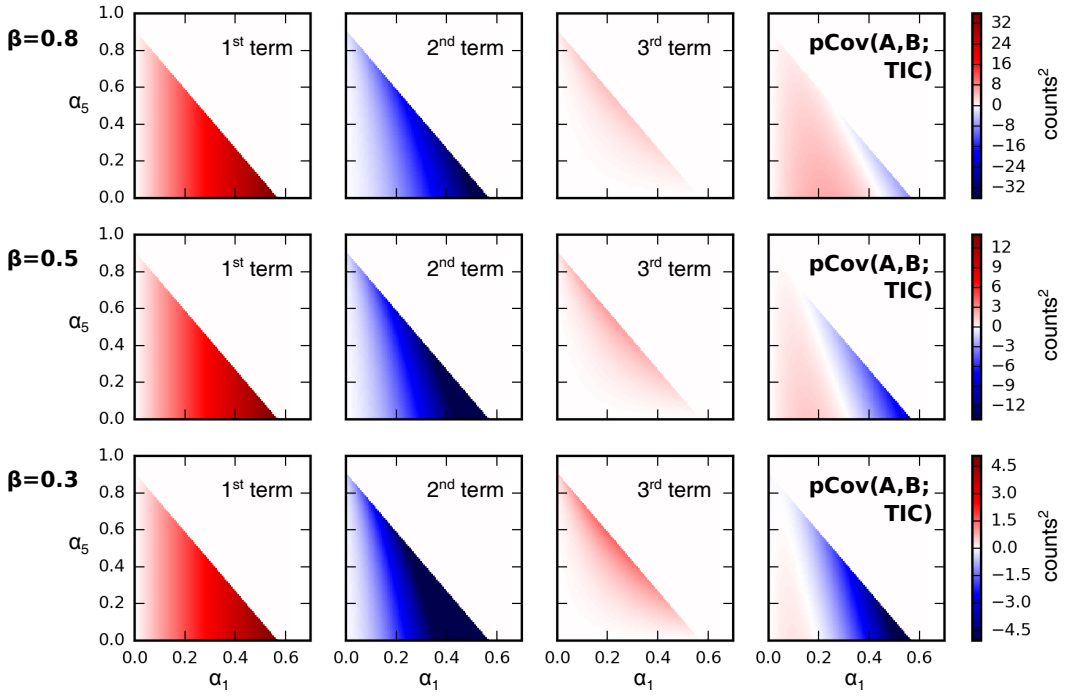


Figure 8.2: Behaviour of Eq. 8.5 in scenario II, for the TIC partial covariance between A and B when A^+ and B^+ are produced in the same reaction and the two signal channels are congested. α_1 and α_5 are scanned, $\alpha_2 = \alpha_3 = 0.3\alpha_1 + 0.05$, $\alpha_4 = 1 - (\alpha_1 + \alpha_2 + \alpha_3 + \alpha_5)$. $\alpha_s = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4$.

For this example case, the value of both α_2 and α_3 becomes $0.3\alpha_1 + 0.05$. The first term, $0.3\alpha_1$, arises from attributing a typical probability of 0.3 to both ions A^+ and B^+ for their undergoing a secondary decomposition following the fragmentation process

which produces both the ions together, for example the loss of a small neutral molecule such as H_2O (particularly common under CID for fragments containing serine (S), threonine (T), glutamic acid (E) and aspartic acid (D)) or NH_3 (particularly common under CID for fragments containing arginine (R), lysine (K), asparagine (N) and glutamine (Q)). The addition of 0.05 accounts for the other possibility, that A^+ and B^+ are produced by another mechanism entirely. This outcome is given the large branching ratio of 0.05 for both fragments A^+ and B^+ to ensure the contribution from this 'contaminating' mechanism is significant. This term also ensures that even when α_1 is zero, there is still some signal measured in both signal channels A and B. In this case it is important that there is no measured partial covariance between A^+ and B^+, because this would represent a false positive.

Fig. 8.2 shows the behaviour of Eq. 8.5 for the case described above. Despite the increase in overall signal measured in the channels A and B, the absolute value of the TIC partial covariance reduces for given values of α_1, α_5. The dominant positive term is still the first term, and there is a shift down in the threshold for α_1 at a given detection efficiency for which the partial covariance becomes negative. Nonetheless, for the small values of α_1 that we would expect for pC-2DMS experiments, the partial covariance remains positive as desired. Moreover, when $\alpha_1 = 0$ so that we would hope to measure no partial covariance between A^+ and B^+, this is indeed the case for the resultant values of α_2 and α_3 in this particular case.

Scenarios I and II represent situations where A^+ are B^+ are produced in the same reaction, so we would like the measured value of $pCov(A, B; TIC)$ to be positive. For the TIC partial covariance described by Eq. 8.5, we also require that it does not produce false positives. In other words, if A^+ and B^+ are not produced in the same reaction ($\alpha_1 = 0$), although both ions may be created separately *via* other fragmentation pathways, we do not want to measure positive partial covariance between A and B. For the two sample cases considered above where α_1 falls to zero at certain values of α_2 and α_3, this is the case. To investigate the false positive rate for Eq. 8.5 more fully, in Fig 8.3 we now set α_1 to zero and determine the value of the TIC partial covariance for all different values of α_5 and α_2, which is given the same value

as α_3. This is scenario III, and in this case the value of the first term $\beta^2\nu_0\alpha_1$ is always trivially zero, so it is not plotted. It can be clearly seen from Eq. 8.5 that it is possible to extract a positive factor of $\nu_0(\alpha_1+\alpha_2)(\alpha_1+\alpha_3)$ from the full expression for the TIC partial covariance when the first term is zero. From here, we can see that setting $\alpha_2 = \alpha_3$ allows the full branching ratio parameter space to be tested for false positive TIC partial covariance by scanning $\alpha_2 = \alpha_3$ from 0 to their maximum physically permissible value given α_5. This is because following factorisation of this positive term, the remaining expression for the TIC partial covariance only has a dependence on α_2 and α_3 through the factor $\alpha_s = \alpha_1+\alpha_2+\alpha_3+\alpha_4$, which is a sum and so is insensitive to the relative contribution from α_2 and α_3.

$$pCov(A,B;TIC) = \underbrace{\beta^2\nu_0\alpha_1}_{1^{st}\text{ term}} - \underbrace{\frac{\beta(1+\beta)\nu_0(\alpha_1+\alpha_2)(\alpha_1+\alpha_3)}{2(\alpha_s)}}_{2^{nd}\text{ term}} + \underbrace{\frac{\nu_0(\alpha_1+\alpha_2)(\alpha_1+\alpha_3)\alpha_5\beta(1+\beta+\nu_0\sigma^2\beta(2\alpha_s+\alpha_5(1+\beta)))}{2\alpha_s(2\alpha_s+\alpha_5+(2\alpha_s+\nu_0\sigma^2(2\alpha_s+\alpha_5)^2)\beta)}}_{3^{rd}\text{ term}}$$

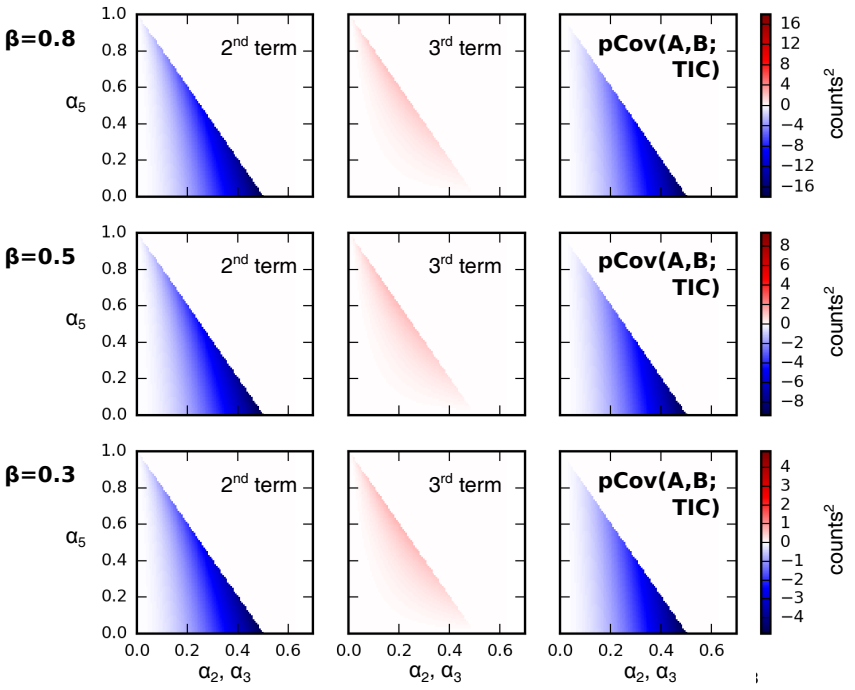


Figure 8.3: Behaviour of Eq. 8.5 in scenario III, for the TIC partial covariance between A and B when A^+ and B^+ are not produced in the same fragmentation pathway but the two signal channels are congested due to other reactions. α_5 and $\alpha_2(=\alpha_3)$ are scanned, $\alpha_1 = 0$ and $\alpha_4 = 1-(\alpha_1+\alpha_2+\alpha_3+\alpha_5)$. $\alpha_s = \alpha_1+\alpha_2+\alpha_3+\alpha_4$.

Fig. 8.3 shows that when $\alpha_1 = 0$, for all values of α_5 and α_2, α_3 and at each

of the considered values for detection efficiency β, the negative contribution from the second term in Eq. 8.5 outweighs the contribution from the third positive term so that the value of the TIC partial covariance is never positive. The greater the branching ratios of the reactions which produce A^+ and B^+ individually, the greater the absolute value of the negative TIC partial covariance, almost independent of α_5. This is an important result because it tells us that the TIC partial covariance is robust to false positives - positive TIC partial covariance measured between two ions A^+ and B^+ is a valid indicator that they were produced in the same fragmentation reaction.

As touched on above, for the experiments probing the decomposition of large biopolymers examined in this thesis, the size of the fragmenting molecules means that the number of competing fragmentation pathways is very large. The competition between these different pathways means that the value of the branching ratio for the fragmentation reaction producing two particular products A^+ and B^+ (α_1) will be very small compared to the combined branching ratios for all other fragmentation pathways ($= \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5$). In the case of negligible α_5 (i.e. complete parent ion fragmentation, as observed for the majority of experiments in this work), it can be straightforwardly seen from Eq. 8.5 that the third term can be disregarded. Furthermore, provided high enough detection efficiency, the denominator ($2\alpha_s = 2\left(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4\right)$) dominates the second term. This means the TIC partial covariance can be approximated as:

$$pCov(A, B; TIC) \sim \beta^2 \nu_0 \alpha_1, \tag{8.6}$$

which tells us that **for measurements of fragmenting biopolymers, the partial covariance between two fragment ions A^+ and B^+ is directly proportional to both the central event rate and the branching ratio of the fragmentation reaction which produces them together.** As such, the value of the TIC partial covariance provides a linear mapping of the abundance (concentration) of the parent ion species, and so TIC partial covariance mapping can be used for quantitative analysis.

It can be seen from Figs. 8.1 & 8.2 that even when α_5 is significant, provided α_1 remains small the dominant positive contribution to $pCov(A, B; TIC)$ comes from the first term, and this in turn overwhelms the negative contribution from the second term, allowing Eq. 8.6 to remain accurate. The validity of TIC partial covariance for quantitative analysis is experimentally tested and confirmed in Chapter 9.

Chapter 9

Resolution of mixtures of combinatorial isomers

Combinatorial isomers, also known as structural isomers, are multiply modified biomolecules with the same sequence where the same type of modification occurs the same number of times, but on different combinations of sites across the molecule. It is important to be able to characterise different combinatorial isomers because, as described below, there is a large body of evidence indicating that different combinations of the same modification can have disparate, significant biological functions.

Although the importance of characterising combinatorial isomers is not limited to protein sequences (for example, methylation patterns of DNA are challenging to resolve and believed to be of great epigenetic importance [123, 124]), the best studied systems in which different combinations of modifications are known to have biological significance are histone proteins. Histone proteins are found in all eukaryotic animals, and interact intimately with the genetic information (DNA) in cells. Octamers of the four core histone proteins (H2A, H2B, H3 and H4) act as spools around which the DNA winds itself, and so function to compact and package the DNA into strands of a higher-order structure known as chromatin. Long thought of as structural elements which participated minimally in biological function, histone proteins are now known to play a key role in both the protection of the genetic information and the regulation of its expression and replication, making them a key area of interest for

the study of epigenetics (i.e. the study of changes in gene expression which are not mediated by changes in the DNA sequence itself [125]). In the mid 1990s the Allis and Schreiber groups demonstrated that enzymes which modify histones play a direct role in the regulation of gene expression [126, 127], signalling the start of a huge research effort focused on understanding the intricate details of the role played by histone modifications.

At the start of this century, the idea that modifications on the 'tails' (last few tens of residues on the N-terminus) of these histone proteins can act sequentially or in concert to code for different biological functions was first articulated [128, 129] on the basis of a number of different experimental observations suggesting different downstream functions for different combinations of histone modifications [130–133]. This hypothesis is generally known as the 'histone code'. The potential complexity of such systems is vast, in accordance with the explosive power of combinatorics - for only four potential modification sites and two different types of modification, the number of possible combinations already grows to $3^4 = 81$. Since this first suggestion that histone modifications may act sequentially or in combination, there has been much focus on the correct characterisation and understanding of these modification patterns in order to be able to map biological function to fundamental chemical structure. These experiments have provided a critical mass of evidence to support the original histone code hypothesis, of which one interesting example demonstrates that certain interaction sites (known as bromodomains) on large proteins that act, among other things, to remodel chromatin and control transcription[1] specifically interact with particular motifs of histone tail modifications [134]. There have been numerous other discovered examples of chromatin-modifying proteins 'talking' with particular patterns of histone tail modifications [135–140].

As the general method of choice for the analysis of protein sequence and PTMs, much of the experimental effort to 'crack' the histone code has been focused on the development of mass spectrometric techniques, which has proved highly success-

[1]The first step of gene expression, where the DNA sequence is read into RNA (which is subsequently read into a protein sequence).

ful [29–31, 141–149]. However, perhaps the most significant remaining problem in the mass spectrometric analysis of combinatorial modifications is the resolution and quantification of structural isomers [29–31]. Very similar chemical properties make these isomers extremely challenging to separate upstream from the mass spectrometer [29] (e.g. they elute at very similar or identical times in high-performance liquid chromatography experiments) and separation in the mass spectrometer itself is not possible since this can only be performed on m/z, which is of course identical for different isomers. Finally, because these isomers only differ in the position of their post-translational modifications, each isomer produces an extremely similar fragmentation pattern and so the inevitable co-fragmentation of multiple isomers produces highly complex tandem mass spectra with numerous overlapping peaks. Analysis of these complex spectra is extremely challenging and it is difficult and often impossible, even in principle, to find any fragment ion signals in the MS/MS spectrum which are able to identify one particular isomer without the risk that they could also have been produced by a different isomer. As such, identification and quantification of different structural isomers, with different biological functions, presents a huge challenge to the field.

There have been a number of diverse efforts to disentangle the complex MS/MS spectra of different structural isomers of combinatorially modified histones. Abshiru et al. [31] compared the overlapping MS/MS spectra of different, co-fragmented isomers to the individually obtained spectra of every possible isomer in its pure form, and attempted to deconvolve the mixed spectra by modelling the deconvolution as a network flow problem. Although this was successful for some mixtures, the deconvolution and quantification failed catastrophically in more complex cases (e.g. mixtures of diacetylated isomers of the H4 histone tail, see Supplementary Table 3 in [31]), and the requirement for measurement of the pure form of every individual isomer is restrictive. Phanstiel et al. [30] solved systems of coupled linear equations to deconvolve the mixed spectra of co-eluting histone structural isomers, but again this method becomes unstuck in the complex case of mixtures of diacetylated[2] H4 histone

[2]In the publication these are referred to as triacetylated forms because the authors include the

tails (deconvolution and quantification of which is conceded to be 'mathematically impossible' using the 1D MS/MS spectrum, see Fig. 2 in [30]).

Fragment ion	$K5_{Ac}K16_{Ac}$	$K8_{Ac}K12_{Ac}$	$K5_{Ac}K12_{Ac}$	$K8_{Ac}K16_{Ac}$	Fragment ion	$K5_{Ac}K16_{Ac}$	$K8_{Ac}K12_{Ac}$	$K5_{Ac}K12_{Ac}$	$K8_{Ac}K16_{Ac}$
$[b/c/a]_2$	x	o	x	o	$[b/c/a]_{i(4-5)}$	x	o	x	o
$[b/c/a]_3$	x	o	x	o	$[b/c/a]_{i(4-6)}$	x	o	x	o
$[b/c/a]_4$	x	o	x	o	$[b/c/a]_{i(4-7)}$	x	o	x	o
$[b/c/a]_5$	x	x	x	x	$[b/c/a]_{i(4-8)}$	x	o	x	o
$[b/c/a]_6$	x	x	x	x	$[b/c/a]_{i(4-9)}$	✓	✓	x	x
$[b/c/a]_7$	x	x	x	x	$[b/c/a]_{i(4-10)}$	✓	✓	x	x
$[b/c/a]_8$	x	x	x	x	$[b/c/a]_{i(4-11)}$	✓	✓	x	x
$[b/c/a]_9$	o	x	x	o	$[b/c/a]_{i(4-12)}$	✓	✓	x	x
$[b/c/a]_{10}$	o	x	x	o	$[b/c/a]_{i(4-13)}$	x	o	x	o
$[b/c/a]_{11}$	o	x	x	o	$[b/c/a]_{i(5-6)}$	x	o	x	o
$[b/c/a]_{12}$	o	x	x	o	$[b/c/a]_{i(5-7)}$	x	o	x	o
$[b/c/a]_{13}$	x	x	x	x	$[b/c/a]_{i(5-8)}$	x	o	x	o
$[y/z/x]_1$	x	x	x	x	$[b/c/a]_{i(5-9)}$	✓	✓	x	x
$[y/z/x]_2$	o	x	x	o	$[b/c/a]_{i(5-10)}$	✓	✓	x	x
$[y/z/x]_3$	o	x	x	o	$[b/c/a]_{i(5-11)}$	✓	✓	x	x
$[y/z/x]_4$	o	x	x	o	$[b/c/a]_{i(5-12)}$	✓	✓	x	x
$[y/z/x]_5$	o	x	x	o	$[b/c/a]_{i(5-13)}$	x	o	x	o
$[y/z/x]_6$	x	x	x	x	$[b/c/a]_{i(6-7)}$	x	x	x	x
$[y/z/x]_7$	x	x	x	x	$[b/c/a]_{i(6-8)}$	x	x	x	x
$[y/z/x]_8$	x	x	x	x	$[b/c/a]_{i(6-9)}$	o	x	x	o
$[y/z/x]_9$	x	x	x	x	$[b/c/a]_{i(6-10)}$	o	x	x	o
$[y/z/x]_{10}$	x	o	x	o	$[b/c/a]_{i(6-11)}$	o	x	x	o
$[y/z/x]_{11}$	x	o	x	o	$[b/c/a]_{i(6-12)}$	o	x	x	o
$[y/z/x]_{12}$	x	o	x	o	$[b/c/a]_{i(6-13)}$	x	x	x	x
$[y/z/x]_{13}$	x	x	x	x	$[b/c/a]_{i(7-8)}$	x	x	x	x
$[b/c/a]_{i(2-3)}$	x	o	x	o	$[b/c/a]_{i(7-9)}$	o	x	x	o
$[b/c/a]_{i(2-4)}$	x	o	x	o	$[b/c/a]_{i(7-10)}$	o	x	x	o
$[b/c/a]_{i(2-5)}$	x	x	x	x	$[b/c/a]_{i(7-11)}$	o	x	x	o
$[b/c/a]_{i(2-6)}$	x	x	x	x	$[b/c/a]_{i(7-12)}$	o	x	x	o
$[b/c/a]_{i(2-7)}$	x	x	x	x	$[b/c/a]_{i(7-13)}$	x	x	x	x
$[b/c/a]_{i(2-8)}$	x	x	x	x	$[b/c/a]_{i(8-9)}$	o	x	x	o
$[b/c/a]_{i(2-9)}$	o	x	x	o	$[b/c/a]_{i(8-10)}$	o	x	x	o
$[b/c/a]_{i(2-10)}$	o	x	x	o	$[b/c/a]_{i(8-11)}$	o	x	x	o
$[b/c/a]_{i(2-11)}$	o	x	x	o	$[b/c/a]_{i(8-12)}$	o	x	x	o
$[b/c/a]_{i(2-12)}$	o	x	x	o	$[b/c/a]_{i(8-13)}$	x	x	x	x
$[b/c/a]_{i(2-13)}$	x	x	x	x	$[b/c/a]_{i(9-10)}$	o	x	x	o
$[b/c/a]_{i(3-4)}$	x	x	x	x	$[b/c/a]_{i(9-11)}$	o	x	x	o
$[b/c/a]_{i(3-5)}$	x	o	x	o	$[b/c/a]_{i(9-12)}$	o	x	x	o
$[b/c/a]_{i(3-6)}$	x	o	x	o	$[b/c/a]_{i(9-13)}$	x	x	x	x
$[b/c/a]_{i(3-7)}$	x	o	x	o	$[b/c/a]_{i(10-11)}$	x	x	x	x
$[b/c/a]_{i(3-8)}$	x	o	x	o	$[b/c/a]_{i(10-12)}$	x	x	x	x
$[b/c/a]_{i(3-9)}$	✓	✓	x	x	$[b/c/a]_{i(10-13)}$	o	x	x	o
$[b/c/a]_{i(3-10)}$	✓	✓	x	x	$[b/c/a]_{i(11-12)}$	x	x	x	x
$[b/c/a]_{i(3-11)}$	✓	✓	x	x	$[b/c/a]_{i(11-13)}$	o	x	x	o
$[b/c/a]_{i(3-12)}$	✓	✓	x	x	$[b/c/a]_{i(12-13)}$	o	x	x	o
$[b/c/a]_{i(3-13)}$	x	o	x	o					

Table 9.1: Table demonstrating that regardless of fragmentation mechanism, sensitivity, mass accuracy or resolution and even provided complete parent ion fragmentation, there are, as a matter of principle, no 1D signal channels able to resolve an arbitrary mixture of the four isomers $K_{5,Ac}K_{16,Ac}$, $K_{8,Ac}K_{12,Ac}$, $K_{5,Ac}K_{12,Ac}$ and $K_{8,Ac}K_{16,Ac}$. Rows denote fragment ions resulting from cleavages of any kind (a/x, b/y or c/z) between the same amino acid residues. Columns denote each of the four diacetylated isomers considered. Structural isomers which share an "x" or "o" across a row produce isomeric fragment ions upon cleavage of the particular bonds(s). Ticks denote the limited set of unique 1D signal channels available to two of the four isomers $K_{5,Ac}K_{16,Ac}$ and $K_{8,Ac}K_{12,Ac}$, whilst it can be seen that there are no possible unique 1D signal channels available to isomers $K_{5,Ac}K_{12,Ac}$ or $K_{8,Ac}K_{16,Ac}$.

The general problem for the resolution of mixtures of combinatorial isomers by 1D MS/MS is the lack of unique signal channels, to uniquely identify a particular

trivially detectable N-terminus acetylation in their modification count.

combination of modifications from other different patterns present in a mixture of co-fragmented isomers. Although this problem (and the solution provided by pC-2DMS) is general, here it is illustrated by consideration of the particularly challenging case of resolving mixtures of the co-eluting [31] diacetylated isomers of the histone H4 tail [150], where the previous attempts described above have been unsuccessful [30, 31]. Experiments are performed on combinatorial isomers of the histone H4 fragment 4–17 $^{4}\mathrm{GK_5GGK_8GLGK_{12}GGAK_{16}R}^{17}$, acetylated at different combinations of two of the four lysine (K) residues K_5, K_8, K_{12} and K_{16}.

Table 9.1 demonstrates that, regardless of fragmentation method, sensitivity, mass accuracy or resolution and allowing for complete fragmentation of the parent ion, 1D mass spectrometry is, as a matter of principle, unable to resolve mixtures of the four diacetylated isomers $K_{5,Ac}K_{16,Ac}$, $K_{8,Ac}K_{12,Ac}$, $K_{5,Ac}K_{12,Ac}$ and $K_{8,Ac}K_{16,Ac}$[3], K_{Ac} = N$^{\epsilon}$-acetyllysine, because of the lack of available unique 1D signal channels. Every possible terminal or internal ion and every possible backbone fragmentation (considering a-, b- and c-type N-terminal and internal ions and x-, y- and z-type C-terminal ions) produced by each of the isomers is considered in an arbitrary mixture of the four. Structural isomers sharing an "x" or "o" across a row produce isomeric fragment ions upon cleavage of the given bond(s), which therefore cannot identify an individual isomer in a mixture. It is seen that there are no possible 1D signal channels which are able to identify the presence of isomers $K_{5,Ac}K_{12,Ac}$ or $K_{8,Ac}K_{16,Ac}$. A tick denotes a unique 1D signal channel for $K_{5,Ac}K_{16,Ac}$ and $K_{8,Ac}K_{12,Ac}$, however these amount to a select few internal ions which are highly challenging to interpret for 1D MS and typically appear at significantly lower intensity than terminal ions in any fragmentation method, precluding their reliable identification above spectral noise in 1D MS. In the 1D CID measurements of the four-fold mixture which is successfully resolved in Fig. 9.4, no 1D ion that could even potentially correspond to a unique 1D signal channel appears above $\sim 2\%$ relative abundance.

Three different mixtures of these diacetylated histones were prepared, for which

[3]The two additional diacetylated isomers $K_{5,Ac}K_{8,Ac}$, $K_{12,Ac}K_{16,Ac}$ are not considered because they produce an abundance of unique identifying fragment ions thanks to the asymmetric distribution of their acetylation patterns, and can therefore be straightforwardly resolved by 1D MS/MS.

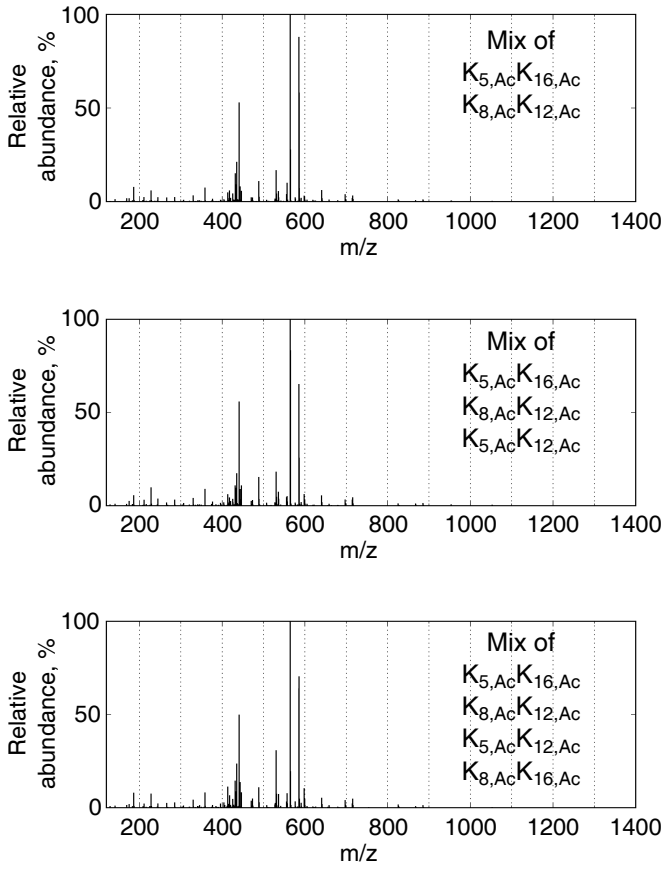


Figure 9.1: Measured 1D MS/MS spectra for each of the three mixtures of the diacetylated H4 histone peptides.

the 1D MS/MS CID spectra are shown in Fig. 9.1. These spectra are only shown to demonstrate the spectral shape of the fragment ion spectra for the mixtures under our low energy CID experimental conditions, and not to prove the already-established fact, demonstrated in Table 9.1, that these mixtures are, as a matter of principle, impossible to resolve using 1D MS on any instrument. As well as there being no possible 1D signal channels to identify any of the three mixtures, the relative abundance of each fragment ion remains highly similar across the mixtures[4].

The two-dimensional character of pC-2DMS allows for the immediate and straighforward deconvolution of mixtures of combinatorial isomers. Figure 9.2 demonstrates

[4]This is probably a major contributing factor to the failure of the authors of reference [31] to resolve and quantify such mixtures.

the general principle of the resolution of structural isomers using pC-2DMS, using the specific example of the $K_{5,Ac}K_{16,Ac}$, $K_{8,Ac}K_{12,Ac}$, $K_{5,Ac}K_{12,Ac}$ and $K_{8,Ac}K_{16,Ac}$ diacetylated isomers. As discussed above, in an arbitrary mixture of any of these co-eluting [31] isomers, all of them will be isolated and fragmented at the same time, producing an overlapping fragment ion spectrum from which it is challenging, and often impossible, to determine the modification state of the species under analysis. Because the same fragment ion can be produced by one or more different structural isomers, there are often no unique 1D signal channels to enable the identification of a particular isomer. pC-2DMS solves this problem by correlating two non-unique 1D fragment signals to produce a 2D marker ion correlation that is unique to an individual isomer. In panel A of Fig. 9.2, measurement of the $y_{5,Ac}$ ion indicates the presence of either the $K_{5,Ac}K_{16,Ac}$ or $K_{8,Ac}K_{16,Ac}$ isomer, but because this fragment could be generated by either isomer it cannot discriminate between these two different modification states. Equally, measurement of the internal $b_{i(3-8),Ac}$ ion[5] indicates the presence of either the $K_{8,Ac}K_{12,Ac}$ or $K_{8,Ac}K_{16,Ac}$ isomer, but is unable to distinguish one from the other because the fragment could have been produced by either.

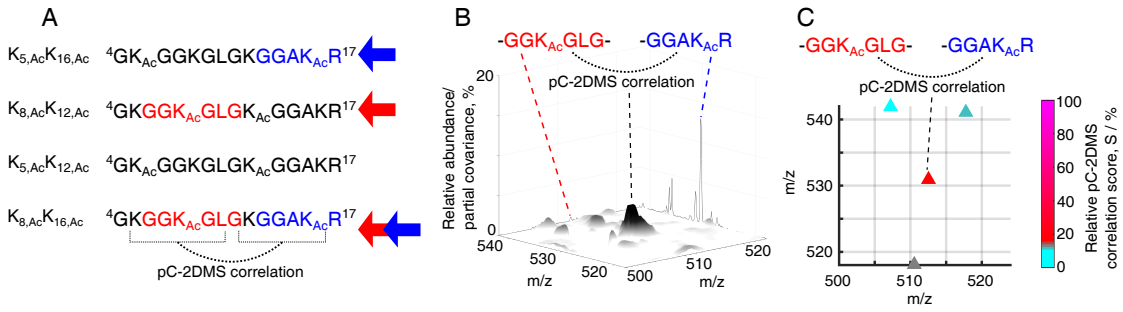


Figure 9.2: Illustration of the principle of the resolution of structural isomers for pC-2DMS. The measured pC-2DMS correlation shown on the pC-2DMS map in panel B and on the scatter plot of resampled features in panel C unambiguously identifies the $K_{8,Ac}K_{16,Ac}$ isomer, whilst panel A shows that measurement of either of the correlated ions in the 1D mass spectrum is unable to do so.

[5]Analysis of internal ions in 1D mass spectrometry is highly challenging due to high levels of potential ambiguity but this is greatly improved by pC-2DMS, see e.g. the unambiguous identification of the $b^{+}_{i(5-8)}$ ion by signals B, D and E in Fig. 6.3 and the dramatic reduction in false positive matching rate for internal ions in Chapter 10.

However, the pC-2DMS correlation observed between the two ions $b^{+}_{i(3-8)}$[6] and $y_{5,Ac}$ indicates the two fragments originate from the same decomposition process of the same parent ion, and therefore could **only have been produced by the $\mathbf{K}_{8,Ac}\mathbf{K}_{16,Ac}$ isomer**. Panel B of Fig. 9.2 shows this experimentally measured unique signal channel on a section of the pC-2DMS map, which is shown in panel C as a scatter plot of features with colour scale indicating their calculated pC-2DMS correlation score (see section 6.2).

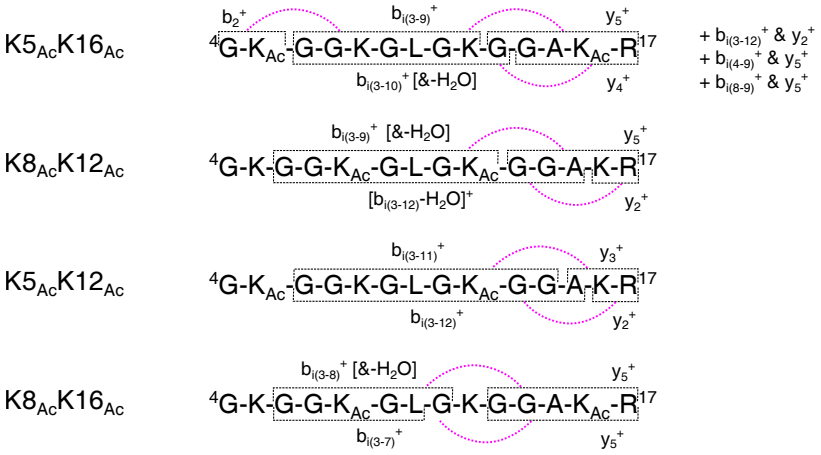


Figure 9.3: Measured unique identifying pC-2DMS signal channels for each of the structural isomers $K_{5,Ac}K_{16,Ac}$, $K_{8,Ac}K_{12,Ac}$, $K_{5,Ac}K_{12,Ac}$ and $K_{8,Ac}K_{16,Ac}$. Magenta dashed line indicates a correlation signal between two different fragment ions. All correlation signals shown could only have been produced by the structural isomer in question.

The pC-2DMS spectra of each of the four diacetylated isomers were measured, and it was verified that each isomer produces an abundance of unique 2D correlation signals. Those unique 2D signals which were found to be in the top 50 pC-2DMS correlation score-ordered features for the measurement of each pure isomer are shown in Fig. 9.3 (there are many more possible unique 2D signal channels, which could be measured at different experimental conditions).

[6]The identity of the internal $b_{i(3-8),Ac}$ fragment has been confirmed by MS^3 measurements, see Fig. B.1.

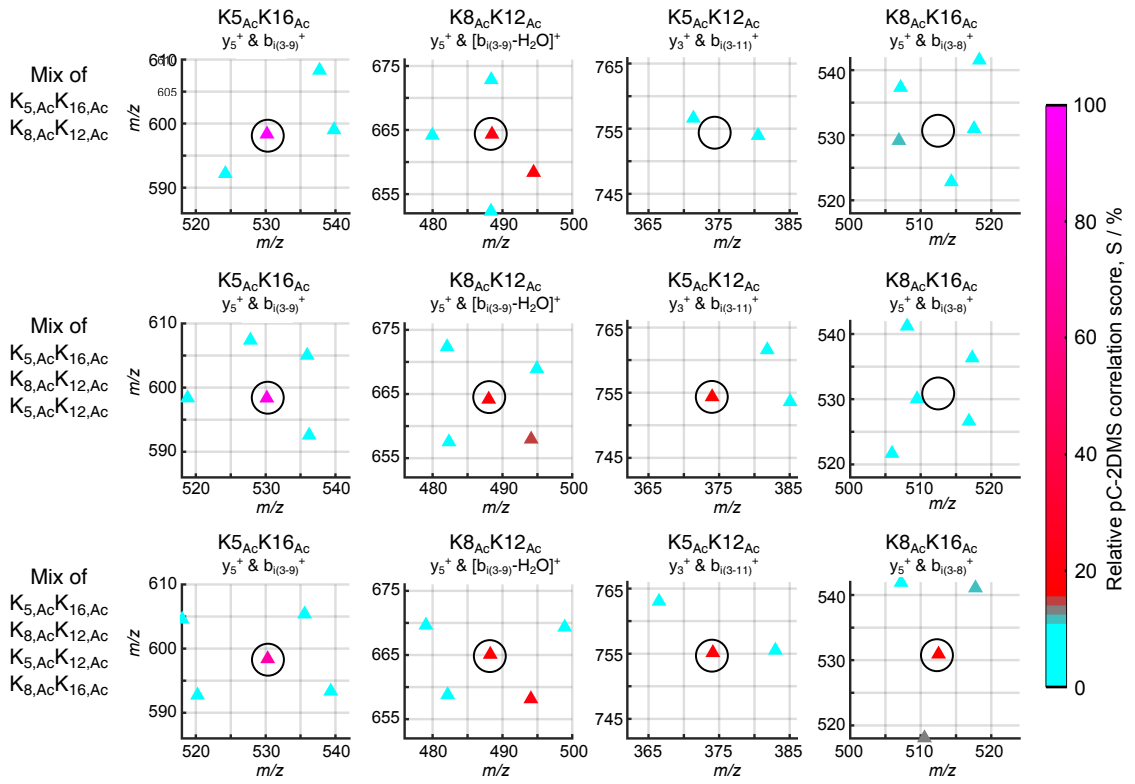


Figure 9.4: Direct resolution of the three mixtures of diacetylated histone H4 peptides using pC-2DMS. For each structural isomer, only one of the measured unique signal channels is required and therefore shown.

Thanks to these unique signal channels, pC-2DMS was used to straightforwardly resolve different mixtures of the four diacetylated histones. In Fig. 9.4, pC-2DMS unambiguously resolves each of the three prepared mixtures of structural isomers which cannot be resolved by 1D MS/MS (Table 9.1, [30, 31]) into its constituent parts. By interrogating only one of the many measured unique pC-2DMS signal channels for each isomer, the presence or absence of each of the four isomers in the hitherto indistinguishable mixtures is directly revealed. This is the first successful experimental realisation of the resolution of these complex and biologically significant [30, 31, 150] mixtures, and the fully general underlying principle of correlating two non-unique fragments to produce a unique 2D signal channel should allow mass spectrometry to provide hitherto unavailable insight into extant and important biological questions regarding combinatorial modification patterns [29, 145, 149].

Importantly, pC-2DMS also enables the straightforward quantitative analysis of these mixtures. On a simple covariance map the volume of a covariance feature depends linearly on the probability of the physical process which produces the feature happening [42, 43], which can be useful in e.g. extracting the branching ratio of different physical processes [151]. In the case of Poissonian fluctuations, the covariance inherits the useful property of the variance, that it is directly proportional to the mean of the Poisson variables for which it is calculated. In the less straightforward case of the total ion count partial covariance used for pC-2DMS, analysis of the formal theory presented in Chapter 8 predicts that, in the case of a fragmenting biopolymer, the volume of a TIC partial covariance feature is directly proportional to the central event rate ν_0 (i.e. the average number of parent ions fragmented at each scan) and the branching ratio of the process producing the two correlated fragments. This suggests that pC-2DMS is also amenable to quantitative analysis, and as well as being used to extract the branching ratios for different fragmentation pathways (which could prove useful for e.g. the ongoing research into the fragmentation pathways of fragmenting biomolecular ions introduced in section 5.3) can also be used to determine the concentration of different species in a mixture (which will be proportional to ν). To demonstrate this capability, pC-2DMS was used to trace the concentration of the isomer $K_{5,Ac}K_{16,Ac}$ in a four-fold mixture with the other isomers $K_{8,Ac}K_{12,Ac}$, $K_{5,Ac}K_{12,Ac}$ and $K_{8,Ac}K_{16,Ac}$, see Fig. 9.5. Here, the molar concentration of $K_{5,Ac}K_{16,Ac}$ has been progressively increased from 1 to 16 arbitrary units and measured the ratio of the volume of the highest scoring partial covariance feature unique to $K_{5,Ac}K_{16,Ac}$ ($y^+_{5,Ac}$ & $b^+_{i(3-9)}$, found in the top 25 pC-2DMS scored correlations for each of the five mixtures). This volume is normalised to the volume of a partial covariance feature that cannot be produced by $K_{5,Ac}K_{16,Ac}$ and can therefore be taken as a reference. Three such features are found in the top 25 pC-2DMS scored correlations for all five mixtures: y^{2+}_5 & $b^+_{9,2Ac}$, y^+_5 & $[b_{9,2Ac}\text{-}NH_3]^{2+}$ or y^+_5 & $[b_{9,2Ac}\text{-}C_7H_{11}NO\text{-}CO]^+$). It is possible to quantify any component of a mixture in the same manner, using the volume of one of its unique 2D signal channels.

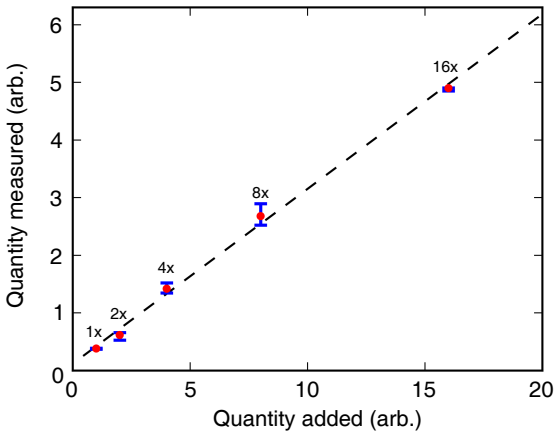


Figure 9.5: Quantitative analysis of four-fold mixture of structural isomers. 'Quantity measured' is the ratio between the volume of the highest scoring partial covariance feature unique to $K_{5,Ac}K_{16,Ac}$ and the volume of a partial covariance feature that cannot be produced by $K_{5,Ac}K_{16,Ac}$ and so can be used as a reference. Error bars show the dependence of the quantification on the particular choice of reference correlation. 'Quantity added' is the actual relative increase in the concentration of $K_{5,Ac}K_{16,Ac}$. The dashed line represents the best linear fit. As predicted by the formal theory of TIC partial covariance (see Chapter 8), the volume of a pC-2DMS feature provides an accurate quantitative measure of the relative concentration of the analyte.

Chapter 10

Reduction in false positive rate for signal identification

Intuition suggests that, given similar m/z accuracy and resolution, a 2D spectrum of a biomolecule should be more specific than a corresponding 1D spectrum; ambiguities in a given 1D spectrum can be resolved along the second dimension. It should also be realised, however, that the position of fragment ion correlation signals on a pC-2DMS map is restricted to varying degrees by the character of the signal - for example, complementary ion correlations arrange themselves along so-called mass conservation lines. In order to investigate the structural specificity of pC-2DMS correlation signals, numerical simulations were performed to determine the structural specificity of pC-2DMS signals *vs*. the corresponding 1D fragment ion signals, by determining the false positive rate for the matching of 2D correlation signals/1D fragment ion signals with theoretical signals generated *in silico* from a random selection of peptide sequences derived from the UniProtKB/Swiss-Prot database [152]. The false positive rate for matching signals is the fundamental measure for the likelihood that a given full sequence match between an experimental spectrum and a protein sequence in a database search engine is a random match (false positive) [153], which is a measure of confidence in protein identification used by a number of database search engines such as Mascot [154, 155]. The smaller the false positive rate for matching individual signals, the more likely an experimental sequence will be to match with the correct

peptide sequence and the smaller the likelihood that the match is simply a random event.

The numerical simulations were all performed in the Python language, making use of part of the software developed for the automated database search engine described in Chapter 11. First, all protein sequences in the UniProtKB/Swiss-Prot database were downloaded and, ignoring the 0.47% of protein sequences with ambiguous amino acid residue coding, subjected to an *in silico* tryptic digest. The peptide digestion products were limited to a minimum length of 5 residues and a maximum molecular weight of 2000 Da. From these peptides, five sets of 10,000 sequences were selected at random (five different sets were chosen in order to be able to check the consistency of results across different sets). The peptide sequences in each randomly chosen set were fragmented *in silico* to form every possible (b^+ ion, internal b^+ ion, y^+ ion) triplet that could in theory be produced from the fragmentation of the triply protonated ion of the sequence. These triplets were then used to generate the example 1D fragment ion signals and 2D fragment ion correlation signals for which the false positive rate was determined. Herein, these signals are referred to as the test signals.

The chronology presented is chosen for clarity of explanation and does not represent the chronology of the algorithm itself, which was optimised on computational efficiency. The procedure outlined below was repeated for each test signal, from each different class of test signals analysed (see the first column of Table 10.1 for the classes analysed). First, the relevant test signal was extracted from the triplet created above (for example, if the class of 2D internal b^+ ion/y^+ ion correlations were being analysed, the m/z values of the internal b^+ ion, y^+ ions were extracted and formed a fragment ion correlation). Then, the randomly selected parent sequence from which the triplet was produced was identified, and in accordance with standard database searches [156] the entire pool of peptidic products from the tryptic digest of the protein database was searched to identify all sequences producing a 3+ peptide ion of m/z value within 7 parts-per-million (ppm) of the m/z value of the 3+ ion of the randomly selected parent peptide. The 3+ precursor m/z value was also used to determine whether the test signal was measurable in a linear ion trap (this was

simply determined by whether the m/z value of the test signal was greater than $\frac{1}{3}$ of the value of the 3+ ion of its parent sequence [93]). In a typical database search the set of peptides which are close in mass to a measured peptide under analysis, and so selected as possible matches for an experimental spectrum, are known as candidate peptides. This terminology is adopted here. The set of candidate sequences identified for each peptide sequence is negligibly sensitive to the choice of parent charge state used to generate the considered m/z values, but the 3+ m/z value was chosen to match the parent charge state which would produce a triplet of singly charged fragment ions, as above.

Next, each of the candidate sequences was fragmented *in silico* to produce the same triplets of every possible (b^+ ion, internal b^+ ion, y^+ ion) as generated above for the five sets of 10,000 randomly chosen peptides. From these triplets, according to the test signal being analysed (see Table 10.1), the relevant signal(s) were extracted for comparison to create a large pool of candidate signals.

Finally, the test signal in question was compared to each of the generated candidate signals under a specified m/z tolerance to identify whether they matched in m/z value. The selected tolerances were chosen to represent typical values used in database searches for different classes of mass analyser: ± 0.8 Da - linear ion trap mass analyser, ± 0.05 Da - time-of-flight mass analyser, ± 0.02 - Orbitrap mass analyser, ± 0 Da - theoretical ideal mass analyser (only isomeric fragments matched). The false positive rate (FPR) for each of the test signals x was calculated according to:

$$FPR(x) = \frac{M(x, s_c)}{N(P_c)}, \tag{10.1}$$

where $M(x, c)$ is the number of matches between the test signal x and all candidate signals s_c generated from the list of candidate peptide sequences P_c, and $N(P_c)$ is the number of candidate peptide sequences P_c determined for the parent ion of x. For each class of signal and each m/z tolerance, the result was averaged over all considered test signals x of length between 2 and 15 residues. The results for test signals x involving fragment ions of length 1 were omitted, because the selective cleavage of

trypsin resulted in a significant artificial increase in the FPR for signals involving y_1 ions. Trypsin cleaves C- terminal to lysine (K) and arginine (R) residues, provided neither is preceded by a proline (P), which means that the overwhelming majority of y_1 ions generated for either the test signals or candidate signals, both of which were generated from a tryptic digest, were either $[K]^+$ or $[R]^+$ and therefore generated a large number of false positive matches.

Ion type	Matched against
1D fragment ions	
b-ion	b-ion, y-ion
y-ion	b-ion, y-ion
internal b-ion	b-ion, y-ion, internal b-ion
2D fragment ion correlations	
b-ion / internal b-ion	b-ion / internal b-ion, internal b-ion / y-ion
internal b-ion / y-ion	b-ion / internal b-ion, internal b-ion / y-ion
b-ion / y-ion	b-ion / y-ion

Table 10.1: Table showing the fragment ion types for which false positive analysis was performed

Table 10.1 identifies the different classes of signals for which the FPR was estimated, and the classes of signal against which each different signal was compared. The classes of signal used for comparison directly relate to the database searches which exploit such matching. In a database search, the test signals are the experimentally measured signals and the candidate signals are those produced *in silico* from the database(s) being searched. The 1D b-ion and y-ion test signals were only compared with 1D b-ion and y-ion candidate signals, which mimics the typical 1D search engine where b-ion and y-ion signals are only considered as a result of the higher false positive rate associated with matching internal b-ions (demonstrated below in Fig. 10.1). The 1D internal b-ions were compared with 1D b-ion, y-ion and internal b-ion signals, now simulating the situation where a database search engine also considers internal b-ions but is, inevitably, unable to identify whether a fragment signal comes

from a b-ion, y-ion, internal b-ion from the m/z value alone. The situation changes slightly in the case of 2D test signals. As discussed in Chapter 6, the geometric position of a signal on a pC-2DMS map provides information on the physical origin of that signal. The fact that correlations between complementary ions (i.e. terminal b-ion/y-ion) arrange themselves along well defined mass conservation lines means that a) an experimental 2D b-ion/internal b-ion or internal b-ion/y-ion signal can be immediately identified as not being a b-ion/y-ion signal by virtue of not falling along the mass conservation line, and b) the correlated m/z value corresponding to the terminal b-ion or y-ion can be experimentally identified by virtue of the fact it aligns with a corresponding b-ion/y-ion complementary correlation on the mass conservation line. This means that in a database search, b-ion/internal b-ion and internal b-ion/y-ion correlations avoid being matched with b-ion/y-ion correlations thanks to their position on the map, although it is not straightforwardly possible to distinguish between b-ion/internal b-ion and internal b-ion/y-ion correlations in the same way. It should be noted that mass conservation lines can appear from secondary fragmentation, so can appear due to e.g. different b-ion/internal b-ion correlations resulting from the fragmentation of a b-ion primary fragmentation product, but the primary mass conservation lines carrying the complementary terminal b-ion/y-ion correlations is easily identifiable using Eq. 13.2 (see page 150).

Figure 10.1 shows the estimated FPR calculated for 1D b^+ ions (green bars), internal b^+ ions (magenta bars) and y^+ ions (blue bars) at the different m/z tolerances described above. The dotted red line, just visible above the x-axis, shows that by correlating exactly the same set of fragment ions to form fragment correlations between b^+ ions and internal b^+ ions and internal b^+ ions and y^+ ions, the false positive rate for signal matching is dramatically reduced. This reduction is achieved even at the modest m/z tolerance of ± 0.8 Da to which the 2D calculations were restricted to remain consistent with the mass accuracy of the linear ion trap mass analyser used in this work. It is clear from Fig 10.1 why many 1D search engines choose to emit matching internal ions [107] - this is due to the significant FPR associated with matching these ions (magenta bars). The increased FPR is primarily a result of the

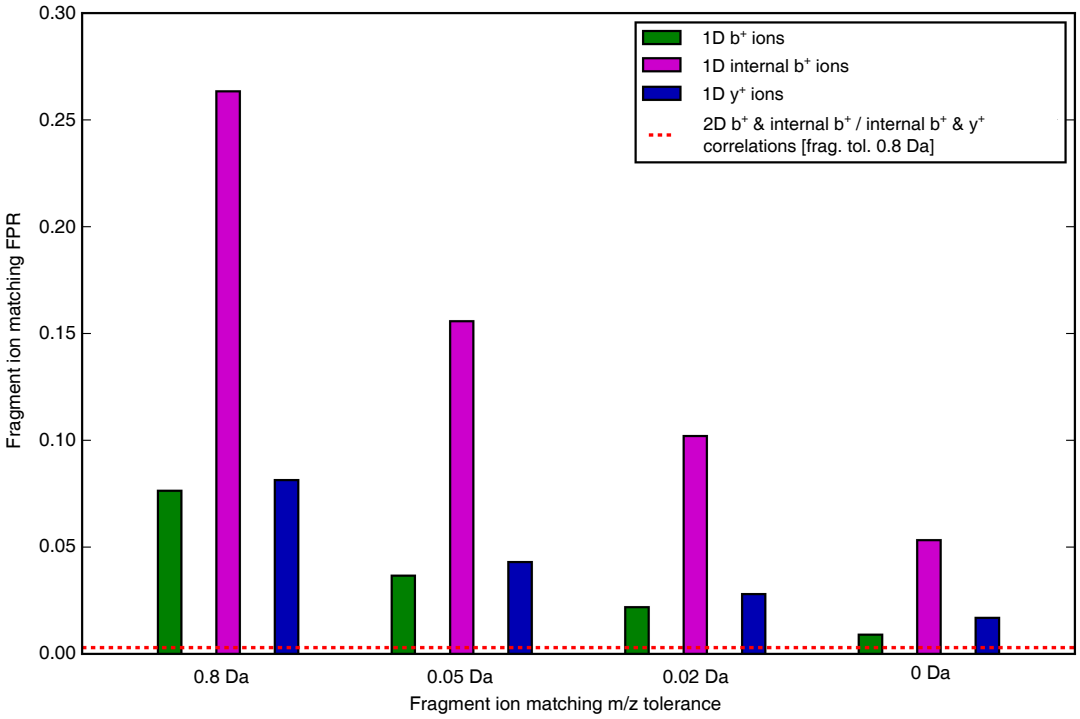


Figure 10.1: Calculated estimated FPR for 1D fragment ion signals (bars) *vs* 2D fragment ion correlations involving internal fragments (dashed red line) at different m/z matching tolerances. The structural specificity of 2D internal ion/terminal ion correlation signals, at the modest fragment ion tolerance of ± 0.8 Da used for the work in this thesis, is dramatically higher than the structural specificity of the 1D fragment ion signals even in the 'ideal' case of infinite mass accuracy and resolution for 1D measurements.

large number of different internal sub-sequences it is possible to generate from a given peptide sequence; assuming a minimum length of 2 for standard b-type ions and 1 for y-type ions [108] this number is $(n-2)(n-3)/2$, compared to $n-1$ possible b-ion sub-sequences and n possible y-ion sub-sequences. However, at an m/z tolerance of ± 0.8 Da, the FPR for matching such an internal b^+ ion (0.26) drops by almost two orders of magnitude when the ions are correlated with one of their corresponding terminal b^+ ions or y^+ ions (FPR of 0.0046 and 0.0011 respectively, giving an average of 0.0030[1]). Even the 'ideal' 1D case of matching only b^+ ions and y^+ ions at *infinite* m/z resolution (0.0089 and 0.017 respectively, average 0.012) provides a FPR of more than 4 times greater than matching b^+/internal b^+ and internal b^+/y^+ ions. Despite the astounding success that improvement in mass accuracy and resolution has brought [4, 36], these results demonstrate (albeit at a highly theoretical level, necessitating a huge amount of further work) the fundamental limits of high mass resolution techniques and the potential of pC-2DMS to outperform 1D MS on decades-old hardware.

These results were confirmed to be highly replicable, with an average relative standard deviation (RSD) of 0.68% between the five sets of 10,000 randomly chosen peptides used to generate the test signals, across all 1D and 2D signal classes. The maximum RSD was for the 2D internal b^+/y^+ ion correlations at ± 0.8 Da (2.0%) and the minimum was for 1D internal b^+ ions at ± 0.05 Da (0.12%). The slightly larger FPR across all m/z tolerances for y^+ ions *vs* b^+ ions is an artefact of the tryptic digest used to produce both the test signals and the candidate peptides - as discussed above, the selective cleavage of trypsin C-terminal to lysine and arginine produces digestion products with highly similar C-termini.

The FPR for matching b^+/y^+ correlations between two complementary terminal ions, at an m/z tolerance of 0.8 Da, was also investigated. This was found, at 0.039, to be lower than the corresponding FPRs for matching the individual b^+ and y^+ ions at the same m/z tolerance (0.076 and 0.081 respectively) but did not replicate the

[1]This is the average FPR across all b^+ and y^+ ions, instead of simply the mean of the two individual values.

dramatic reduction observed for correlations featuring terminal ions. This can be understood by the tight 7 ppm parent mass tolerance used to select the candidate sequences, which was performed to reproduce the behaviour of a standard database search engine [156]. As a result, every test signal was only matched against candidate signals generated from a candidate sequence very close in mass to the parent sequence that produced the test signal. In this case, each time a test signal b^+ ion or y^+ ion matches with a candidate signal b^+ ion or y^+ ion, the probability that the complementary y^+ ion or b^+ ion from the candidate sequence matches the complementary y^+ ion or b^+ ion from the test sequence is artificially increased, because the mass of the 'remaining' part of the test sequence will be extremely close to the mass of the 'remaining' part of the candidate sequence. This can be visualised by noting that the primary (complementary terminal ion) mass conservation lines for peptides of very similar mass all lie very close to each other on the pC-2DMS map. Without this pre-selection on parent mass, the FPR for 2D b^+ ion/y^+ ion correlations can also be expected to show a dramatic decrease compared to the corresponding 1D case.

Chapter 11

Development of automated database search engine

The requirement for high-throughput protein identification has driven a large research effort into the automatic analysis of MS and MS/MS spectra [157]. A standard bottom-up experiment can produce thousands of individual MS/MS spectra over a full run, orders of magnitude more than could be reasonably analysed manually. The different methods for automatic analysis of MS/MS spectra can be broadly divided into three categories, where protein sequences and post-translational modifications are inferred from the experimentally measured spectra by either: i) comparing to large libraries of previously recorded experimental spectra of known sequences [19], ii) comparing to theoretical spectra generated *in silico* from databases of known protein and/or genetic sequences, according to generalised peptide fragmentation rules [2, 17, 18] or iii) so-called *de novo* sequencing, where a first-principles reconstruction is performed directly from the measured experimental spectrum without recourse to any database [20, 158].

Although all three methods provide different advantages, the lack of suitably large spectral databases and strong dependence of fragmentation patterns on fragmentation method and experimental conditions renders option i) challenging, and option iii) is generally limited by the incomplete information available from a standard MS/MS spectrum. As a result, the most prevalent method of automated spectral interpreta-

tion in use today is the comparison of experimentally measured spectra to theoretical spectra generated *in silico* from sequence databases, which nonetheless suffers from numerous drawbacks. A highly variegated problem is the variability of peptide decomposition patterns, which are affected by the particular amino acid sequence, the presence of PTMs, peptide length, tertiary structure, etc. [24–26]. This frequently leads to considerable deviation of the experimental mass spectra from those predicted by the simplified peptide fragmentation rules. Additionally, many of the peptide fragments predicted by these simplified fragmentation rules (e.g. b- and y-type ions for CID) [2] can appear at very low relative abundances, and so are often missed during the spectrum-to-structure matching procedure. This results in incorrect assignments even from mass spectra displaying standard fragmentation patterns. Moreover, 1D m/z-based MS analysis is compromised by false identifications caused by fragments being attributed to isobaric (within the given mass tolerance) and isomeric ions of incorrect structures, see Chapter 10 for further information. Nonetheless, despite these numerous problems, many database search engines based on theoretically generated spectra have proved highly successful for the correct analysis of vast quantities of spectral data [154, 159, 160].

Fig 11.1 shows a simple schematic for a typical database search. The experimental spectrum is matched against a large number of different theoretical spectra, and the database sequence which provides the best matching theoretical spectrum is typically designated as the experimentally measured sequence. Some of the most successful search engines, such as Mascot by Matrix Science [154], provide a corresponding score for the identified sequence which measures the confidence that the matched theoretical sequence is correct. In the case of Mascot, the 'ions score' which is reported for an MS/MS match is the value of $-10 \log_{10} P$, where P is the estimated probability that the match between an experimental and a theoretical spectrum is a random event. A threshold score can be used, according to a user-specified p-value, above which a match is deemed to have a sufficiently high score to denote a sequence identification. For example, if the user desires the probability that the match is due to a random event is one in twenty or less (corresponding to a p-value of 0.05) and there are

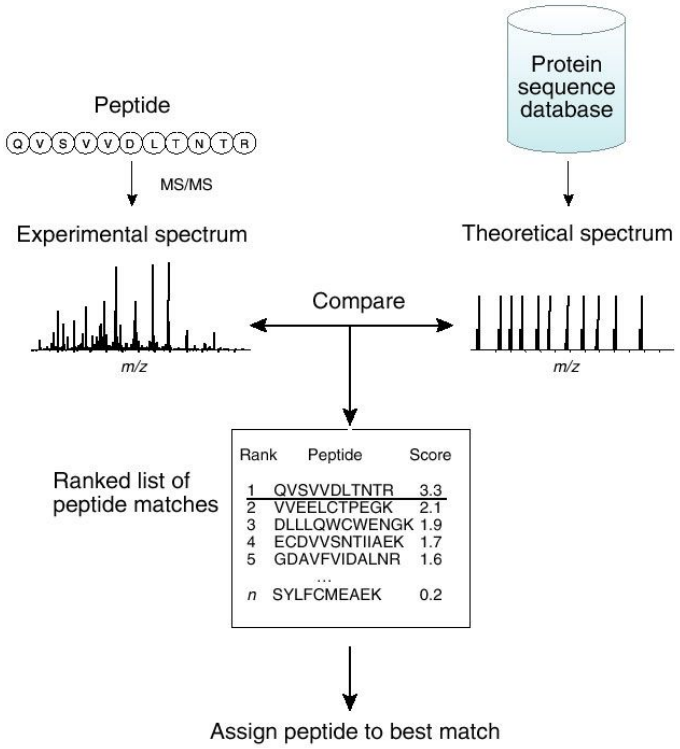


Figure 11.1: Figure adapted from [157]. A simple schematic of method ii) for the analysis of MS/MS spectral data. Experimental spectra are compared with theoretical spectra generated *in silico* from databases of known sequences and standardised peptide fragmentation rules. As illustrated by the uniform peak heights in the theoretical spectrum, the highly complex task of predicting so-called ion intensity relationships (IIRs) from database sequences [25] dictates that the theoretical spectra are generally considered as a list of m/z values, with no prediction made on fragment ion relative abundances.

1000 database sequences being tested, for a fully random set of database sequences the threshold score would be[1] $-10\log_{10}\left(\frac{1}{20\times1000}\right) \approx 43$. Another widely used search engine, SEQUEST [159], calculates a compound cross-correlation function between the experimental spectrum and each theoretical spectrum, which increases with the number of fragment ion matches, the abundances of the experimentally measured fragment ions, the number of consecutive fragment ions matched and the number of characteristic immonium ions observed, and is normalised to the overall number of theoretical predicted sequence ions to account for the increase in the probability of a random match which follows an increase in the number of possible matches.

Whichever method is used to compare experimental and theoretical spectra, the fundamental unit of similarity between the two are matched fragment ion signals. Therefore, the success of all such search engines ultimately rests on the selectivity of the fragment ion matching process. As demonstrated in Chapter 10, where the false positive rate for matching 1D fragment ion signals was calculated to be almost two orders of magnitude higher than the rate of false positives for matching corresponding pC-2DMS fragment ion signals, the selectivity provided by pC-2DMS fragment ion correlations is dramatically higher than that provided by 1D fragment signals. This was even the case in the theoretically ideal scenario of matching terminal b- and y-ions at infinite m/z tolerance. This suggests that a database search engine based on matching 2D fragment ion correlations alone could provide unprecedented success in selecting the correct sequence from a list of potential database matches.

To investigate this further, a prototypical pC-2DMS search engine was created with spectral matching based entirely on pC-2DMS correlations. All parts of the search engine were written by the author in Python. The search engine accepts as input a protein sequence database and list of experimental pC-2DMS correlations with their associated pC-2DMS correlation scores. It outputs a ranked list of peptide or protein sequences (according to whether the experimentally measured data were from a peptide or intact protein sequence), each with their own similarity score in

[1]This simple calculation has been found in practice to be too high, so an empirical correction of -13 is applied to these threshold values.

comparison to the experimental spectrum. As shown below, it has been found that in the tested cases the search engine provides a single sequence with an outstanding similarity score, which can be confidently assigned as the correct match to the experimental spectrum.

Algorithmic Details Figure 11.2 shows a detailed schematic of the prototypical pC-2DMS search engine. The search engine first accepts as input a database of protein amino acid sequences. The capability to translate genetic sequences to amino acid sequences has not been included in the prototype, and is functionality which will be incorporated in future work. This protein database is digested *in silico* according to the user specified 'enzyme' ('enzyme' is given a broad definition, such that for e.g. top down searches the specified 'enzyme' is 'no cleavage' and the digestion simply returns the protein database). In this process, a single protein sequence produces every possible peptidic digestion product that it could according to the specified parameters. As well as standard selective enzymatic digests, which only cleave between given combinations of amino acid residues (e.g. trypsin, which cleaves C-terminal to lysine [K] and arginine [R] except when either is N-terminal to a proline [P]), the search engine is also able to perform a non-selective digest which cleaves between every possible pair of residues in a protein sequence. In the implementation of the *in silico* digest, the computational overhead associated with the vast number of digestion products produced by a non-specific digest is limited by performing m/z-based precursor selection (see below) on-the-fly.

If a modification has been specified by the user, the resultant peptide digestion products are also 'modified *in silico*'. In line with 1D database search engines [160], the user can specify the chemical form of the modification to be considered (phosphorylation, acetylation, amidation, etc.) and the residue on which this occurs (serine, lysine, the peptide C-terminal residue, etc.). These modifications can be specified as fixed, meaning every occurrence of the specified residue features the modification, or variable, meaning all possible combinations are considered with the specified residue either modified or not. The number of combinatorial possibilities which need to be

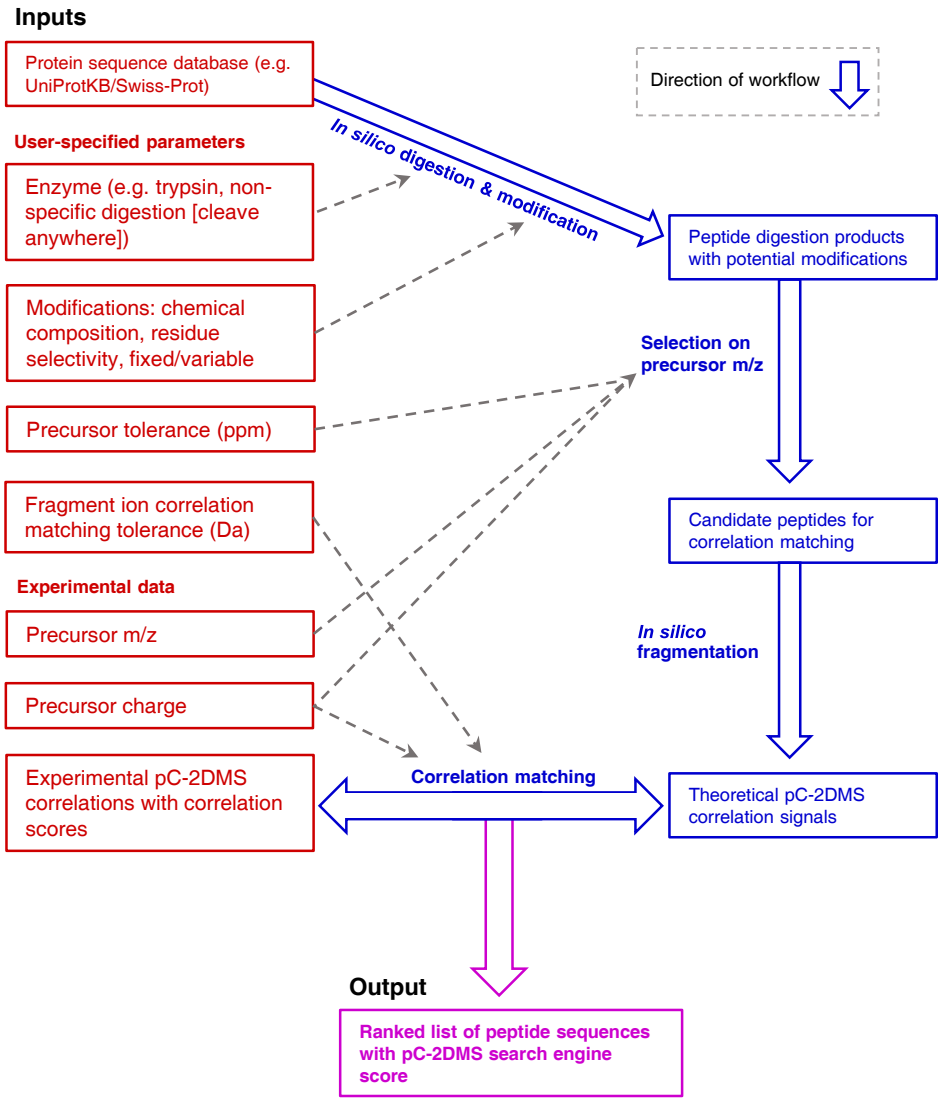


Figure 11.2: Schematic for the prototypical pC-2DMS search engine based on fragment ion correlations.

considered for variable modification increases the search space significantly.

The resultant peptide digestion products are a pool of all possible peptide sequences and modification states which could be produced from the user-specified enzyme and modifications. These peptide digestion products are filtered according to a user-specified precursor tolerance, to produce a pool of candidate peptide sequences which produce a peptide ion of m/z value within a certain distance from the experimentally measured precursor m/z value at the experimentally measured precursor charge state. These precursor measurements (often known as full MS scans) are typically performed by a high-resolution mass analyser such as an Orbitrap[2], affording a tight precursor tolerance (e.g. 7 parts-per-million), and therefore the number of filtered candidate peptides is orders of magnitude smaller than the total number of digestion products.

Finally, the candidate peptides are subjected to pC-2DMS correlation matching. The pC-2DMS correlation scores for the fragment ion correlations, which have already been filtered so that they feature the top scoring signals only, are currently normalised so that they sum to unity[3]. Next, each of the remaining possible candidate peptide sequences from the database is fragmented *in silico* according to experimentally derived fragmentation rules. The three following categories of fragment ion are produced by the *in silico* fragmentation to be considered for comparison with the experimental correlation signals:

i) terminal b-ion & y-ion correlations

ii) terminal b-ion & y-ion correlations followed by a neutral loss of either H_2O or NH_3 from one or both of the correlated fragment ions, either with or without CO loss from the b-ion producing an a-ion

iii) terminal b-ion & internal b-ion and internal b-ion & terminal y-ion correlations, including those followed by a neutral loss of either H_2O or NH_3 from one or both

[2]For testing of both 1D and 2D search engines, such high-accuracy measurements were simulated by manual adjustment of the precursor m/z measured in a linear ion trap (this was not done for the searches in Fig. 14.7 where the precursor mass was determined by the Hough transform).

[3]Identification of an absolute threshold for the pC-2DMS correlation score, see section 7.1, would preclude this step.

of the correlated fragment ions, either with or without CO loss from any of the b-type fragment ions producing an a-type ion.

For correlations in category i) and ii), all fragment ion correlations where the charge state sums to the measured precursor charge state Z or lower are considered. For correlations in category iii), all fragment ion correlations where the charge state sums $Z-1$ are considered, except when $Z=2$ in which case all fragment ion correlations where the charge state sums to Z are considered. Limiting the possible charge states for type iii) correlations in this way was found to increase selectivity of the 2D search engine, and reflects the fact that for precursor ions with $Z \geq 3$, type iii) correlations are rarely observed without being accompanied by the loss of at least one charge. Note that although CO loss can only occur from a b-type ion (either terminal b-ion or internal b-ion), loss of both H_2O and NH_3 is not dependent on the type or amino acid sequence of the fragment ion. This is a departure from some 1D database search engines, which restrict these small molecule neutral losses to only the most likely residues (see page 62) in order to try to improve search specificity. It was found that for the pC-2DMS prototype search engine, this increase in specificity was negatively offset by the fact that this often caused important correlation signals to be missed.

Every candidate peptide sequence is given three similarity scores, corresponding to the three different categories of fragment ion correlation. For every theoretical correlation produced by the *in silico* fragmentation of a candidate database sequence that matches an experimentally measured correlation within the user-specified m/z tolerance (± 0.8 Da for all linear ion trap measurements in this work), the relevant similarity score for the database peptide is incremented by the normalised pC-2DMS correlation score for the experimental signal. Once all possible theoretical-to-experimental correlation matches have been considered[4], the final pC-2DMS search engine score

[4]As was done for the false positive analysis in Chapter 10, the experimental correlations could be pre-separated according to their geometric position on the map, which enables one to immediately identify e.g. terminal b-ion & y-ion correlations and so negate the requirement to test these correlations against theoretical type ii) or iii) correlations. In practice, the precursor m/z filtering already nullifies the danger of any such avoidable inter-category matching so any pre-separation would be for computational efficiency only and is not performed here.

for each peptide sequence is calculated by the weighted sum of the three individual similarity scores. The weights for this sum were empirically optimised across all sample peptide sequences which were experimentally measured, and are as follows[5]: 0.8 for type i) correlations, 0 for type ii) correlations and 1 for type iii) correlations. A weight of zero for type ii) correlations is a result of the fact that in pC-2DMS, the appearance of a b-ion & y-ion correlation with at least one ion suffering neutral loss of a small molecule (H_2O, NH_3, CO) without detection of the corresponding intact b-ion & y-ion correlation is extremely rare. The highest score for type iii) correlations is in agreement with the higher calculated estimated false positive rate for matching these correlations.

Results The pC-2DMS search engine was first tested on a well understood set of sample sequences, for which the state-of-the-art 1D MS/MS search engine technology (Mascot [154]) confidently identifies the correct sequences from the corresponding 1D CID scans. The results of these searches are shown in Fig. 11.3. All searches were performed over a non-specific digest of the UniProtKB/Swiss-Prot database [152], at precursor m/z tolerance 5 ppm and fragment ion tolerance 0.8 Da. The user-specified potential modifications are given in the caption. As is performed for all 1D (Mascot) and 2D database searches throughout this work, the precursor ion mass was manually adjusted in order to simulate the standard MS/MS workflow of a high mass accuracy precursor ion measurement. The results show that the pC-2DMS search engine is able to identify the correct sequence from a huge pool of candidate peptide sequences (note the logarithmic scales on the y-axes in Fig. 11.3) by attributing a considerably outstanding pC-2DMS search engine score to the correct database sequence. This is despite the fact that the pC-2DMS search engine is so far only a prototype, with a naïve sequence scoring method based only on a weighted sum of the pC-2DMS correlation scores of matching fragment correlations. Further developed versions of the pC-2DMS search engine, which could incorporate a more sophisticated function for theoretical-to-experimental sequence matching as in SEQUEST [159] or

[5]Thanks to Nikhil Bachhawat, whose work provided these optimised weights.

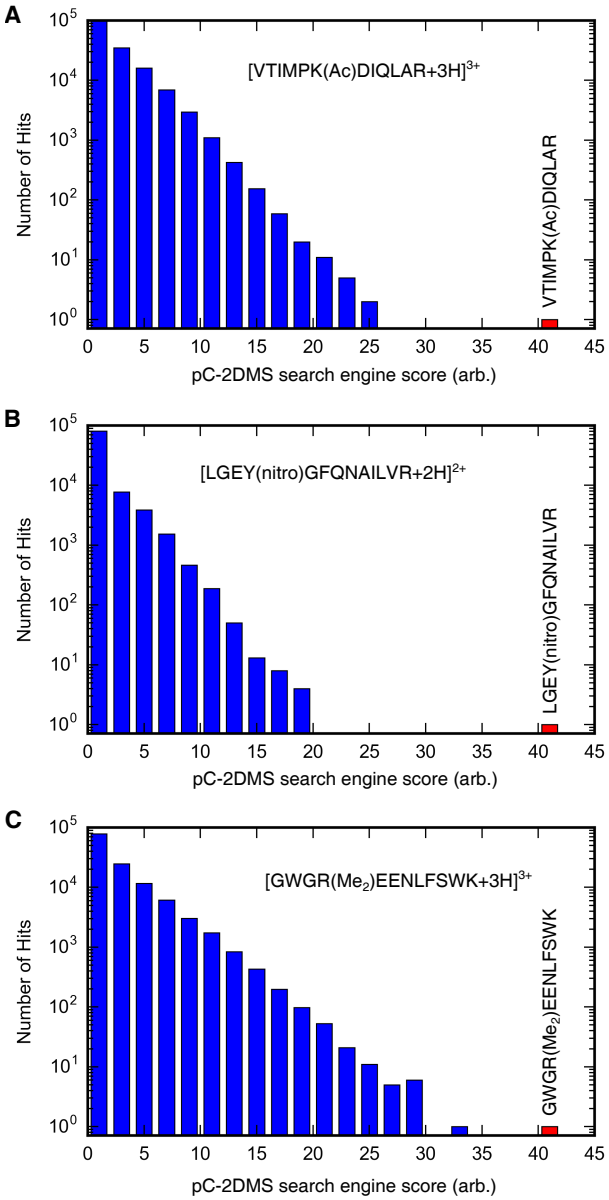


Figure 11.3: pC-2DMS search engine results for three different peptide sequences. Note the logarithmic scale on the y-axes of all plots, showing the outstanding selectivity demonstrated by the search engine. Modifications specified were: **A**: lysine acetylation (variable), **B**: tyrosine nitration (variable), **C**: arginine dimethylation (variable). In panel **B** the isomeric database sequences LGEY(nitro)GFQNALLVR and LGEY(nitro)GFQNALIVR, which were trivially awarded the same pC-2DMS search engine score, are not plotted on the histogram. For 2+(3+) parent ions, the top 40(50) pC-2DMS correlation score-ordered signals were provided to the search engine.

a probabilistic approach with an associated confidence threshold as for Mascot [154], can be reasonably expected to vastly outperform this prototype.

The performance of the pC-2DMS search engine in complex situations where the state-of-the-art fails has also been investigated. Although Mascot was found to perform extremely well in the majority of situations, the introduction of contaminant fragment ion signals disrupts its success. To simulate the common scenario of co-isolation and co-fragmentation of contaminant species [161], the amyloid beta peptide GSNKGAIIGLM was prepared in an equimolar mixture with its unnatural (and so not in any sequence database) palindromic isomer MLGIIAGKNSG. Because precursor ion selection and isolation for MS/MS is performed on m/z, it is impossible to separate such isomers prior to fragmentation and measurement. The doubly protonated ions were selected for CID, and both standard 1D MS/MS and pC-2DMS scans were taken for the isolated ion ensembles. To ensure optimal signal-to-noise for the 1D MS/MS scans, they were averaged over 5000 microscans. The data from the 1D MS/MS and pC-2DMS scans were fed to Mascot and the prototypical pC-2DMS search engine, respectively, with the same search parameters specified for both: non-selective digest of the UniProtKB/Swiss-Prot database, precursor m/z tolerance 5 ppm, fragment ion m/z tolerance 0.8 Da. As above, the top 40 pC-2DMS correlation score-ordered signals were provided to the pC-2DMS search engine, whilst Mascot iteratively optimises the number of 1D MS/MS fragment ion signals to consider from the 1D MS/MS spectrum[6].

Fig. 11.4 compares the performance of the state-of-the-art 1D search engine (Mascot) with the prototype 2D search engine in the complex case of co-isolation and co-fragmentation of a contaminant ion. The contaminant fragment ion signals cause the MS/MS Ions Search results returned by Mascot (left) to fail to identify the correct natural sequence from the database, which is instead outscored by three different, highly dissimilar sequences. Mascot only returns the ten highest-scoring sequences from the MS/MS Ions Search, presented in the table. In comparison, the enhanced

[6]Implementation of a similar method for correlation signal selection is one way in which the performance of the pC-2DMS search engine could be further improved.

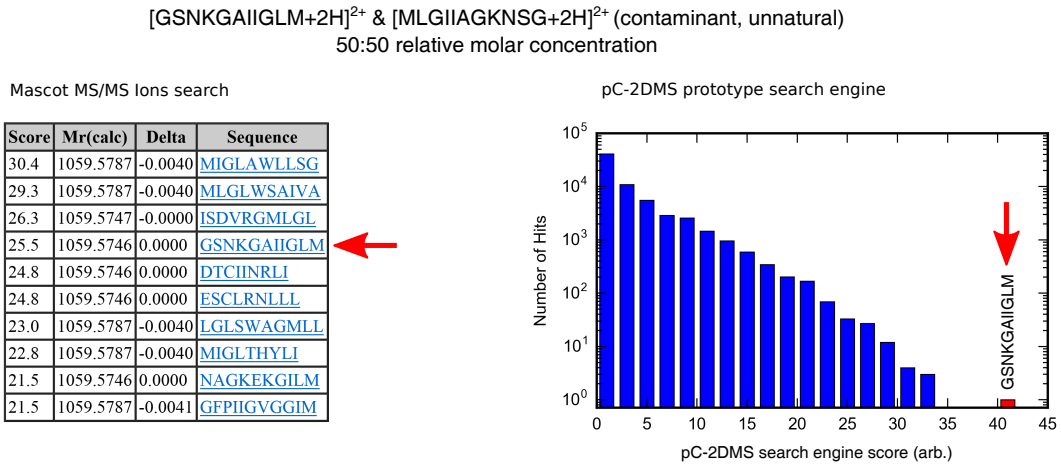


Figure 11.4: The prototypical pC-2DMS search engine outperforms the state-of-the-art 1D database search in the complex case of co-isolation and co-fragmentation of an abundant contaminant ion. The red arrow indicates the correct database sequence.

selectivity of pC-2DMS fragment ion correlation signals (Chapter 10) enables the prototypical pC-2DMS search engine to confidently identify the correct sequence from the equivalent pC-2DMS measurement, by awarding the correct database sequence an outstanding search engine score despite the presence of an abundant contaminant ion (right). The scores for all correlation-matched database sequences are plotted, demonstrating the immense selectivity offered by the engine. It is worth noting again that the results for the pC-2DMS search engine have been achieved without the huge additional layer of performance over simple fragment ion signal matching which is provided by Mascot.

Dependence on number of scans The dependence of the sequence identification provided by the pC-2DMS search engine on the number of scans used to calculate the pC-2DMS map was also investigated. Figure 11.5 shows the dependence on number of scans of the search engine score given to each of the database sequences which are subjected to correlation matching, i.e. those which make it past the precursor m/z-based filtering, for two different pC-2DMS measurements. The results are presented for the straightforward case of the pure peptide ion $[\text{VTIMPKDIQLAR+3H}]^{3+}$ and the complex case of the doubly protonated ions $[\text{GSNKGAIIGLM+2H}]^{2+}$ and

$[MLGIIAGKNSG+2H]^{2+}$ from an equimolar mixture of the two sequences shown in Fig. 11.4, where the state-of-the-art 1D approach fails. The pC-2DMS search engine parameters used are the same as those in Figs. 11.3 and 11.4, with no modification specified for either search.

For the ion $[VTIMPKDIQLAR+3H]^{3+}$, the correct sequence emerges well before 2000 scans and remains the top scoring sequence thereafter, whilst the same happens in the complex mixture of the two palindromic sequences very shortly after 2000 scans. Although this demonstrates that the current number of scans required for confident automatic identification of a peptide sequence in pC-2DMS is five times lower than the number acquired for optimal visual presentation of the pC-2DMS maps, even in the more complex case which the 1D state-of-the-art is unable to resolve, the number is larger than that suggested by analysis in Fig. 7.5 where on average it is possible to extract ~ 25 good correlation signals with high ($> 80\%$) fidelity from well below 1000 scans. This discrepancy is likely a result of the fact that the performance of the prototype search engine has so far been optimised using the signals generated from the full 10,000 scan sets for each sequence, parameters which were maintained for every search performed in Fig. 11.5. For example, the number of pC-2DMS correlation score-ranked signals provided to the search engine was, as optimised for the sample set of pC-2DMS measurements at 10,000 scans, 40 for 2+ ions and 50 for 3+ ions. It is clear from Fig. 7.5 that at lower numbers of scans this figure is an overestimate and will result in some false signals being provided to the search engine. One method of bypassing this problem would be to introduce an absolute threshold for pC-2DMS correlation score to determine true signals from statistical noise. Moreover, as described above, it was empirically discovered at 10,000 scans that for parent ions of charge 3+ and above the performance of the search engine is improved by limiting the *in silico* production of correlations including internal b-ions to cases where there has been at least one charged loss from the molecule. This may not improve the performance in the same way at lower numbers of scans. The inset of panel (**A**) in Fig. 11.5 shows different behaviour at lower number of scans when this condition is loosened, and theoretical correlations are considered regardless

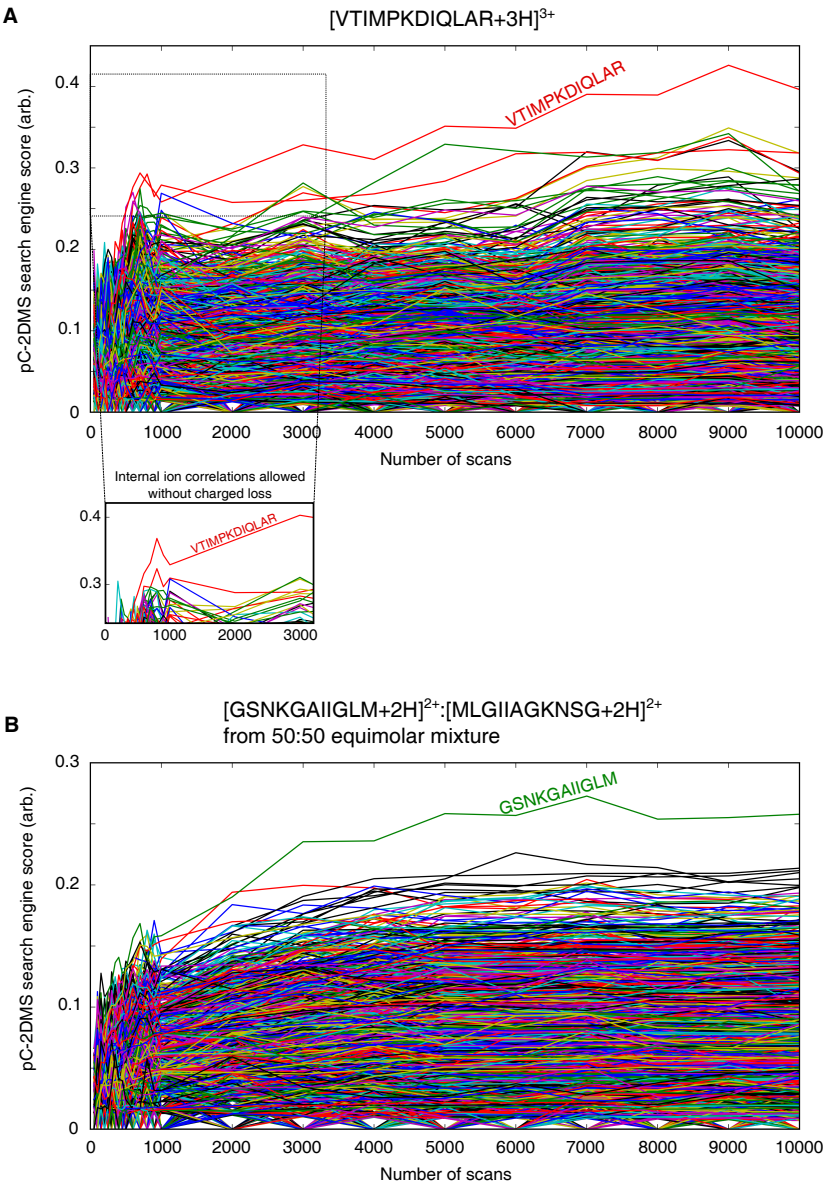


Figure 11.5: Dependence of search engine score on number of scans for all correlation matched sequences from the pC-2DMS search for measurements of (**A**) $[\text{VTIMPKDIQLAR+3H}]^{3+}$ and (**B**) $[\text{GSNKGAIIGLM+2H}]^{2+}$ and $[\text{MLGIIAGKNSG+2H}]^{2+}$ in an equimolar mixture. The correct structure emerges as the highest-scoring sequence after only a small fraction of the full number of scans taken for the measurements, even in the complex case of an equimolar mixture of two isomers. However, it is likely the number of scans at which this happens can be pushed significantly lower (see discussion in text).

of whether charged loss has taken place or not. In this case, the correct sequence emerges well before 1000 scans. It remains the highest scoring sequence up to and including the case of the full 10,000 scans.

In conclusion, a prototypical pC-2DMS search engine, written entirely in-house in Python, has been developed, which identifies an experimentally measured sequence from a database based on the matching of fragment ion correlations. This prototype has been shown to identify the correct sequence from a large sequence database, and even outperforms the otherwise highly successful state-of-the-art 1D database search engine in the complex case of co-isolation and co-fragmentation of an abundant contaminant ion. Being so far only a prototype, the scoring mechanism for database sequences is naïve: it is simply the weighted sum of the pC-2DMS correlation scores of matched experimental correlations. Considering the fruits borne by the vast and ongoing research efforts of last three decades to improve 1D database search engines beyond simple fragment ion matching, the potential of a pC-2DMS search engine is great.

Chapter 12

Identification of fragmentation pathways

The information provided by pC-2DMS correlations gives mechanistic insight into the processes through which the detected fragment ions are formed. All such information is fundamentally inferred from the experimental confirmation provided by pC-2DMS that two fragment ions are produced by the same, often compound, fragmentation reaction of the same parent ion. Fragment ion correlations allow us to probe the behaviour of the *single molecule* in a way that traditional 1D MS, which tells us about the behaviour of the *molecular ensemble*, cannot. By way of example to illustrate this fundamental difference, consider the simple (albeit rather contrived) example of two different sets of fragmentation mechanisms, **A** and **B**, for the production of a complementary b-ion and y-ion pair b_n^+ & y_m^+ from an ensemble of the peptide molecule X^{2+}. In scenario **A**, b_n^+ is produced by the initial fragmentation of X^{2+} to the complementary pair b_{n+1}^{2+} & y_{m-1}, followed by a secondary decomposition of the b_{n+1}^{2+} fragment ion to the b_n^+ ion [25], $X^{2+} \rightarrow \mathrm{b}_{n+1}^{2+}$ & y_{m-1}, $\mathrm{b}_{n+1}^{2+} \rightarrow \mathrm{b}_n^+$. Similarly, y_m^+ is produced by the initial decomposition of X^{2+} to the complementary pair b_{n-1} & y_{m+1}^{2+}, followed by loss from the y_{m+1}^{2+} ion to form y_m^+, $X^{2+} \rightarrow \mathrm{b}_{n-1}$ & y_{m+1}^{2+}, $\mathrm{y}_{m+1}^{2+} \rightarrow \mathrm{y}_m^+$. In scenario **B**, the X^{2+} molecule dissociates to form the b_n^+ & y_m^+ pair directly, $X^{2+} \rightarrow \mathrm{b}_n^+$ & y_m^+. This is, of course, the more likely scenario for the production of such a pair from a doubly charged parent molecule, but this example focuses on what

is directly observable from the measurement of the spectra.

Now consider a measurement of the two different scenarios **A** and **B**, each displaying different single-molecule behaviour, using both 1D MS and pC-2DMS. In 1D MS, the observable for both scenarios is the same - there will be one peak measured at the m/z value of b_n^+ and one at the m/z value of y_m^+. In this way, 1D MS is unable to identify between two different behaviours of the single molecule because the response of the molecular ensemble (i.e. b_n^+ ions are produced and y_m^+ ions are produced) is the same. However, a pC-2DMS measurement is directly able to discriminate between **A** and **B**. In scenario **B**, a correlation between b_n^+ and y_m^+ is observed, and in scenario **A** it is not (in contrast, correlations can be expected between b_n^+ & y_{m-1}^+, b_{n+1}^+ & y_{m-1}^+, b_{n-1}^+ & y_m^+ and b_{n-1}^+ & y_{m+1}^+). By the same argument explaining the absence of negative covariance between competing pathways (see Chapter 2), none of the X^{2+} molecules in the highly diffuse measured ensemble are able to 'communicate' with each other and therefore a pC-2DMS correlation will only appear between two fragment ions if they happen to be produced in the same fragmentation mechanism, *of the same molecule*. It is in this way that pC-2DMS is able to provide information on the single molecule behaviour response which is inaccessible to a 1D measurement of the same fragment ion signals. Of course, there have been a huge number of complex mechanistic studies performed using 1D MS/MS from which is it has been possible to determine the fragmentation pathways of a large range of biopolymers [9, 25, 162–165], but the unique information directly available from the pC-2DMS map is unmatched in direct comparison with a 1D MS/MS spectrum.

It is possible to use the information from pC-2DMS correlations to directly infer much more complicated fragmentation pathways than the given example. Under linear ion trap CID (normalised collision energy 35%), the doubly protonated ion of the peptide sequence GSNKGAIIGLM (peptide **4**) produces a series of high pC-2DMS correlation scoring features which correlate a canonical expected m/z value (corresponding to b[/a]-ions, y-ions, internal b[/a]-ions and related small neutral losses) with an unexpected m/z value which does not feature in the expected set of m/z values. These correlations are the labelled red features in the 3D scatter plot shown

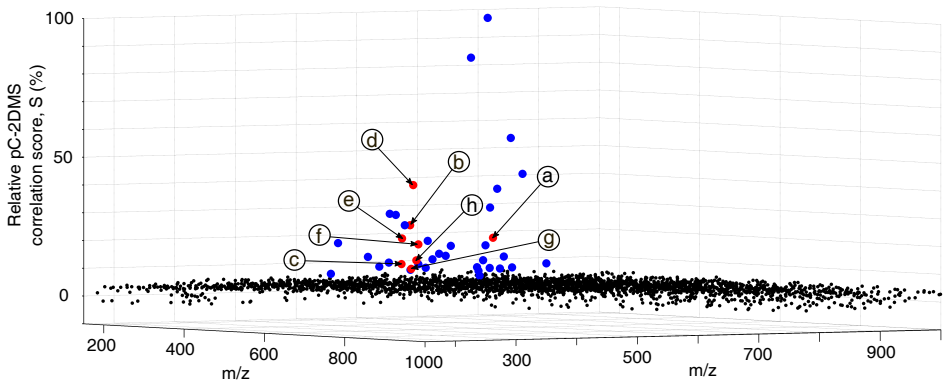


Figure 12.1: Scatter plot of the pC-2DMS correlation features for the doubly protonated peptide ion of sequence **4**, $[\text{GSNKGAIIGLM+2H}]^{2+}$. The horizontal x- and y-axes denote the m/z values of the two correlating fragments, and the vertical axis is the pC-2DMS correlation score calculated according to Eq. 6.2. The features contributing to Fig. 12.3 are plotted in red, with labels consistent across both figures. All these signals appear in the top 40 correlation score-ordered features, the rest of which are shown in blue.

in Fig. 12.1. Despite the presence of these unexpected correlation features, the specificity of pC-2DMS correlations enables the pC-2DMS search engine to unambiguously identify the correct primary structure (Fig. 12.2).

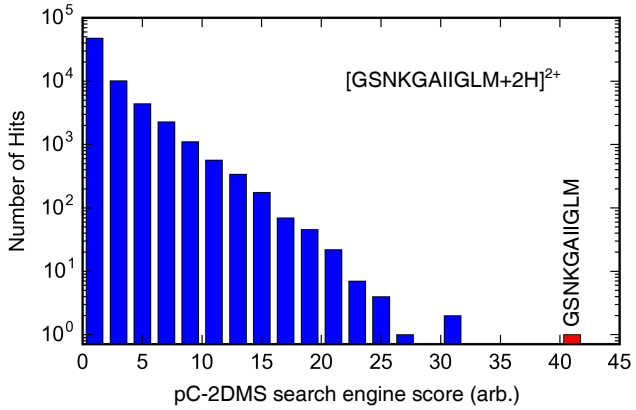


Figure 12.2: The specificity of pC-2DMS correlations allows the pC-2DMS database search engine to unambiguous identify peptide **4**, despite the presence of unexpected 'sequence scrambling' products.

Moreover, following identification of the correct sequence, the fact that the unexpected m/z values are correlated with canonical fragment ions allows for both the unambiguous identification of the unexpected fragment ion m/z's and consequently

a full reconstruction of the fragmentation history of these signals. The reconstructed fragment ion formation tree for each of the labelled signals in Fig. 12.1 is shown in Fig. 12.3. Here, the pC-2DMS correlations enable the identification of four individual 'sequence scrambling' routes [166] (highlighted in purple) of the original peptide sequence. The identity of the sequence scrambling products has been confirmed by MS^3 scans, for which some representative examples are given in appendix B.

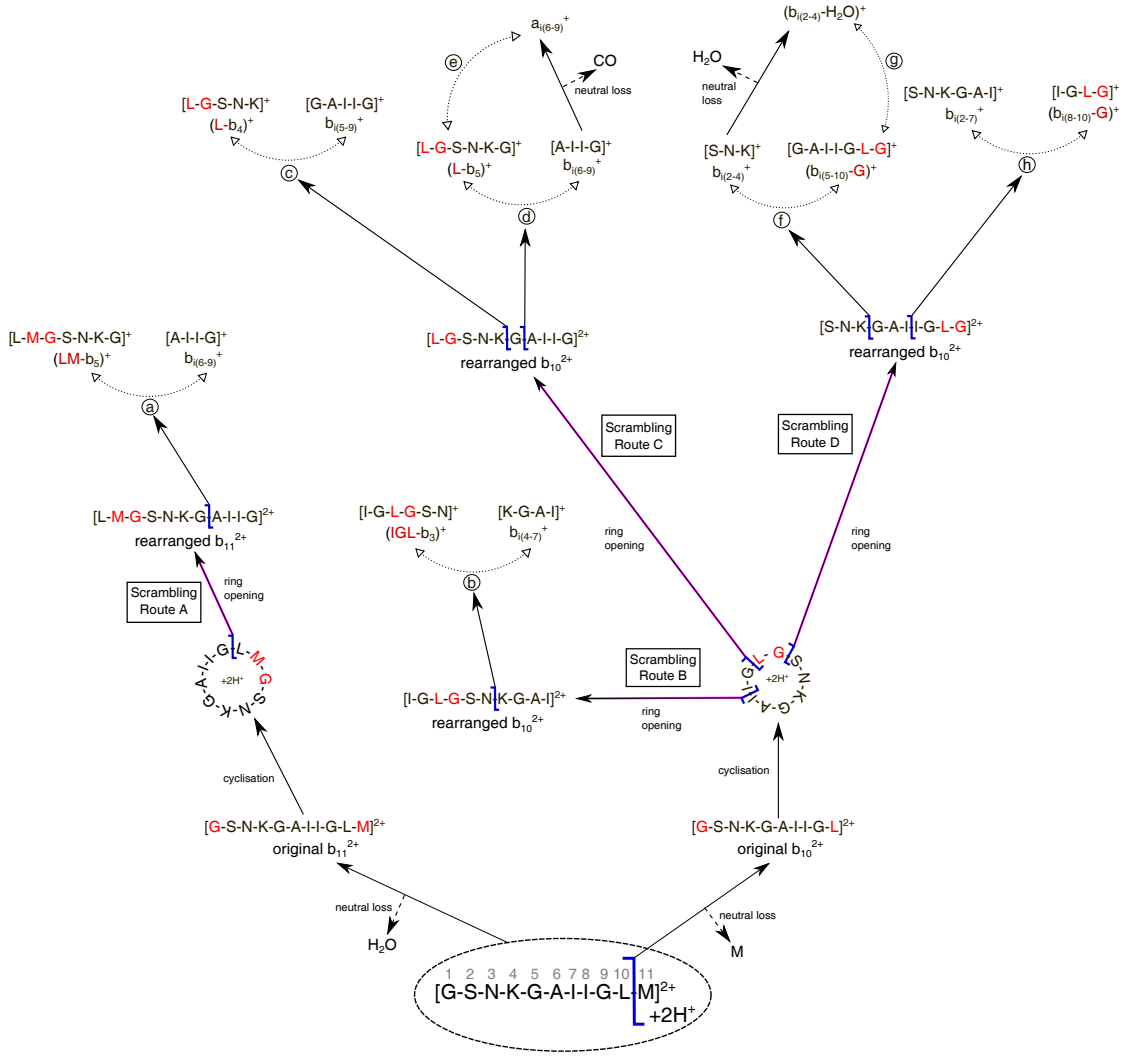


Figure 12.3: Fragment ion formation tree for products of sequence scrambling under CID, constructed from the pC-2DMS spectrum of the doubly charged ion of peptide **4**, $[GSNKGAIIGLM+2H]^{2+}$.

Chapter 13

pC-2DMS determination of charge state

The first requirement for the successful analysis of fragment ion spectra is the correct identification of the fragment ion masses from the measured m/z ratios, which of course requires determination of the fragment ion charge states. pC-2DMS provides a robust and novel two-dimensional method to determine the charge state of both fragment ions and the parent ion directly from the pC-2DMS map, as well as the mass of the fragmenting parent ion(s).

The canonical method for identification of fragment ion charge states in 1D MS/MS is to exploit the small natural abundance of heavier isotopes (e.g. ^{13}C at $\sim 1.1\%$ and ^{15}N at $\sim 0.4\%$) to determine the charge of a fragment ion by measuring the m/z difference between the resultant isotopic peaks which are separated in mass by 1 atomic unit. The pattern produced by the small abundance of these heavier isotopes is commonly referred to as the isotopic envelope, and consecutive peaks in this isotopic envelope are separated in m/z by $\frac{1}{z}$, where z is the charge of the fragment ion. In order to correctly infer the charge state of the fragment ion from the isotopic envelope, an accurate measurement of the m/z difference between two peaks in the isotopic envelope is required. This becomes very challenging for low mass resolution instruments or highly charged fragment ions, and the relatively low mass resolution of the linear ion trap mass analyser means that fragment ions of charge 5+ or greater are

typically impossible to identify, because the isotopic envelope is unresolved (see e.g. Fig. 14.3). Hitherto, this has significantly limited the application of linear ion trap mass spectrometers, desirable for their speed, sensitivity and relative affordability [55], to the analysis of intact protein molecules and highly charged parent ions.

The pC-2DMS method presents a novel way to extract the charge state from the fragment ion spectrum, without the requirement of sufficiently high mass resolution or sufficiently low charge state to resolve the isotopic envelope. Instead, the gradient of the mass conservation line upon which a correlation falls can be used to determine the charge states of both correlating ions.

As discussed in Chapter 6, all correlations of primary fragment ions originating from the same parent ion but due to cleavages at different bonds in the ion are arranged along so-called mass conservation lines. This is a direct result of the fundamental physical principles of mass and charge conservation. Take a correlation between two m/z values m_1/z_1 & m_2/z_2. Due to the symmetry of pC-2DMS maps this same physical correlation produces exactly two correlation islands on the pC-2DMS map; at $x = m_1/z_1, y = m_2/z_2$ and $x = m_2/z_2, y = m_1/z_1$. For the following discussion we will consider the set of correlation islands on the pC-2DMS map described by $x = m_1/z_1, y = m_2/z_2$, where for each physical correlation z_1 and z_2 are determined by $z_1 \geq z_2$. This set covers all physical correlations measured on the pC-2DMS map.

If the two correlated ions measured at m_1/z_1 & m_2/z_2 are the complete primary products of a single dissociation of a parent ion with mass Mr_{Par}, then the sum of the masses of these two correlated ions must be equal to the mass of the parent ion that dissociated to produce them. In other words, the conservation of mass and charge dictates that the following equality holds true:

$$((m_1/z_1) \times z_1) + ((m_2/z_2) \times z_2) = Mr_{Par} \tag{13.1}$$

This means that all correlation islands which correlate the primary products from the decomposition of a parent ion of mass Mr_{Par} will fall along one of the straight lines

of the form:

$$(x \times z_1) + (y \times z_2) = Mr_{Par}$$

which can be rearranged to:

$$y = -\frac{z_1}{z_2} \times x + \frac{Mr_{Par}}{z_2}, \tag{13.2}$$

and therefore defines a straight line with gradient $-\frac{z_1}{z_2}$ and y-intercept $\frac{Mr_{Par}}{z_2}$.

It follows that all correlation islands which are due to the correlation of the primary dissociation products of a fragmented ion[1] of mass Mr_{Par} and charge $z_{Par} = z_1 + z_2$ will appear on the pC-2DMS map along a straight line of gradient $-\frac{z_1}{z_2}$ and y-intercept $\frac{Mr_{Par}}{z_2}$. For the reciprocal set of correlation islands where $x = m_2/z_2, y = m_1/z_1$ and $z_2 \geq z_1$, all correlation islands will fall on a straight line with gradient $-\frac{z_2}{z_1}$ and y-intercept $\frac{Mr_{Par}}{z_1}$.

The result of the above is that if it is possible to determine a set of correlation islands lying on a particular mass conservation line, both the ratio between the charge states $\frac{z_1}{z_2}$ and the value of $\frac{Mr_{Par}}{z_y}$ can be identified. Provided limited knowledge on the charge state of the parent ion (e.g. an upper limit), this generally enables the charge state of each of the correlated fragment ions, as well as the charge state and mass of the parent ion responsible for the mass conservation line (if not already known), to be inferred.

Shown in Fig. 13.1 is a scatter plot of the top 50 significance ordered correlation islands from the pC-2DMS map of the dissociation of the triply protonated peptide **1** ([VTIMPK$_{\mathrm{Ac}}$DIQLAR+3H]$^{3+}$) under CID at normalised collision energy 35%. There are 10 correlation islands (red circles) lying along the red line, which has gradient=-2 and y-intercept=1428.6. Following the above analysis, all correlation islands lying along this line with coordinates $x = m_1/z_1, y = m_2/z_2$ result from the correlation of two ions at m/z m_1/z_1 and m_2/z_2, with $z_1 = 2+$ and $z_2 = 1+$. Additionally, from the gradient and y-intercept of this mass conservation line all 10 correlation

[1]This doesn't have to be the intact precursor ion, and could e.g. be a terminal fragment ion which further fragments to a series of terminal ion/internal ion correlations.

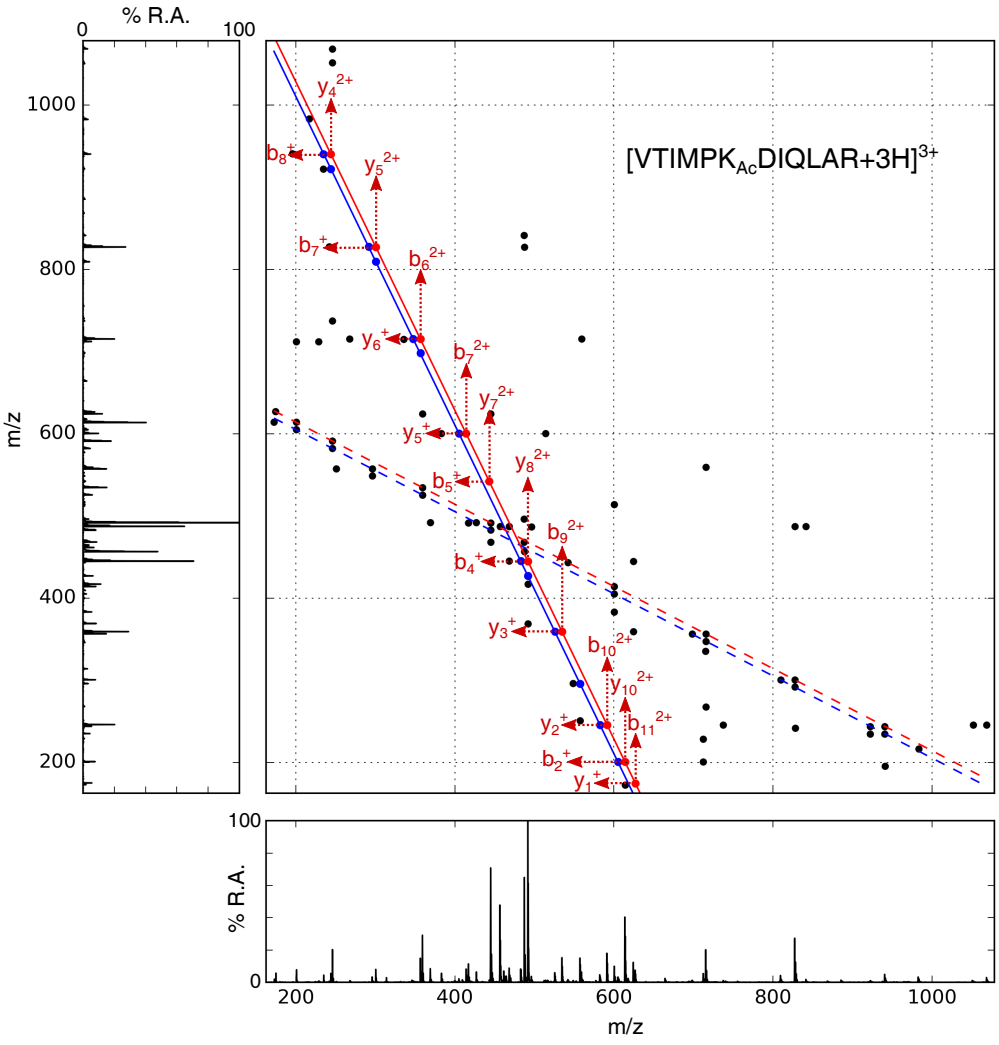


Figure 13.1: Scatter plot of the top 50 pC-2DMS correlation score-ordered correlation islands from the pC-2DMS map of the dissociation of the triply protonated ion of peptide **1**, showing the mass conservation lines formed due to the dissociation of the parent molecule (red), and dissociation of the parent molecule minus small neutral loss (blue). The averaged 1D mass spectra are plotted along the each side of the pC-2DMS scatter plot.

islands are inferred to have arisen from the dissociation of a parent molecule of charge $z_{Par} = z_1 + z_2 = 3+$ and mass $Mr_{Par} = z_1 \times y\text{-}intercept = 1 \times 1428.6 = 1428.6$ Da, which corresponds to the mass of the M+3H parent ion (expected mass 1428.8 Da). This has been confirmed by the manual assignment of all 10 correlation islands along the mass conservation line, with fragment ion charge state verified using the isotopic envelope in the 1D mass spectrum.

The blue line, along which 13 blue correlation islands are plotted, has gradient=-2 and y-intercept=1410.7. This corresponds to a parent ion of charge 3+ and $\mathrm{Mr_{Par}}$ 1410.7, i.e. the correlation of two primary products from the parent ion minus neutral loss of a water molecule (expected at $1428.8 - 18.0 = 1410.8$). The assignments of these correlation islands are not plotted for visual clarity. The two dashed lines show the mass conservation lines for the reciprocal set of correlation islands where $x = m_2/z_2$ and $y = m_1/z_1$.

Chapter 14

Extension to larger species

'Top-down mass spectrometry' refers to the tandem mass spectrometric analysis of intact protein ions [5], bypassing the digestion step[1] in the canonical bottom-up proteomic workflow. The rapidly developing field of top-down proteomics is driven by a desire to achieve full proteoform coverage which is inaccessible to bottom-up proteomics, including holistic information on the sequence and its post-translational modifications (PTMs) across the entire protein structure. Such information is often destroyed by enzymatic digest in bottom-up workflows. Following sample preparation, the first step in the mass spectrometric analysis of intact protein molecules is their introduction into the gas phase in ionised form, for which, as for smaller peptides, electrospray ionisation (ESI) [14] (see section 5.1) has emerged as the method of choice. However, due to the larger number of residues available to be protonated/deprotonated on longer sequences, the gaseous ions of intact protein molecules typically appear at higher charge states than peptide ions, which significantly complicates the analysis of both their molecular weight and their fragment ion spectra. The fragmentation spectra of intact protein ions, containing a complex mix of ions at different charge states, have until now required isotopic resolution of fragment ion peaks in order to facilitate correct interpretation of the MS/MS spectra. The necessary resolving power is only available from Fourier transform-based mass analysers such as Fourier transform ion cyclotron resonance (FT-ICR) and Orbitrap analysers

[1]The cleaving of whole proteins into smaller peptide molecules by enzymes [167]

[36] or quadrupole time-of-flight mass analysers [168]. This is a forced departure from the favoured mass analyser for bottom-up MS/MS scans, the linear ion trap [55] (see section 5.2), preferred due to its unprecedented cycle-time and sensitivity as well as its relative affordability. The significantly lower mass resolution available from the linear ion trap has limited its successful application to top-down proteomics studies to a few select cases [169–171].

pC-2DMS offers a novel workflow for the analysis of intact proteins using a linear ion trap mass analyser. This makes use of the two-dimensional method, described in Chapter 13, for the charge state identification of both parent and fragment ions directly from the fragment ion spectrum without the requirement to resolve of the isotopic envelope, and also exploits the ability of pC-2DMS to unambiguously identify complementary fragment ions. Top-down pC-2DMS analysis is demonstrated on protein molecules weighing up to 17 kDa, and its power is further shown by using it to immediately deconvolve the complex so-called chimera spectrum[2] resulting from the co-fragmentation of two co-isolated intact protein molecules. Because the *in silico* pC-2DMS separation of multiple co-isolated and co-fragmented parent ions relies solely on the phenomenon of mass conservation lines, it is entirely general to any biomolecular species, e.g. oligonucleotides, peptides or proteins. As well as addressing the very common problem of unintentional chimera spectra [161, 172], this opens the possibility for improving throughput and sensitivity by performing pC-2DMS multiplex measurements. In multiplex measurements [173], multiple parent ions are intentionally co-isolated and co-fragmented and separation of the different fragment ion peaks is performed post-acquisition. Successful realisations of such measurements are currently limited to smaller peptides [174], and even for these simpler molecules the multiplexing capabilities using an ion trap mass analyser are highly restricted by the low mass accuracy of such measurements [175, 176].

The particularly challenging analysis of the fragment ion spectra of whole pro-

[2]In Greek mythology the Chimera is a fearsome fire-breathing hybrid animal, often depicted as a lion with the head of a goat protruding from its back and a snake for a tail. The term is used to describe the fragment ion spectrum of overlapping peaks from the fragmentation of two or more different parent structures.

teins is in part due to the high degree of spectral congestion from a large number of overlapping fragment ions, which is relieved by the introduction of a second fragment m/z dimension in pC-2DMS as demonstrated in e.g. Figs. 14.5 and 14.7. The other main contributing factors to the complexity of top down fragment ion spectra are the large number of different, high charge states which the fragment ions can have, and the much greater possibility at longer sequences for non-canonical sequence ions to be created causing misinterpretation of the fragment ion spectrum [95]. As described in Chapter 13, pC-2DMS provides a uniquely two-dimensional method for extracting the charge state of fragment ions, as well as the mass and charge state of the parent ion, using the phenomenon of 'mass conservation lines'. Moreover, the ions that fall along the primary mass conservation lines are necessarily complementary ions (e.g. b- and y-ions in ion trap CID), enabling the immediate extraction of these useful sequence ions from amongst the non-canonical fragment ions present in these complex spectra.

Given knowledge of the parent ion mass and charge, it is possible to immediately define the set of primary (i.e. correlating two ions resulting from a primary fragmentation event, such that their mass sums to the mass of the parent ion) mass conservation lines for a given parent ion. However, it is not always possible to obtain a high resolution precursor ion scan to detect the charge state (and therefore mass) of the parent ion. It is also common for mass conservation lines to appear along which the mass of the two correlated ions does not sum to the mass of the parent ion, resulting from the secondary decomposition of primary fragment ions. Examples of such cases are the fragmentation of a terminal fragment ion to a smaller terminal fragment ion and internal fragment ion, or the correlation between two terminal fragment ions whose mass sum to the mass of the parent ion, minus a small neutral loss (e.g. H_2O, NH_3 or CO). For such mass conservation lines, an *a priori* definition is not possible. Moreover, the correct identification of these mass conservation lines would provide information to extract the charge state and mass of the ion which decomposed to produce the correlation islands which lie along it, providing otherwise unavailable structural and mechanistic information.

14.1 Restricted Hough transform

The Hough transform provides a means of identifying correlation islands on a pC-2DMS map which lie along the same straight line, and can therefore identify mass conservation lines arising from the decomposition of an unknown or uncertain parent molecule. Further, the Hough transform is able to identify the vertical or horizontal series of correlation islands that appear when one fragment ion A correlates with another fragment ion B as well as a series of derivative ions of B due to neutral losses.

The Hough transform is a computational method first proposed by Hough in 1959 [177] and developed into the form in which it is used here by Duda and Hart in 1972 [178]. It provides a computationally efficient algorithm for feature detection in image analysis, and although since generalised to identify the positions of arbitrary shapes [179] it was originally developed to identify straight lines, which is how it is employed here. Specifically, we make use of the Hough transform to identify straight lines passing through multiple pC-2DMS correlation islands. In this implementation, it works by constructing an accumulator array in ρ, θ parameter space and identifying maxima in this accumulator space as lines which pass through the greatest number of points.

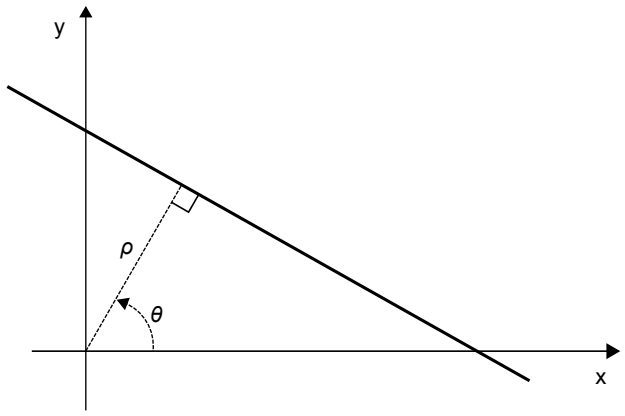


Figure 14.1: The Hesse normal parametrisation of a straight line.

To describe the Hough transform, we begin with the Hesse normal parametrisation

of a straight line. Use of this parametrisation, instead of the classic slope-intercept description of a line, is one of the steps made by Duda and Hart [178]. The parametrisation is illustrated in Fig. 14.1. A straight line is described by the angle between its normal and the x-axis (θ), and the shortest distance from itself to the origin (ρ). Use of Hesse normal form allows for the simpler handling of lines with gradient 0 (horizontal) or ∞ (vertical), an ability which is important for its application to pC-2DMS because such lines correspond to the correlation island series due to one ion correlating with another and a series of neutral losses from the same ion. It also allows for a well-bounded parameter space to be used (all possible lines in an image can be described by a θ in the interval $[0, \pi)$ and $|\rho| \leq d$, where d is the length of the longest straight line that can be drawn across the image), whereas analysis of any image using the slope-intercept parametrisation inevitably requires consideration of intercepts from $-\infty$ to ∞.

A description of the use of the Hough transform to identify sets of points lying along a straight line is now given, using the specific example of pC-2DMS correlation features (the 'points') on a pC-2DMS map (the 'image'). For a mass conservation line in Hesse normal form, ρ and θ are given by:

$$\rho = \frac{Mr_{Par}}{z_1 \cdot \cos\theta + z_2 \cdot \sin\theta};\ \theta = \frac{\pi}{2} - \tan^{-1}(\frac{z_1}{z_2}). \tag{14.1}$$

A two-dimensional ρ, θ accumulator space, like that plotted as a colour map in panel **a** of Fig. 14.2, is initialised as an array of zeros, with suitable sampling[3] chosen for the intervals in ρ and θ. For each feature on the pC-2DMS map, all points in the accumulator space are queried, and for every ρ, θ describing a line that passes through (practically, closer than a given distance from) the feature in question, the corresponding element of the accumulator space is incremented by one. Each feature ends up tracing out a sinusoidal curve in the accumulator space, as can be seen in panel **a** of Fig. 14.2, which shows the accumulator space for all features plotted on the pC-2DMS map in Fig. 13.1. The process is repeated for each feature on the map, so that

[3]A step size of 0.5 Da for ρ and 0.1° for θ were used for this explanatory plot.

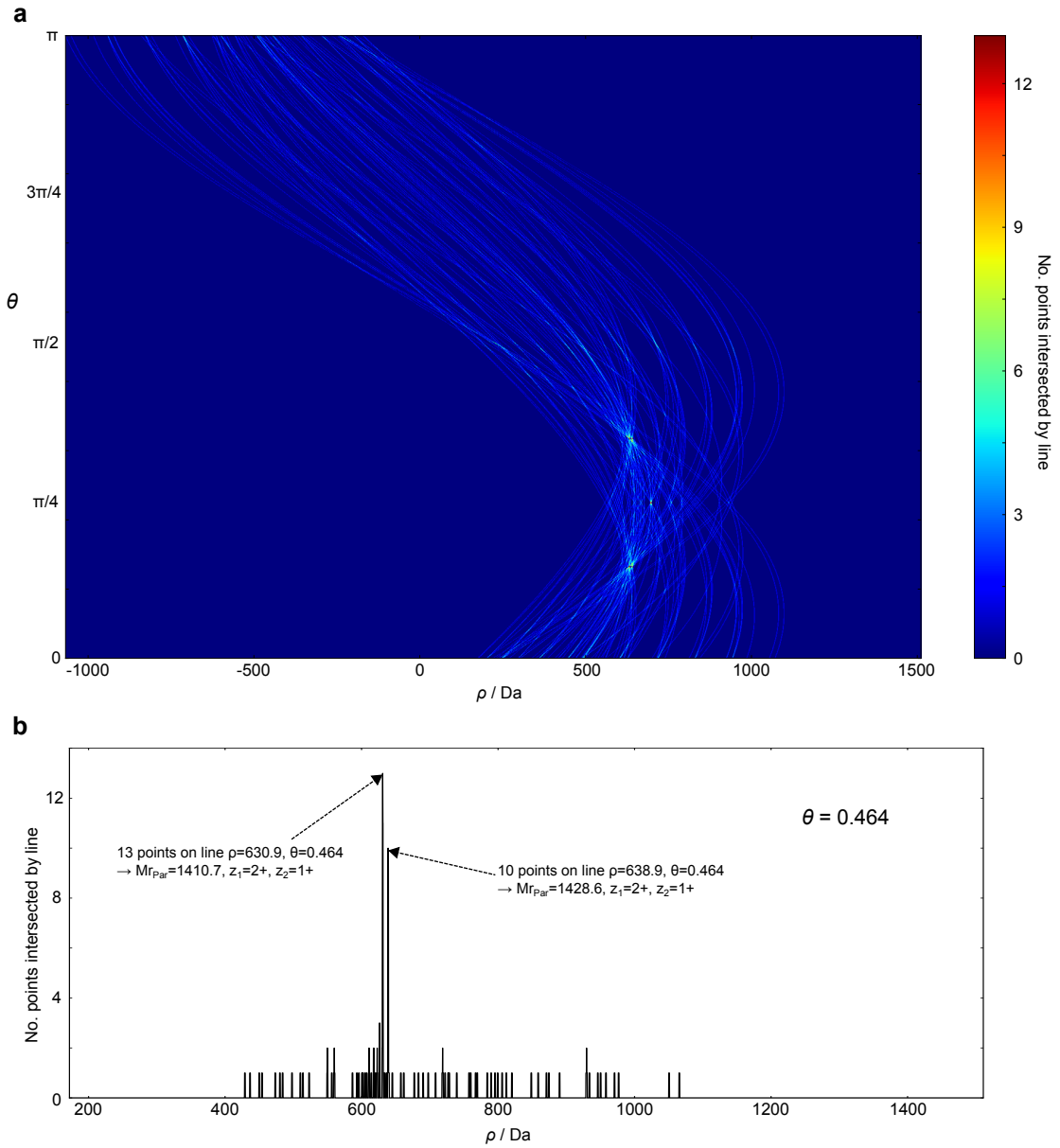


Figure 14.2: Accumulator spaces for the Hough transform. **a**, 2D accumulator space, plotted as a colour map, constructed for all points on the pC-2DMS map in Fig. 13.1. The value of the accumulator space at each ρ, θ is the number of points through which the corresponding straight line passes. **b**, 1D accumulator space plotted for $\theta = 0.464$. Restricting θ also restricts the possible values that ρ can take, as can be seen on the x-axis of this plot compared to the x-axis of panel **a**.

at the end of the procedure the value of each element of the accumulator is equal to the number of features through which the straight line described by the corresponding ρ, θ values passes. Therefore, all lines passing through a certain number of features can be directly read off the accumulator. A script to implement this procedure was written by the author in the Python language, making use of the NumPy numerical computing library.

For the purposes of identifying mass conservation lines on a pC-2DMS map, it is possible to greatly reduce the computational work required to perform the Hough transform. This can be seen by recognising that since all mass conservation lines have gradient $-\frac{z_1}{z_2}$, where z_1 and z_2 each correspond to a discrete charge state and it has been chosen that $z_1 \geq z_2$, mass conservation lines are restricted to a limited set of gradients and therefore a limited set of corresponding θ values. This means that only a very small subsection of the full parameter space need be scanned to identify all mass conservation lines, namely only those slices in θ which correspond to a possible gradient of a mass conservation line. These values of θ are determined only by a maximum value for charge of the parent ion originally fragmented. Restriction of θ also allows for a tightening of the boundaries for ρ, which in the first instance can be chosen as a truncation from $-(m/z)_{max} \leq \rho \leq \sqrt{2 \cdot ((m/z)_{max})^2}$ to $(m/z)_{min} \leq \rho \leq \sqrt{2 \cdot ((m/z)_{max})^2}$ when the interval for θ shrinks from $[0, \pi)$ to $[0, \frac{\pi}{2})$. This restricted Hough transform was implemented as set of consecutive scans across the one-dimensional ρ space, each at a different set value of θ (see panel **b**, Fig. 14.2), and was found to significantly increase the computational speed of the procedure ($\sim 240\times$ speed-up for the pC-2DMS map in Fig. 13.1, $\Delta\theta = 1.0^\circ$ for the full Hough transform).

Due to the symmetry of pC-2DMS maps, every map features two correlation islands corresponding to the same physical correlation. For the case of mass conservation lines where $z_1 = z_2$, both such correlation islands fall on the same mass conservation line because $-\frac{z_1}{z_2} = -\frac{z_2}{z_1} = -1$. Therefore, for $\theta = \frac{\pi}{4}$, the values of the accumulator are halved prior to further processing in order that they represent the true number of physical correlations lying along a mass conservation line.

Limits of 1D MS for top-down measurements in a linear ion trap Figure 14.3 demonstrates the limits of 1D top down measurements using a linear ion trap such as the mass analyser in the LTQ XL. The measurements presented have been taken using the high mass range function of the LTQ XL, which extends the maximum measurable m/z from 2000 Da to 4000 Da. The highest possible resolution for these scans is known as 'Zoom Scan', which corresponds to a scan rate of 1100 Da/s. The top panel shows the measurement of the protonated 13+ ion of intact myoglobin, which has expected (average) m/z 1305.98 Da. The resolution of the linear ion trap is entirely unable to resolve the isotopic envelope of this parent ion, which renders the determination of the parent ion charge state (and therefore mass) impossible with standard 1D MS on the linear ion trap. The eight smaller panels show the isotopic envelopes, as measured by the LTQ XL in the highest resolution 'Zoom Scan', of example fragment ions at charges 4+ to 11+[4].

The example fragment ions all appear at fairly high relative abundance. Whilst the Zoom Scan function of the LTQ XL allows for some resolution of the isotopic envelopes of the y_{35}^{4+} and y_{44}^{5+} ions (thanks greatly to their high relative abundances) which enables albeit uncertain diagnosis of charge state, the isotopic envelopes for ions at charge states greater than 7+ are completely unresolved by the instrumental resolution afforded by the linear ion trap. The insets to each fragment ion peak show the isotopic envelope from the much quicker 'Turbo Scan' mode (125 000 Da/s) which was used to take the pC-2DMS measurements, demonstrating the ability of pC-2DMS to resolve fragment (and therefore parent) charge states even in the complete absence of isotopic resolution. Even for the 4+ charge states, Fig. 14.4, showing Zoom Scan measurements of three different 4+ fragment ions whose identities and charge states have been resolved by pC-2DMS (Fig. 14.5) for high pC-2DMS-scoring correlation features, demonstrates that at lower relative abundances the 1D measurements are unable to resolve the isotopic envelope and therefore the charge state of the fragment ions. As can be seen in the insets for the fragment ion isotopic envelopes in Fig.

[4]The charge states for all these fragment ions has been determined using pC-2DMS (see Fig. 14.5).

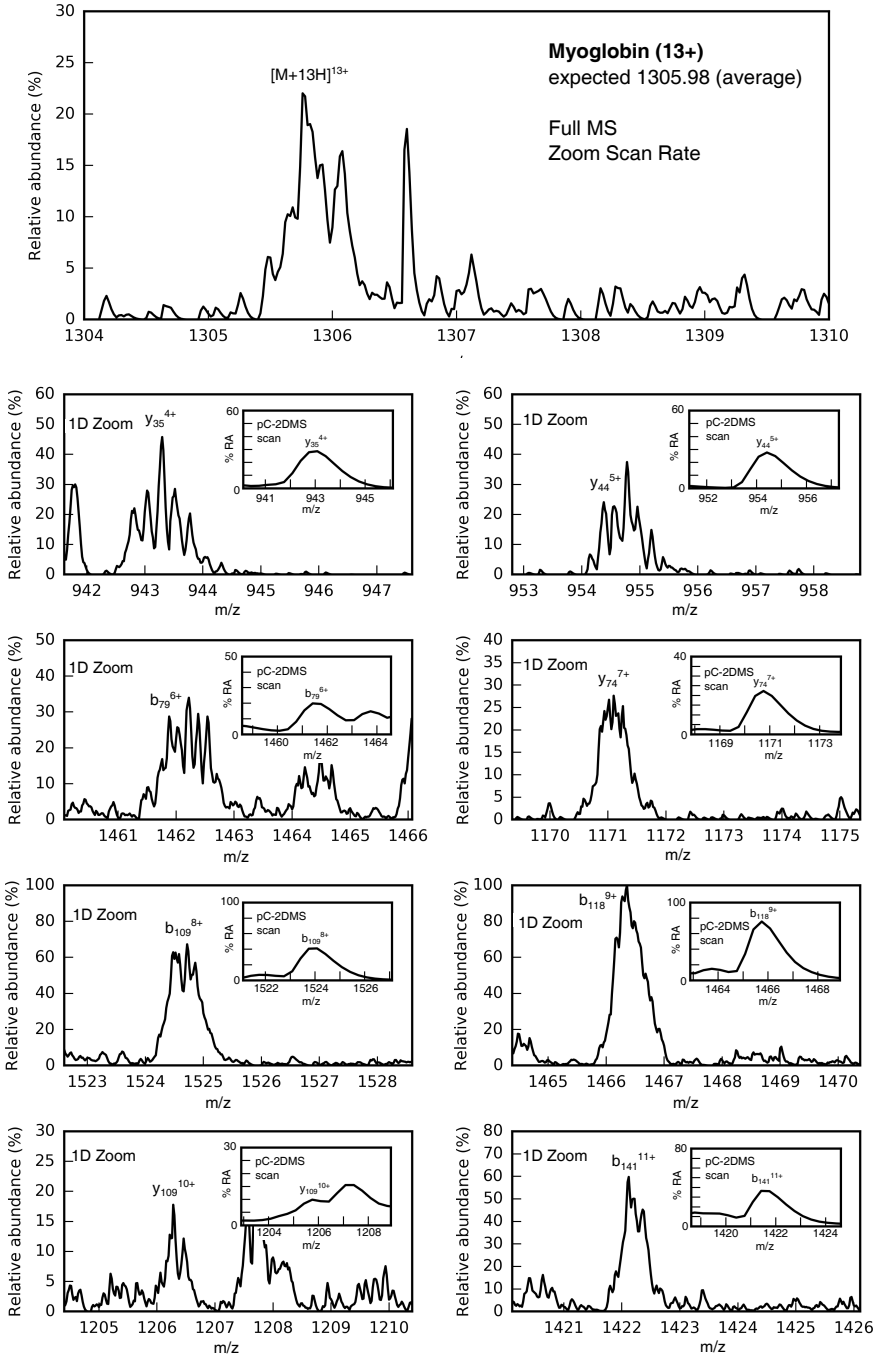


Figure 14.3: Isotopic envelopes measured on the LTQ XL at the highest possible resolution demonstrate the inability of the linear ion trap mass analyser to identify the charge state of ions with charge state greater than 5.

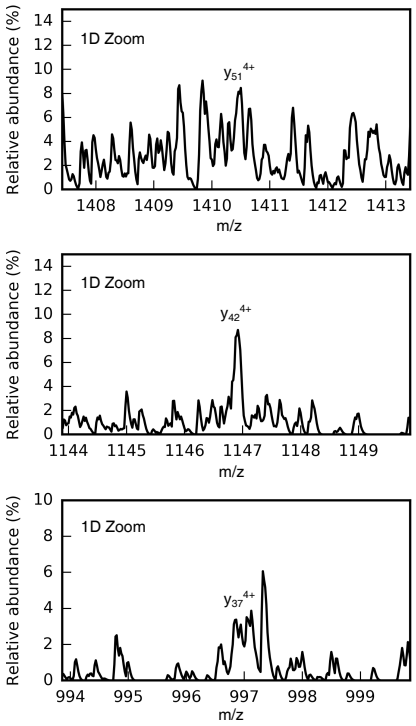
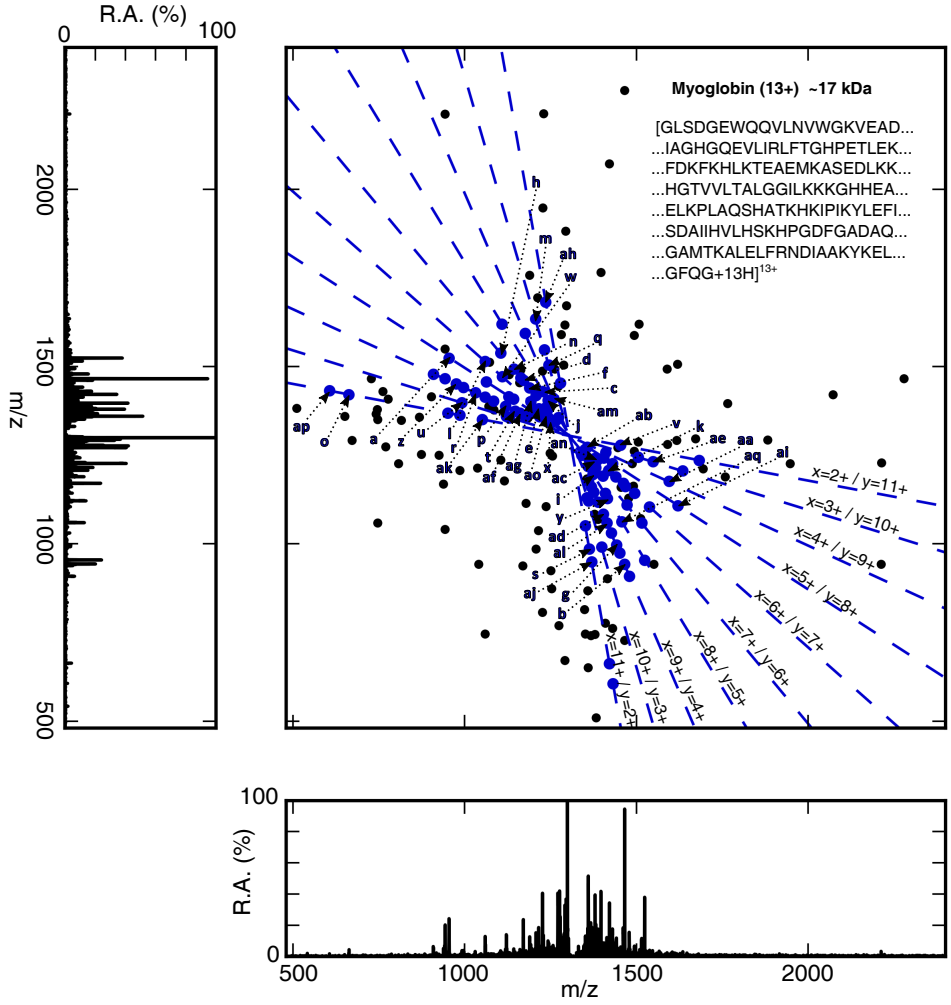


Figure 14.4: Inability of linear ion trap mass analyser to resolve 4+ fragment charge states for lower abundance fragment ions of myoglobin (13+) in highest resolution 'Zoom Scan' mode.

14.3, the high scan rate scans which are taken for the pC-2DMS measurements afford no resolution whatsoever of the fragment ion isotopic envelope. This is problematic for standard 1D MS, which relies on the use of the monoisotopic mass of fragment ions in order to match them. The use of the monoisotopic mass stems from the requirement for very high mass accuracy, necessary for 1D MS in order to reduce the number of deleterious false positive fragment matches. Provided sufficiently high mass resolution, the accurate determination of the m/z ratio of a single monoisotopic peak is more straightforward and repeatable than the accurate measurement of the average mass from the monoisotopic envelope. However, as demonstrated in Chapter 10, the two-dimensional nature of pC-2DMS hugely diminishes the requirement for high mass accuracy in order to reduce false positive ion matches and instead shifts the emphasis onto increased scan rate, in order to be able to require a sufficient number of scans to perform pC-2DMS analysis in a suitably short period of time. The overwhelming

benefits of sacrificing mass accuracy in favour of a larger number of scans to generate pC-2DMS correlations are demonstrated in Fig. 10.1 on page 125. However, the structure of the pC-2DMS peaks (which, for any covariance or partial covariance map, always matches the structure of the correlated 1D signals) does not provide the monoisotopic peak and therefore does not permit the monoisotopic mass to be read off.

pC-2DMS top-down measurements Figure 14.5 shows the successful application of the restricted Hough transform to identify the primary complementary ions, and their charge states, resulting from the fragmentation under CID of the 13+ ion of the intact protein myoglobin from equine skeletal muscle (153 amino acid residues long). The 100 highest-ranking pC-2DMS correlation signals, calculated from the measurement of myoglobin (13+) under CID at NCE 70%, have been used. These features were passed to the restricted Hough transform described above, with a threshold value of 6 points required to be lying along a line to identify it as a mass conservation line. The Hough transform has identified 5 different mass conservation lines, each of which correspond to a different charge partition for the complementary ions from a parent ion of mass 16958.0 and charge 13+. This parent mass, determined directly from the pC-2DMS map and the restricted Hough transform, shows a deviation of 0.04% from the expected theoretical mass, well within the 0.1% parent mass tolerance typically used for top down database searches [180, 181]). All correlations determined to be complementary correlations are plotted in blue, and the 43 annotated signals have all been identified by our automatic pC-2DMS signal matching software as being correlations between complementary b_a^{i+} & y_b^{j+} ions where $a+b=153$ and $i+j=13$. The alphabetical order of the annotated correlations reflects their ranking according to the pC-2DMS correlation score, and the labels are provided in the table beneath the map. The fragment ion matching was performed with a fragment ion tolerance of 0.8 Da.



Label	b/y complementary ion correlation	Label	b/y complementary ion correlation
a	b_{109}^{8+} & y_{44}^{5+}	**w**	b_{93}^{7+} & y_{60}^{6+}
b	b_{118}^{9+} & y_{35}^{4+}	**x**	b_{73}^{6+} & y_{80}^{7+}
c	b_{63}^{5+} & y_{90}^{8+}	**y**	b_{102}^{8+} & y_{51}^{5+}
d	b_{79}^{6+} & y_{74}^{7+}	**z**	b_{117}^{9+} & y_{36}^{4+}
e	b_{87}^{7+} & y_{66}^{6+}	**aa**	b_{56}^{4+} & y_{97}^{9+}
f	b_{78}^{6+} & y_{75}^{7+}	**ab**	b_{60}^{5+} & y_{93}^{8+}
g	b_{116}^{9+} & y_{37}^{4+}	**ac**	b_{98}^{8+} & y_{55}^{5+}
h	b_{83}^{6+} & y_{70}^{7+}	**ad**	b_{113}^{9+} & y_{40}^{4+}
i	b_{99}^{8+} & y_{54}^{5+}	**ae**	b_{42}^{3+} & y_{111}^{10+}
j	b_{62}^{5+} & y_{91}^{8+}	**af**	b_{42}^{4+} & y_{111}^{9+}
k	b_{77}^{6+} & y_{76}^{7+}	**ag**	b_{63}^{6+} & y_{90}^{7+}
l	b_{96}^{7+} & y_{57}^{6+}	**ah**	b_{122}^{11+} & y_{31}^{2+}
m	b_{44}^{3+} & y_{109}^{10+}	**ai**	b_{73}^{5+} & y_{80}^{8+}
n	b_{81}^{6+} & y_{72}^{7+}	**aj**	b_{17}^{2+} & y_{136}^{11+}
o	b_{141}^{11+} & y_{12}^{2+}	**ak**	b_{135}^{11+} & y_{18}^{2+}
p	b_{112}^{9+} & y_{41}^{4+}	**al**	b_{114}^{9+} & y_{39}^{4+}
q	b_{41}^{3+} & y_{112}^{10+}	**am**	b_{102}^{9+} & y_{51}^{4+}
r	b_{115}^{9+} & y_{38}^{4+}	**an**	b_{75}^{6+} & y_{78}^{7+}
s	b_{136}^{11+} & y_{17}^{2+}	**ao**	b_{86}^{7+} & y_{67}^{6+}
t	b_{111}^{9+} & y_{42}^{4+}	**ap**	b_{142}^{11+} & y_{11}^{2+}
u	b_{126}^{10+} & y_{27}^{3+}	**aq**	b_{47}^{5+} & y_{106}^{8+}
v	b_{27}^{2+} & y_{126}^{11+}		

Figure 14.5: pC-2DMS map of 13+ ion of intact protein myoglobin from equine skeletal muscle (∼17 kDa). The blue dashed lines are the mass conservation lines identified by the restricted Hough transform, showing different charge partitions across the complementary ion correlations which fall along them (blue dots). Identities of the annotated complementary ion correlation signals, which are each labelled on only one side of the autocorrelation line for visual clarity, are given in the table.

Although the routine developed for determination of the monoisotopic m/z for peptide and oligonucleotide fragments provided mass accuracy well within the normal specified mass accuracy of $+/-0.8$ Da for a linear ion trap mass analyser (see e.g. Chapter 11), Fig. 14.6 demonstrates the systematic inaccuracy accumulated by the determined monoisotopic masses for fragment ions at higher m/z.

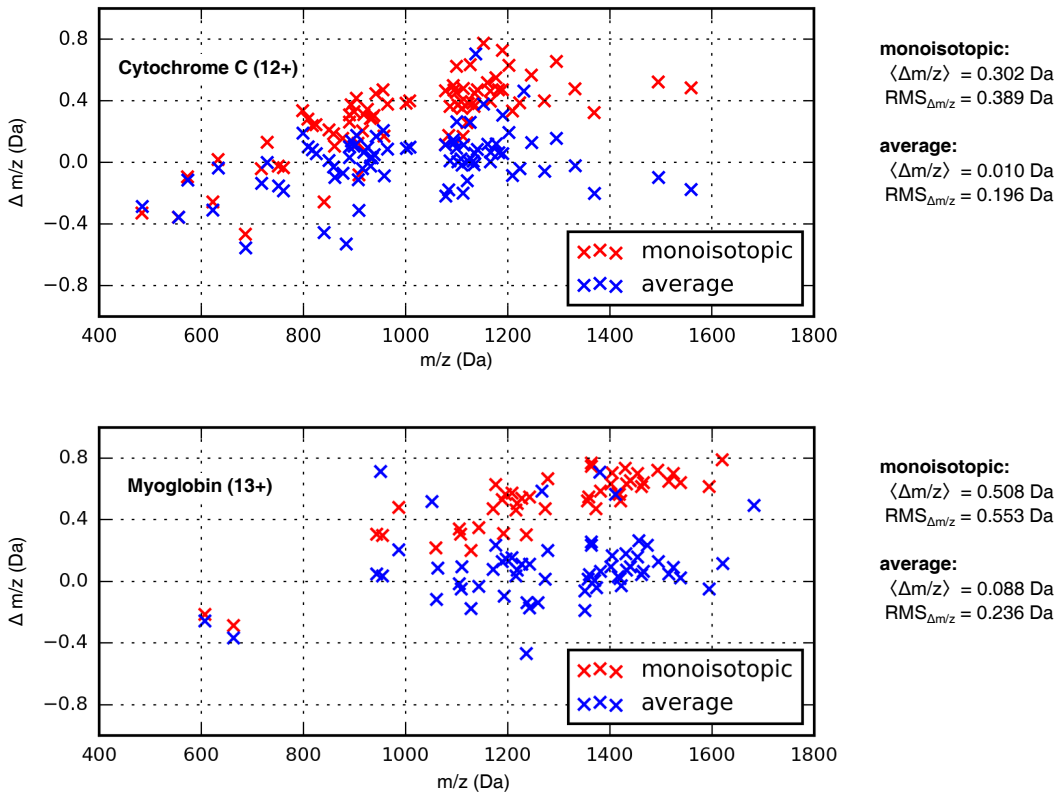


Figure 14.6: Deviations in m/z for matched pC-2DMS correlations in top down measurements, monoisotopic mass (red) *vs.* average mass (blue).

For fragment ions in correlations which are able to be automatically assigned by the pC-2DMS search engine software (and so fall within the specified m/z tolerance of ± 0.8 Da), the fragment ion m/z is plotted on the x-axis whilst the deviation of that value from the expected (theoretical) m/z for that value is plotted on the y-axis. The monoisotopic m/z's as determined by the standard procedure described above show a steadily increasing deviation with increasing m/z for both the parent ions plotted.

Subtraction of a term linearly dependent on the fragment ion m/z is a possibility, but there exists a simple and elegant solution which should be robust enough to account for an arbitrary increase in mass for fragment ions. This solution is motivated by the information available from the unresolved peak provided by the high scan rate measurements taken for pC-2DMS. Investigation of the third moment (skew) of this envelope function for features on the pC-2DMS map did not show any significant differences between fragment ions of peptide and those of full protein molecules, currently precluding the possibility of being able to accurately infer the monoisotopic peak from the centre of mass of the peak and the value of the skew (in theory, the smaller the mass of the ion, the greater the abundance of its lower mass isotopes and therefore the greater the value of its [positive] skew). However, the centre-of-mass of the measured peaks should directly correspond to the average mass of the measured ions. Indeed, by taking the centre-of-mass of all pC-2DMS peaks and matching these values with the m/z values corresponding to theoretical average masses, it is possible to achieve consistently high m/z accuracy which remains robust across the entirety of the mass spectrum (blue crosses is Fig. 14.6). As well as the identification of 43 complementary ion correlations in Fig. 14.5, it is demonstrated below that use of these average masses, at the standard m/z tolerance of ± 0.8 Da, enables the correct identification of two different, co-fragmented, intact protein molecules from their chimera spectrum (Fig. 14.7).

14.2 *In silico* deconvolution of mixtures of intact proteins by pC-2DMS

As well as charge state determination and identification of signal origin, mass conservation lines and the Hough transform can also be used to deconvolve the spectra of mixed biomolecules. Biological samples are typically complex mixtures of more than one protein, and separation of these mixtures prior to top down 1D MS/MS analysis is essential to avoid the insurmountably difficult task of identifying proteins

from the overlapping 1D fragment ion signals from the simultaneous decomposition of several protein molecules. Liquid chromatography (LC) is the preferred method for separation of complex mixtures prior to top-down analysis because it is straightforwardly automated and is able to couple directly to a mass spectrometer for online analysis. Reversed-phase liquid chromatography (so-called because, in contrast to the first LC systems, the stationary phase is hydrophobic) in particular has experienced a remarkable amount of success in the separation of mixtures of small (up to ~ 3 kDa) peptide molecules [4]. However, different realisations of liquid chromatography present various different problems for top-down analysis, and when it comes to intact proteins the technique suffers from limited separation capabilities and poor, to undetectably low, protein recovery [182].

pC-2DMS allows for the unprecedented *in silico* separation of protein mixtures which have been co-isolated and co-fragmented, precluding the costly, wasteful and throughput-limiting process of upstream separation. According to Eq. 13.2, the complementary ions produced by the fragmentation of parent molecules of different mass and/or charge state fall along uniquely defined mass conservation lines. Therefore, all that is required for the separation of overlapping fragment ions in the pC-2DMS map is the identification of the different mass conservation lines appearing on the map, which, as demonstrated above, is straightforwardly performed by the Hough transform.

It is useful that the type of fragment ion correlation which can be confidently identified and separated on the pC-2DMS map (complementary b/y ions) is especially suitable for protein identification in mass spectrometry [183]. This is particularly true in the case of top-down mass spectrometry, due to the larger number of amino acid residues in the sequences which are fragmented. As well as complementary b- and y-type ions, protonated peptides and proteins fragmenting under CID can also produce internal ions due to more than one cleavage along the peptide backbone. Whilst such fragment ions can be extremely useful (see, for example, the identification and localisation of combinatorial PTMs in Chapter 9), they can also be problematic in e.g. database searches for top-down mass spectrometry, where they cause the

probability for incorrect assignment of experimental spectral peaks from an unknown sequence to explode. This is because the number of possible internal subsequences that it is possible to produce from a sequence of length n grows with the square of the sequence length[5], and the higher charge states encountered in top-down mass spectrometry further increases the number of different internal ions it is possible to produce from the fragmentation of one parent ion. In such cases, identification of complementary fragment ions from a fragment ion spectrum is particularly valuable for protein identification from the fragment mass spectrum.

Figure 14.7 demonstrates the *in silico* separation of the two co-isolated and co-fragmented intact protein ions cytochrome C (13+) and ubiquitin (9+). The top 200 pC-2DMS correlation score-ranked features, along with the roughly determined parent ion m/z values as measured in the precursor scan in the linear ion trap, have been passed to the restricted Hough transform with the same parameters as used for the identification of the mass conservation lines for pure spectrum of myoglobin in 14.5. The Hough transform has identified two sets of mass conservation lines, corresponding to parent ions of average mass 8572.7 Da[6] and charge state 9+ (blue) and average mass 12368.4 Da[7] and charge state 13+ (red). The zoomed-in view of the horizontal 1D MS/MS spectrum illustrates the deconvolution and charge state identification performance of the Hough transform in this particularly congested region of the spectrum. The two different sets of correlation features lying along each of these two sets of mass conservation lines have been individually passed to the pC-2DMS search engine (described in Chapter 11) along with the parent mass and charge state determined by the Hough transform and a typical parent tolerance of 0.1% [180, 181]. All searches were performed against the UniProtKB/Swiss-Prot database [152] under a non-specific digest (which includes intact protein sequences) with N-terminal acetylation and heme C specified as variable modifications, and initiator methionine loss automatically considered as a variable modification by virtue of the non-specific

[5]Assuming a minimum length of 2 as for standard b-type internal ions, this number is $(n-2)(n-3)/2$.

[6]This value, determined by the Hough transform directly from the fragment ion spectrum, represents a deviation of only 0.01% from the theoretical average mass of the ion.

[7]Deviation of 0.04% from theoretical average mass of the ion.

digest. The search was performed using the average masses of the fragment ions, as justified above for pC-2DMS, and at a fragment ion matching tolerance of ± 0.8 Da as for the peptide searches shown in Chapter 11. As illustrated by the inset histograms in Fig. 14.7, the pC-2DMS search engine unambiguously identifies each of the two mixed proteins from the two sets of deconvolved pC-2DMS features.

Both sets of mass conservation lines in Fig. 14.7 cross the line $x = y$ at the same point. This is a result of the fact that both co-fragmented parent ions have very similar m/z values, which is a restriction on the ionic species it is possible to co-fragment using the basic resonant excitation CID capabilities of the linear ion trap instrument used for the experiments performed in this thesis [55]. This is by no means a restriction on the application of the pC-2DMS deconvolution of multiplexed mass spectra, which, as for pC-2DMS measurements themselves, requires only the sufficiently high instrumental detection efficiency [42] which has been demonstrated in this thesis. The technique described above is, for example, expected to be immediately applicable for experimental setups where the fragmentation of different ions across a large m/z range is possible, such as methods of photodissociation [34], electron-based dissociation [32, 33] and sustained off-resonance irradiation CID (SORI-CID) [184, 185]. The application of different resonant excitation voltages to a linear quadrupole, which has been experimentally demonstrated [165] but hitherto enjoyed limited relevance, is also a possibility.

This chapter has described the exploitation of a number of unique features of pC-2DMS (complementary sequence ion determination, 2D fragment and parent charge state identification, parent mass identification) to the successful top-down analysis of intact proteins using a linear ion trap mass analyser. Furthermore, the use of mass conservation lines to assign fragment ions to the parent which produced them opens the possibility for pC-2DMS multiplex measurements on peptides, oligonucleotides (see Chapter 17) and even, as experimentally demonstrated in Fig. 14.7, intact protein ions.

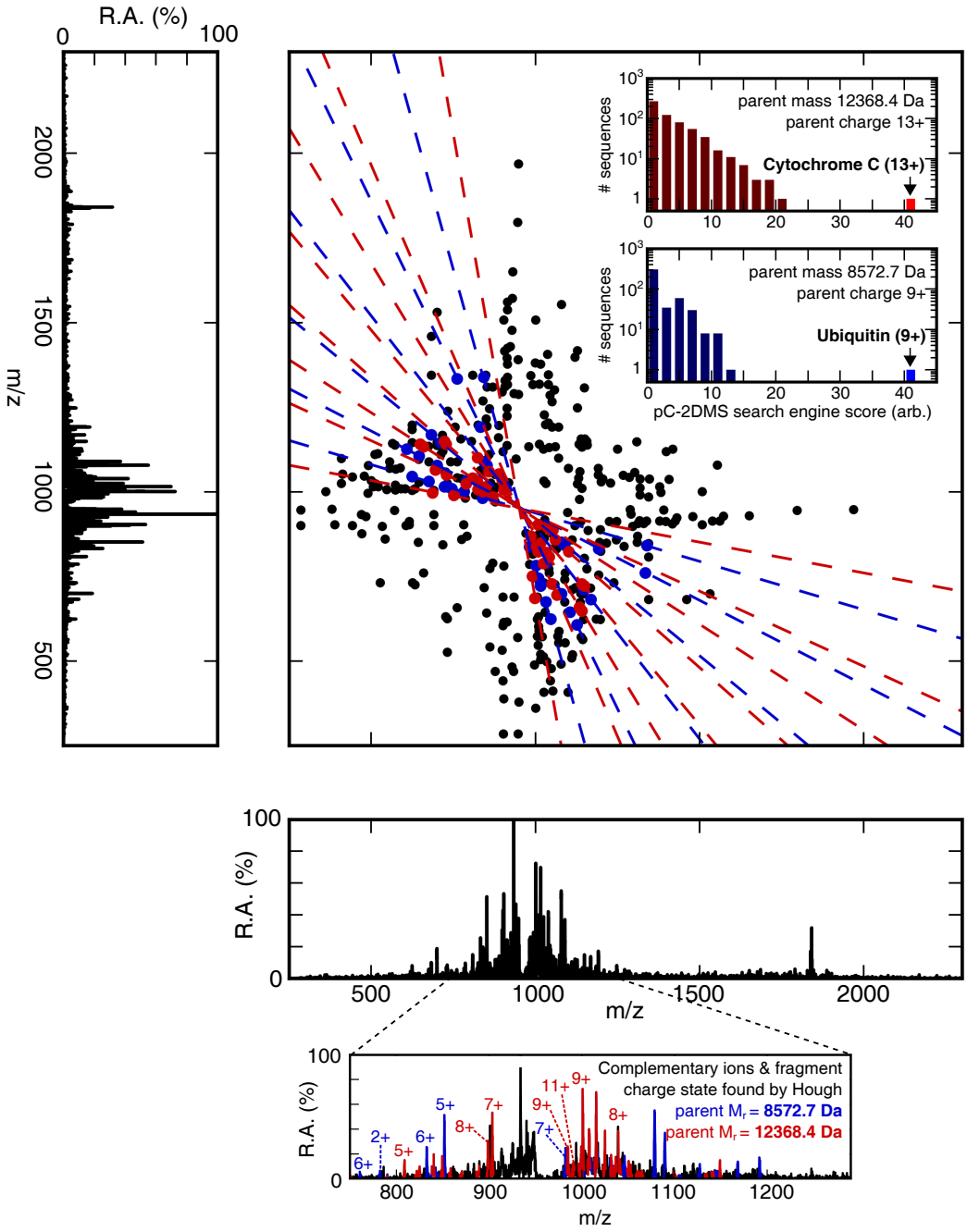


Figure 14.7: Deconvolution of the fragment spectrum of co-fragmented protein parent ions cytochrome C (13+) and ubiquitin (9+) using pC-2DMS and the restricted Hough transform. The restricted Hough transform has identified two different sets of mass conservation lines (red and blue), along which lie the complementary fragment ion correlations of two different intact protein ions. Both protein ions have been successfully identified by the pC-2DMS search engine using the fragment ion correlations, parent mass and parent charge state, all of which have been determined by the restricted Hough transform.

Chapter 15

Two-dimensional diagnostic for fragment spectra from multiple parent ions

A major challenge for the successful identification of peptide sequences by tandem mass spectrometry is the common scenario of two or more parent ions being co-selected for fragmentation simultaneously [161]. This is estimated to occur for up to ∼50% of all fragment ion spectra in a standard high-performance liquid chromatography MS/MS (HPLC-MS/MS) run [161, 172], and results in so-called chimera spectra, with complex patterns of overlapping peaks from the fragment ions of more than one parent ion. It was shown in Chapter 14 that using mass conservation lines and a restricted version of the Hough transform enables the identification and deconvolution of the chimera spectrum of two co-fragmented parent ions, provided the parent ions either have different charge states or a resolvable difference in mass. This general technique should be immediately applicable to a large number of different scenarios where two or more different parent ions have been unintentionally or intentionally co-fragmented. This chapter extends the use of mass conservation lines further, for the identification of chimera spectra for CID measurements on peptide cations of the same charge state and with no measurable difference in mass, up to fully isomeric co-fragmented parent ions.

In 1D MS/MS, the identification of chimera spectra relies on a high quality, high mass resolution scan of the precursor ions prior to fragmentation [172, 183, 186–189]. This requires dedicated high-resolution mass measurement hardware, is dependent on the unreliable quality of a precursor scan, and is fundamentally limited by the requirement of a measurable m/z-separation between the multiple co-isolated ions, precluding the identification of chimera spectra for co-isolated isomeric or near-isomeric parent ions. Furthermore, during a HPLC-MS/MS run, the temporal separation between the precursor scan and related MS/MS scans means it is likely that the precursor scan is not fully representative of the ion ensemble fragmented for the MS/MS measurement (a result of the precursor scan and MS/MS scans taking place at different points in time during the time-resolved chromatographic separation). Even techniques proposed but not yet realised which would in theory not require a high-resolution precursor scan, would still rely on an instrumentally resolvable mass difference between different co-isolated parent ions [183].

This chapter presents a new 2D method for the identification of peptide chimera spectra for CID experiments on protonated peptide ions directly from the fragment ion spectrum. This technique precludes the requirement for a high quality, high-mass resolution precursor scan and is able to diagnose chimera spectra even in the case of different fully isomeric co-isolated parent ions (no potentially measurable mass difference between them). Numerical simulations based on experimental measurements have been performed to demonstrate its sensitivity, the experimental demonstration of the detection of a mixture of two isomeric fully parent ions succeeds even when one ion is at one five-hundredth of the molar concentration of the other.

The concept of the 2D chimera diagnostic is general across all charge states, thanks to the ability to determine fragment ion charge state for correlations falling along the mass conservation line (see Chapter 13). As illustrated in e.g. Fig. 13.1 on page 151, all complementary ion pairs (b/y in CID measurements of protonated peptides [2]) arrange themselves along so-called 'mass conservation lines' on a pC-2DMS map, named because the position of the two complementary ion m/z's along this line is a manifestation of the conservation of mass of the parent ion by the combined

mass of the two fragments. For a parent ion of mass Mr_{Par} and charge z_{Par} which decomposes into two detected complementary fragment ions of m/z ratios m_1/z_1 and m_2/z_2 (measured on the x- and y-axes respectively), the complete set of mass conservation lines along which all the complementary ions fall is defined by:

$$y = -\frac{z_1}{z_2} \times x + \frac{Mr_{Par}}{z_2}, \tag{13.2}$$

replicated from page 150, where z_1 and z_2 are constrained by $z_2 \geq z_1$ and $z_1 + z_2 = z_Par$ (see Chapter 13 for more details). It follows that with knowledge of the charge state and m/z of a parent ion it is possible to define the mass conservation lines for its products *a priori*, which enables immediate determination of the intact complementary ions from the pC-2DMS map. Knowledge of the parent ion is not necessary however; the mass conservation lines can also be identified using computer vision (see section 14.1).

Because isolation is performed on m/z, co-isolated parent ions will have the same or very similar m/z values. If they have a different charge state, Eq. 13.2 dictates that the fragment ions will fall on different mass conservation lines, enabling immediate identification of the chimera spectrum and unambiguous assignment of the fragment ions to the correct parent ion, as performed for intact protein ions of different charge states and similar m/z in section 14.2. If the co-isolated parent ions have the same charge state, but the m/z tolerance of the fragment ion detector is sufficient, the complementary ions from the different co-isolated parents will be immediately separated by virtue of falling on different mass conservation lines. If, however, the charge state of the co-fragmented parent ions is the same and the resolution of the fragment ion detector is not sufficiently high (this includes the case where the co-isolated parents are isomeric), the complementary ions from all co-isolated and co-fragmented parent ions will fall along the same mass conservation line(s). Nonetheless, even in this situation, the identification of chimera spectra resulting in the co-isolation and co-fragmentation of multiple parent ions is straightforwardly possible.

Under CID, which is by far the most preferred ion activation method for tandem

MS [95], the intact complementary ion pairs produced are of b-type (N-terminus) and y-type (C-terminus), corresponding to cleavage of the HN-CO peptide bond along the peptide backbone [2]. The experimental realisation of CID, the fragmentation mechanism and the resultant fragment products are discussed in more detail in section 5.3. Each successive b-ion or y-ion from the backbone cleavage of a particular sequence is separated in mass by the mass of the next amino acid. The least massive amino acid is glycine (G) where the functional R-group is a single H atom and the mass is therefore 57 Da. The 2D chimera diagnostic relies on the following consideration: the presence of three or more intact complementary ions of mass within less than 57 of each other is equivalent to saying there are (at least) two consecutive b-ions or y-ions in the fragment spectrum which are within less than 57 Da mass of each other. This means that not all complementary ions can originate from the same sequence, because two successive complementary ions must be separated by 57 Da or more. As such, if three or more intact complementary ions of mass within less than 57 of each other are identified, the spectrum contains the fragment ions of more than one sequence and can therefore be identified as a chimera spectrum. A chimera 'flag' is defined as a single set of 3 such complementary b- and y-ions of mass within 57 Da of each other, and it is of course possible (and expected) for a chimera spectrum to contain multiple such flags.

Numerical simulations were performed to test the 2D chimera diagnostic and estimate its recall rate. To do so, the pC-2DMS detection probability for b/y complementary ion pairs was first estimated by surveying a representative range of scans on peptide ions (ten doubly charged ions and nine triply charged ions), using the matching procedure from the pC-2DMS search engine described in Chapter 11. From this small sample, 69.07% of all possibly detectable b/y complementary pairs were detected from 2+ parent ions and 74.13% from 3+ parent ions[1]. The simulations were carried out separately for each of these two parent charge states.

[1]'Detection' was defined as appearing in the top 50 pC-2DMS correlation-ranked signals, the standard procedure for selection of true correlation signals as described in section 6.2. The successful use of the pC-2DMS correlation score as an absolute measure, as discussed on page 83, should enable more sensitive detection of true correlation signals in a chimera spectrum where a greater high-scoring pC-2DMS correlation features will be produced.

Next, the UniProt/Swiss-Prot database [152] was downloaded and, ignoring the 0.47% of protein sequences with ambiguous amino acid residue coding, all sequences were subjected to an *in silico* tryptic digest (no missed cleavages). From the resultant digestion products, five different sets of 5,000 peptide sequences of length between 5 and 15 residues were selected at random. For each of these randomly selected sequences, all sequences which give, at the relevant charge, a parent ion which falls within a typical ion trap isolation width of 2.5 Da of the given sequence were identified, and from each of these pools sets of possible two-fold, three-fold and four-fold mixtures of peptides which therefore risk being co-isolated and co-fragmented were drawn, without replacement, at random.

Each set of potentially co-isolated sequences was co-fragmented *in silico*, and the resultant b-ion/y-ion complementary correlations were then 'measured' *in silico* at the experimentally determined detection probability for the relevant parent ion charge. To account for experimental inaccuracies, the *in silico* measurement incorporated a varying deviation from the theoretical m/z which was uniformly distributed across $\pm acc_{m/z}$, with $acc_{m/z}$ the specified instrumental m/z accuracy (here, 0.8 Da as standard for the linear ion trap pC-2DMS measurements). For 3+ parents, which generate a 2+/1+ complementary pair, the 2+ fragment was specified to be the longest (in terms of number of residues) of the two, and if the fragments had equal numbers of residues the extra charge was attributed to the y-ion[2]. The resultant b/y complementary pairs measured for each two-, three- or four-fold chimera spectrum was then queried to identify all 'chimera flags', i.e. all sets of of three or more consecutive complementary ions falling within $57 - 2 \times acc_{m/z}$ of each other for 2+ parent ions, or $57 - 3 \times acc_{m/z}$ for 3+ parent ions. The subtraction of the factors $2 \times acc_{m/z}$ or $3 \times acc_{m/z}$ account for the finite accuracy of the detector[3].

For triply charged ions, using the experimentally determined detection probability

[2]This is justified by the fact that for tryptically cleaved peptides, the proton tends to be sequestered by the highly basic C-terminal lysine (K) or arginine (R) residue.

[3]In an unlikely worst case scenario, two singly charged ions with true m/z's x and y ($x < y$) could be measured at m/z's $x + acc_{m/z}$ and $y - acc_{m/z}$ such that their measured difference in mass is $(y - acc_{m/z}) - (x + acc_{m/z}) = y - x - 2 \times acc_{m/z}$. For 3+ parents, an extra multiple of $acc_{m/z}$ is accumulated as the m/z of the doubly charged fragment ion is propagated to a mass.

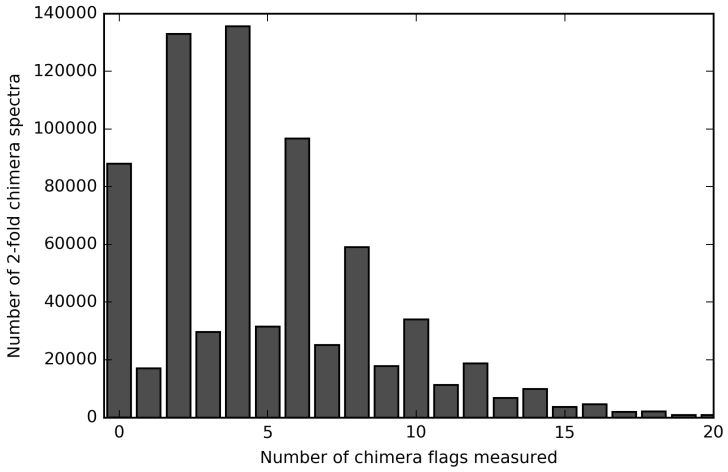


Figure 15.1: The number of chimera flags expected to be measured, according to typical pC-2DMS experimental parameters, for two-fold chimera spectra of 3+ parent ions.

of complementary b/y pairs, numerical simulations gave the the 2D chimera diagnostic an estimated recall rate of 87.94% ($\sigma = 0.24\%$ across the five repeated iterations using different sets of 5000 randomly selected peptides) for two-fold chimera spectra, 99.33% ($\sigma = 0.036\%$) for three-fold chimera spectra and 99.97% ($\sigma = 0.0022\%$) for four-fold chimera spectra. For doubly charged parent ions, the 2D diagnostic was given an estimated recall rate of 82.68% ($\sigma = 0.32\%$) for two-fold chimera spectra, 98.50% ($\sigma = 0.072\%$) for three-fold chimera spectra and 99.90% ($\sigma = 0.015\%$) for four-fold chimera spectra. The A histogram of the number of chimera flags expected to be produced at the experimentally determined b/y correlation measurement probability, for two-fold chimera spectra of triply charged peptides, is shown in Fig. 15.1. In all, there were 7.29×10^5 simulated chimera spectra tested to create this histogram, for which the mean average number of flags produced was found to be 4.99. The strong modulation in favour of an even number of detected flags is explained by the proximity in mass between two co-isolated and co-fragmented parent ions (isolation width 2.5 Da). All of the b-(y-)ions contributing to a given chimera flag will have detected complementary y-(b-)ions, whose mass is the difference between the parent molecule and its complementary b-(y-)ion. Given the parent masses of all co-fragmented ions are constrained to be very similar, in the majority of cases the three detected b- or

Analysed sample	Chimera spectrum identified?	# chimera flags
[GSNKGAIIGLM+2H]$^{2+}$ (**I1**, pure)	N	0
[MLGIIAGKNSG+2H]$^{2+}$ (**I2**, pure)	N	0
[**I1**+2H]$^{2+}$:[**I2**+2H]$^{2+}$ Rel. molar conc. 50:50	Y	14
[**I1**+2H]$^{2+}$:[**I2**+2H]$^{2+}$ Rel. molar conc. 10:90	Y	10
[**I1**+2H]$^{2+}$:[**I2**+2H]$^{2+}$ Rel. molar conc. 5:95	Y	10
[**I1**+2H]$^{2+}$:[**I2**+2H]$^{2+}$ Rel. molar conc. 1:99	Y	6
[**I1**+2H]$^{2+}$:[**I2**+2H]$^{2+}$ Rel. molar conc. 1:499	Y	2

Table 15.1: Table summarising 2D chimera spectrum identification for the doubly charged ions of the two isomeric sequences GSNKGAIIGLM (**I1**) and MLGIIAGKNSG (**I2**) across different relative molar concentrations.

y-ions within 57 Da mass of each other which contribute to a given chimera flag will have three respective complementary ions which also fall within 57 Da of each other, producing their own chimera flag. For both charge states, as expected, the simulated complementary ion spectra for 50,000 tested pure peptides (for which the 'worst-case scenario' of 100% detection probability for all possible complementary b-ion/y-ion pairs was assumed) produced zero chimera flags.

The experimental performance of the 2D chimera diagnostic was also investigated, using the two isomeric sequences GSNKGAIIGLM (isomer 1, **I1**) and MLGIIAGKNSG (isomer 2, **I2**). Identification of such a chimera spectrum lies outside the realms of possibility for 1D mass spectrometry, which requires a measurable difference in precursor m/z's [186–189], even in the ideal case of infinite instrumental resolution. The two sequences were mixed at different relative molar concentrations and the pC-2DMS spectrum was measured for their doubly charged ions. The top 50 pC-2DMS correlation-scored sequences were queried for chimera flags. Table 15.1 demonstrates the remarkable sensitivity of the 2D chimera diagnostic: its performance remains robust even when the **I1** is at a relative molar concentration of *only one five-hundredth* of **I2**. The pC-2DMS map of the co-isolated ions [**I1**+2H]$^{2+}$ and [**I2**+2H]$^{2+}$ at this relative molar concentration is shown in Fig. 15.2. The top 50 pC-2DMS correlation

score-ranked features are plotted, and the this score is coded in the colour map. Use of the pC-2DMS correlation score, which is able to identify complementary ions even at extremely low abundances (see also Fig. 6.4 on page 80), allows for the detection of two chimera flags even at extremely low abundance of one of the precursor ions.

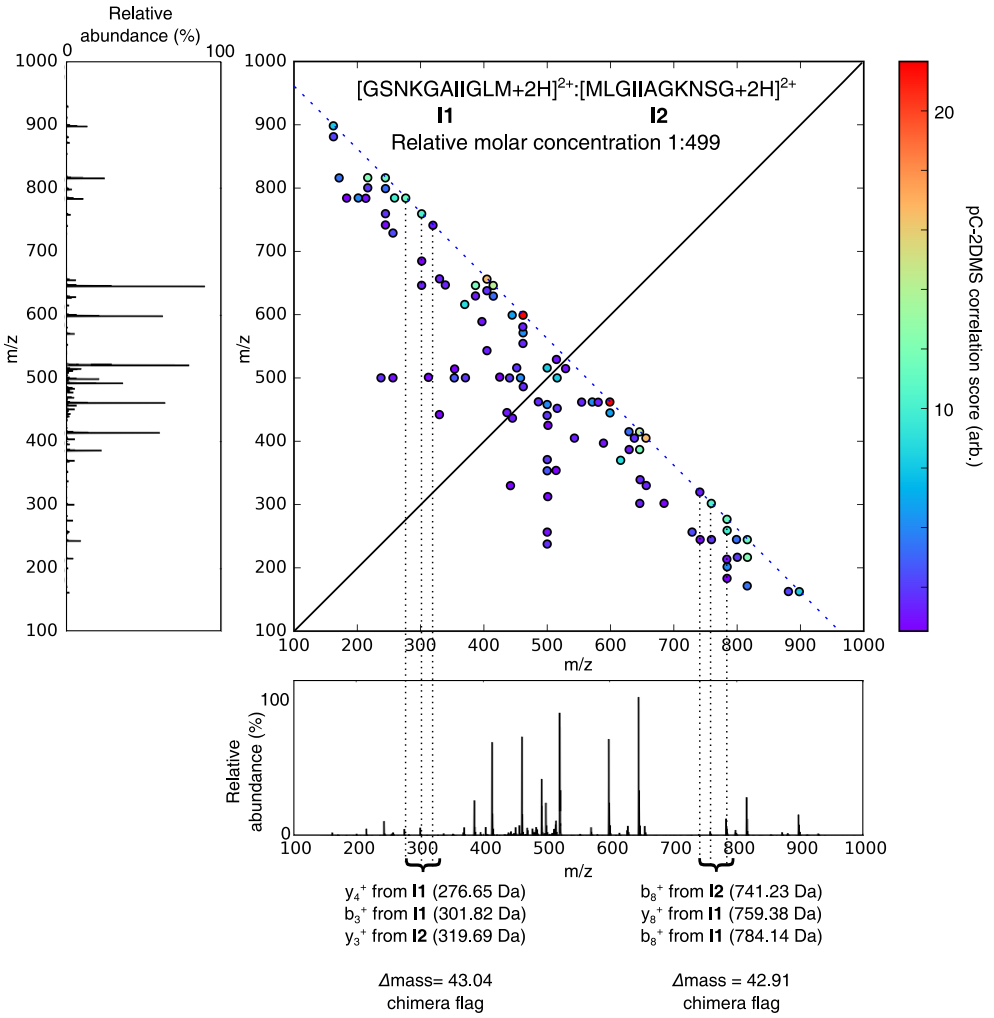


Figure 15.2: pC-2DMS map for the chimera spectrum of $[\text{GSNKGAIIGLM+2H}]^{2+}$ ($\mathbf{I1}^{2+}$) and $[\text{MLGIIAGKNSG+2H}]^{2+}$ ($\mathbf{I2}^{2+}$) at a relative molar concentration of 1:499, showing the two measured chimera flags.

The sensitivity of the 2D chimera diagnostic far surpasses the detection level typically required to identify the most problematic chimera spectra, from which no peptide identifications are possible. This scenario is estimated to occur for 1D MS/MS when the fragment ions of the most abundant peptide in a fragment ion spectrum peptide

are contaminated by the fragments ions from another at a relative abundance of over 20% [161]. The use of the 2D chimera diagnostic in the case of such high abundance contaminants would most likely involve the subjection of the chimera spectra to analysis which expects more than one peptide ion to have contributed to a single fragment ion spectrum. There have been a number of algorithms developed for this purpose for standard 1D MS/MS [186], and the high specificity of the pC-2DMS search engine is also a promising avenue for the identification of more than one peptide sequence from the same fragment spectrum. For example, for the case of the chimera spectrum in Fig. 11.4 (page 139) where a naturally occurring peptide sequence has been intentionally co-fragmented with an unnatural contaminant which is therefore not tested in a database search, under fragment correlation matching the unnatural contaminant receives a slightly higher score than the correctly identified natural sequence, therefore providing two stand-out identifications from one set of measured fragment ion correlations. In the case of identification of low-abundance contaminant chimera spectra (featuring, of course, the implied measurement of the low abundance contaminant fragment ions), the 2D chimera diagnostic opens the door for identifying many more chimera spectra than previously detectable from which more than one peptide sequence could be identified. This opens the door for identifying peptide sequences which have until now been buried by more abundant, co-isolated species, offering greatly improved coverage of the proteome.

Chapter 16

Extension to peptide anions

The vast majority of proteomic MS measurements are performed in the positive ion mode, on protonated cationic peptide molecules. The prevalence of positive ion mode mass spectrometry can be explained by a number of critical advantages which it enjoys: it combines well with a number of mature LC-MS technologies using acidic buffer solutions, measurements of cations are typically more sensitive than anions [191], and through something of a virtuous spiral the body of experimental data (and therefore level of understanding) for tandem MS on positive peptide ions [2, 25] far outweighs the work so far done on negative ions [192]. Furthermore, under the most common ion activation method of collision-induced dissociation [16], deprotonated peptide anions undergo a number of side chain fragmentations [192] (i.e. fragmentations of the functional R^n group of the peptide's amino acids, see Fig. 5.8) which are minimally structurally informative and can complicate the fragment ion spectrum.

The result of this is that those sequences which are successfully analysed are typically basic molecules, which are preferentially protonated in solution prior to ESI in the standard MS workflow. This disfavours the significant portion of the proteome[1] which is acidic, either by virtue of a large number of acidic residues or the presence of biologically important PTMs such as phosphorylation or sulphation. Acidic peptide molecules favour deprotonation in solution, and as such many parts of

[1]The proteome is the entire set of proteins which can be expressed by an organism, or as it is used here, any organism.

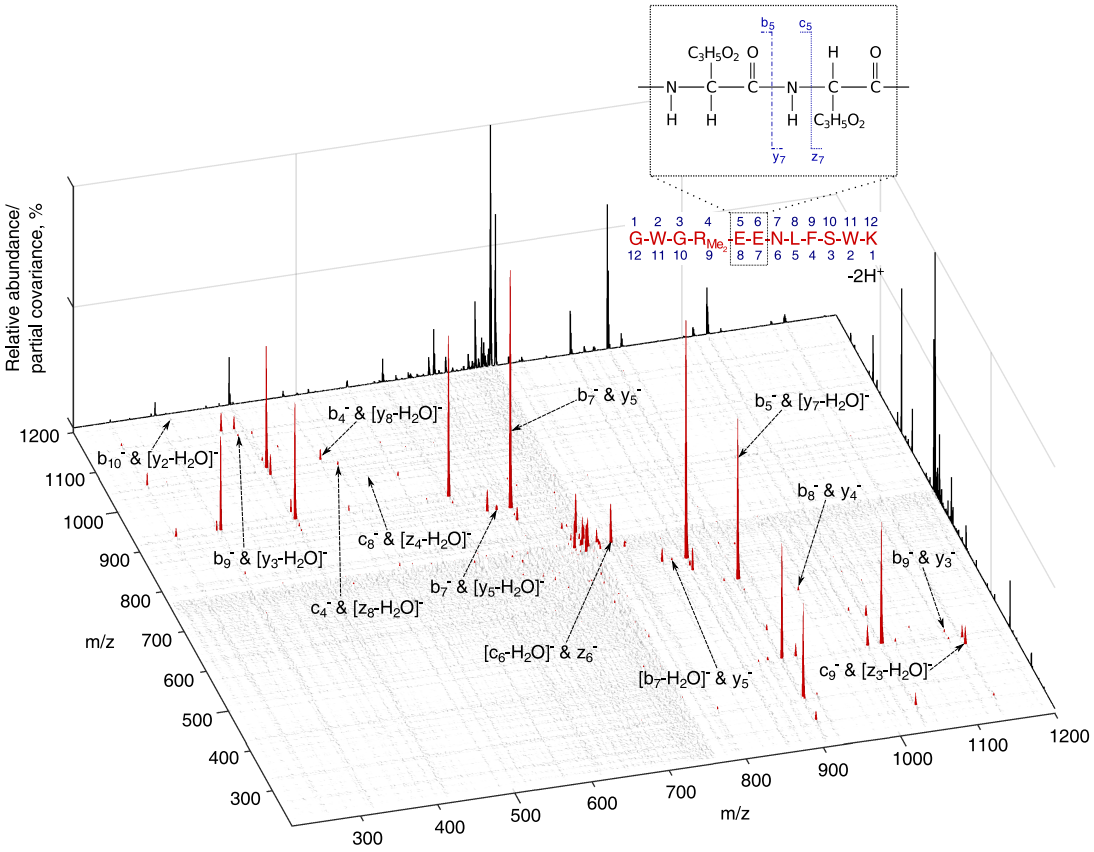


Figure 16.1: pC-2DMS map of the doubly deprotonated peptide ion of the sequence GWGR(Me$_2$)EENLFSWK (peptide **5**), measured under CID at normalised collision energy (NCE) 35%. The autocorrelation line has been removed for clarity. Selected 2D features are annotated and the 1D MS/MS spectrum is plotted against the back walls. Low-volume, true correlation features are unambiguously identified by calculation of the pC-2DMS correlation score (see section 6.2). The fragment ion nomenclature, which is analogous to that used for fragments produced from peptide cations (as proposed in [27] and refined in [190] for the case of the c/z fragmentation), is illustrated in the inset.

the proteome are not detected at sufficient levels to be allow for structural analysis in the standard tandem MS workflow. This is despite the fact that once the anions are present in a solution, electrospray ionisation is able to produce gaseous, intact, anionic biomolecules simply by reversing the polarity of the applied voltages from the typical positive ion setup [14]. Nonetheless, negative ion mode mass spectrometry has been demonstrated to have a number of key advantages over the more common positive ion mode analysis for a number of analytical tasks, such as the localisation of phosphorylated residues within regions of consecutive serine (S) and threonine (T) residues [191] and discrimination between the near-isomeric phosphorylation and sulphation of tyrosine (Y) residues [193] using CID, as well as more generally the extraction of sequence information from acidic peptides [194, 195].

Because pC-2DMS rests on the principle of partial covariance mapping, which requires only sufficiently high detection efficiency for its successful implementation, it is expected to be applicable for the analysis of any biomolecular species provided the fragmentation-to-detection transmission efficiency is high enough [42]. A possible obstacle for the realisation of pC-2DMS for biopolymeric anions comes in the form of the typically lower sensitivity of measurements in negative ion mode mass spectrometry, but modern instrumentation has developed a number of features to alleviate these potential setbacks such as the use of a conversion dynode to improve detection efficiency in negative ion mode, as described for the LTQ XL on page 55. To investigate the applicability of pC-2DMS to the analysis of biomolecular anions, pC-2DMS measurements were performed in the negative ion mode using the same CID parameters as the positive ion mode measurements but simply switching the polarity of the measurement.

The pC-2DMS map of the doubly deprotonated peptide GWGR(Me_2)EENLFSWK (peptide **5**) is shown in Fig. 16.1. The map shows that pC-2DMS can be successfully applied to linear ion trap CID measurements of deprotonated anions, providing a myriad of structurally informative signals resulting from both b/y- and c/z-type cleavages along the peptide backbone. The annotated signals have been identified using an adapted version of the automatic correlation matching software written as part

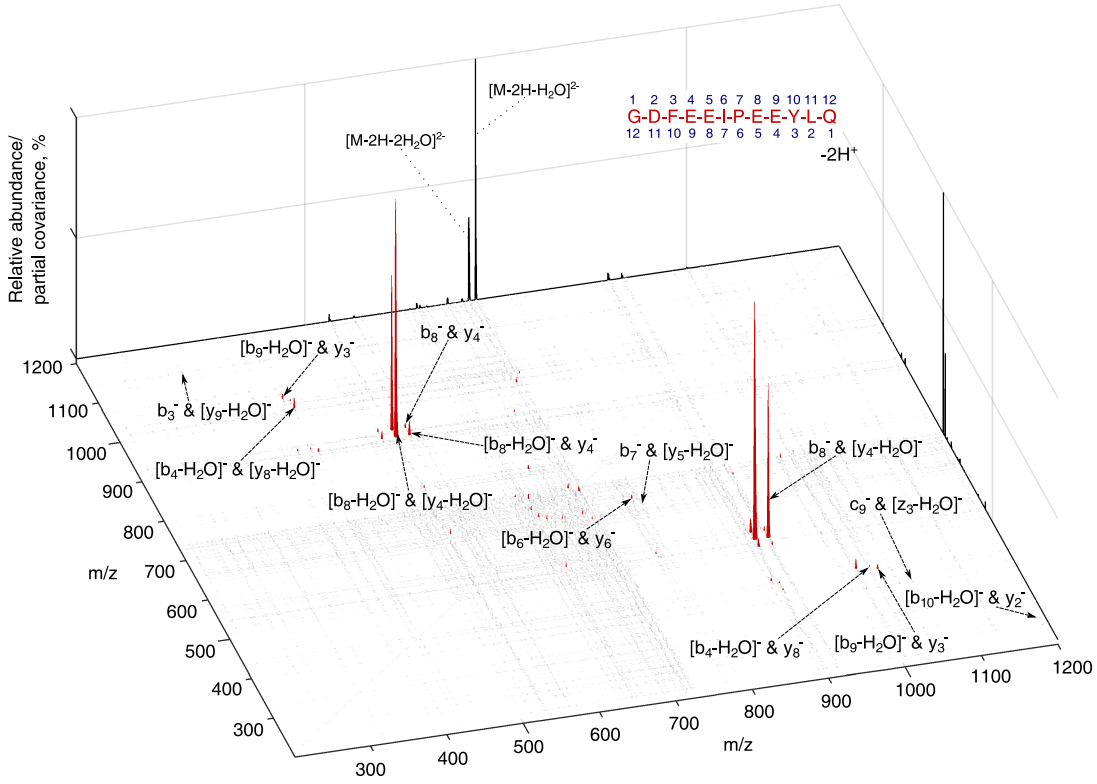


Figure 16.2: pC-2DMS map of the fragmentation of the doubly deprotonated sequence GDFEEIPEEYLQ (peptide **6**) under CID, at 35% NCE. The autocorrelation line has been removed for clarity. Selected 2D features are annotated (fragment ion nomenclature as for Fig. 16.1), as are the dominant neutral loss peaks in the 1D MS/MS spectrum plotted against the back walls of the map. Low-volume, true correlation features are unambiguously identified by calculation of the pC-2DMS correlation score (see section 6.2).

of the search engine described in Chapter 11. The software was adapted according to the known canonical backbone fragmentation rules for deprotonated peptides under CID [192]. Internal ions, which are often subject to sequence scrambling in negative ion CID and therefore risk being minimally structurally informative [192, 196, 197], are not considered by the software. Extremely low-volume, true correlation features have been unambiguously identified by calculation of the pC-2DMS correlation score introduced in section 6.2.

Further measurements were performed on different peptide anions to confirm the generality of pC-2DMS. The pC-2DMS map of the doubly deprotonated sequence

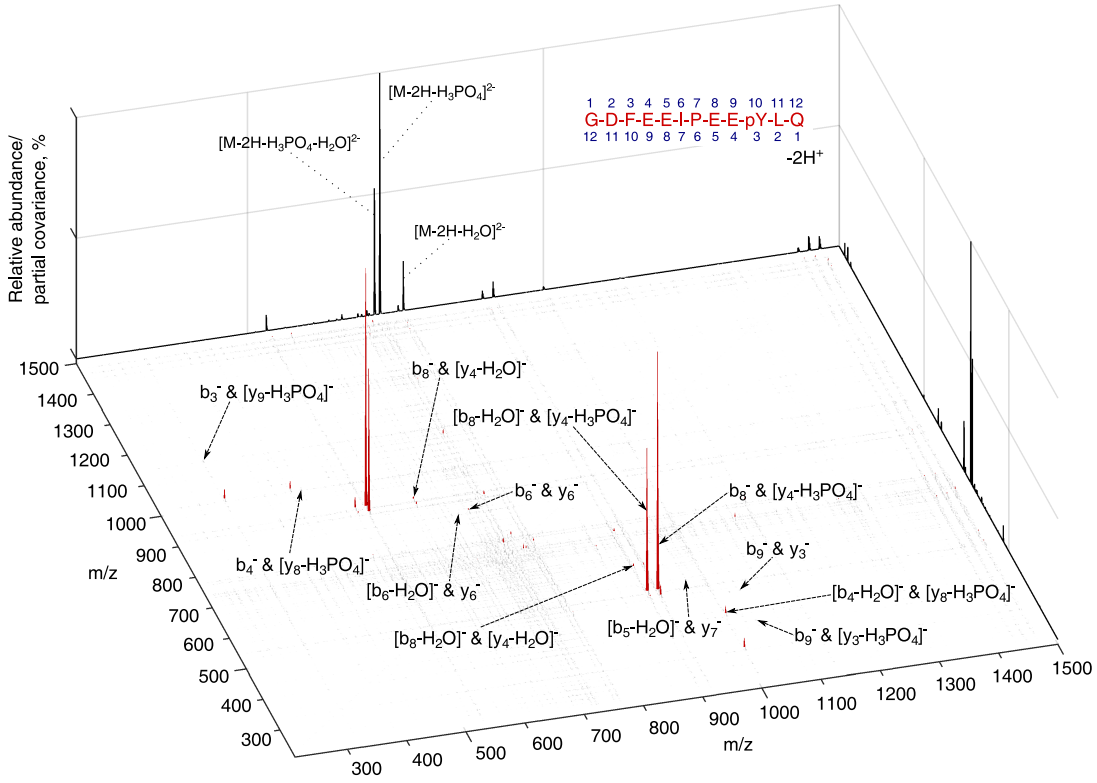


Figure 16.3: pC-2DMS map of the fragmentation of the doubly deprotonated phosphorylated sequence GDFEEIPEEpYLQ (peptide **7**) under CID, at 35% NCE. The autocorrelation line has been removed for clarity. Selected 2D features are annotated (fragment ion nomenclature as for Fig. 16.1), as are the dominant neutral loss peaks in the 1D MS/MS spectrum plotted against the back walls of the map. Low-volume, true correlation features are unambiguously identified by calculation of the pC-2DMS correlation score (see section 6.2).

GDFEEIPEEYLQ (peptide **6**) is shown in Fig. 16.2, and that of its phosphorylated counterpart GDFEEIPEEpYLQ (**7**, pY=phosphotyrosine) in Fig. 16.3. In both cases, the one dimensional spectra are dominated by neutral loss of small stable molecules from the parent ion: for peptide **6**, by elimination of H_2O from one of the abundant acidic residues (aspartic acid [D] or glutamic acid [E]) [194] and for the phosphopeptide **7**, by loss of H_3PO_4 from the phosphorylated tyrosine residue as well [193]. It is a useful general feature of pC-2DMS (see also e.g. Figs. 6.4 & A.5) that the common occurrence of highly efficient neutral loss of a small molecule from the parent ion does not cause the suppression of structurally informative signals by the dominant structurally uninformative neutral loss peak(s), as it does in 1D MS/MS. This is because such decomposition reactions by their very nature do not produce two detectable ions to be correlated (the neutral is undetected). As such, both the pC-2DMS in Figs. 16.2 and 16.3 maps bear a host of structurally informative signals generated by cleavages along the full peptide backbone, as automatically identified by the adapted fragment ion correlation matching software. For the phosphorylated peptide **7** in Fig. 16.3, as also demonstrated in Fig. 6.4, the pC-2DMS correlation score enables the unambiguous identification of correlations involving both intact fragment ions and fragment ions which have suffered a modification-specific neutral loss (e.g. the simultaneous observation of b_9^- & y_3^- and b_9^- & $[y_3\text{-}H_3PO_4]^-$), despite one or both of the relevant fragment ions appearing at extremely low abundance. As for Fig. 6.4, this important information greatly assists in the localisation of the modification to a particular residue.

Chapter 17

Extension to oligonucleotides

pC-2DMS has also been applied to measurements of deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) oligonucleotide ions. DNA is the carrier of genetic information, consisting of two polymeric chains which wrap around each other to form the famous double helix structure. The monomeric unit for each of these chains is the class of molecule known as nucleotides, which appear, for DNA, in four different types: adenine (given the single letter code A), cytosine (C), guanine (G) and thymine (T). Different types of nucleotide are dictated by the identity of the 'nucleobase' group which they hold (often simply referred to as 'bases'), in the same way that the functional group R^n dictates the type of amino acid for proteins (see Fig. 5.8). RNA is also a polymeric chain, but is more typically found as a single strand folded back on itself. It shares three of the same monomeric units with DNA, but with thymine being replaced by the nucleotide uracil (U). In the synthesis of proteins, the process of 'transcription' first creates an RNA sequence from a DNA sequence, and this RNA sequence is then read into a sequence of amino acids according to the 'genetic code'. The genetic code, which is strikingly similar across all organisms, relates a set of three bases to a particular amino acid and thus can be used in protein database search engines such as those described in Chapter 11 to construct a full protein sequence database from libraries of DNA or RNA sequences only. As is the case for most amino acids, the different nucleotides are separable by their mass and so in principle amenable to MS analysis.

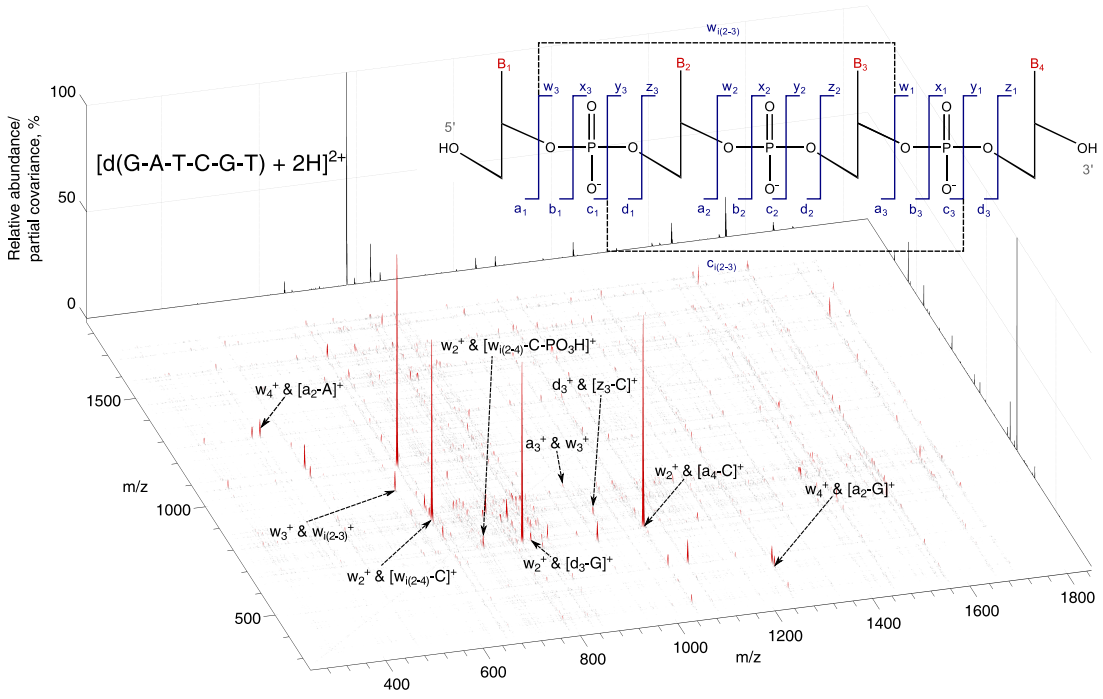


Figure 17.1: pC-2DMS map of the doubly protonated DNA oligonucleotide [d(GATCGT)+2H]$^{2+}$. The inset is an illustration of the oligonucleotide fragment ion nomenclature employed here, which is from reference [7]. Internal c-type fragments are annotated as $c_{i(n-m)}$ where n and m are the 5' and 3' nucleotides of the fragment sequence respectively, or as $c_{i(n)}$ where the fragment is one nucleotide long. The same subscript rules apply for internal w-type fragments. The autocorrelation line has been removed for visual clarity, and the 1D MS/MS spectrum is plotted against the back walls. This represents the first experimental realisation of pC-2DMS on an oligonucleotide ion. Assistance in interpretation of the DNA signals from Ruth Ayers, who also produced the inset illustration.

Mass spectrometry does not dominate the exploration of DNA and RNA primary structure as it does for proteins, thanks in part due to a number of extraordinarily successful, high-throughput methods for nucleotide sequencing based on chemical processes [198, 199]. These chemical sequencing methods are made particularly viable by the limited 'alphabet' (i.e. only five different nucleobases) of nucleotide sequences and the natural phenomenon of 'complementarity', where each nucleotide has one other nucleotide to which it highly preferentially bonds to *via* hydrogen bonding. They have achieved remarkable success, such as the sequencing of almost the entirety of the human genome (i.e. all the DNA information present in humans) [200]. The mass spectrometric analysis of nucleotide structure is nonetheless a well-developed and prosperous field [8, 9, 201]. One area of particular use for MS-based analysis is the study of epigenetic modifications such as DNA methylation [202], which often do not change the chemical properties of a base in way which is detectable by the standard chemical sequencing methods. As with post-translational modifications of peptides and proteins, the mass shift induced by such modifications renders MS a natural tool for their analysis. Of particular note here is the role that combinatorial patterns of DNA methylation patterns are believed play in the 'epigenetic code', of which histone modifications form a vital part [203]. For the same reasons which concern the combinatorial modifications of histone sequences which are described and resolved by pC-2DMS in Chapter 9, the problem of resolving combinatorial patterns of DNA methylation, thought to have significant biological relevance, remains an open and biologically important problem [123, 124].

Figure 17.1 shows the pC-2DMS map of the fragmentation of the doubly protonated DNA oligonucleotide sequence $[d(GATCGT)+2H]^{2+}$ under CID at NCE 35%. The basic structure of both DNA and RNA oligonucleotide sequences is shown the inset of this figure, along with the fragment ion nomenclature used. The different nucleotides, which are distinguished by different bases (indicated by B_n in red), are covalently held together by the PO_4 phosphodiester bond. Fragmentation at different points along this phosphodiester backbone produce the sequence ions whose nomenclature, proposed in [7], is illustrated. This terminology bears many similarities to

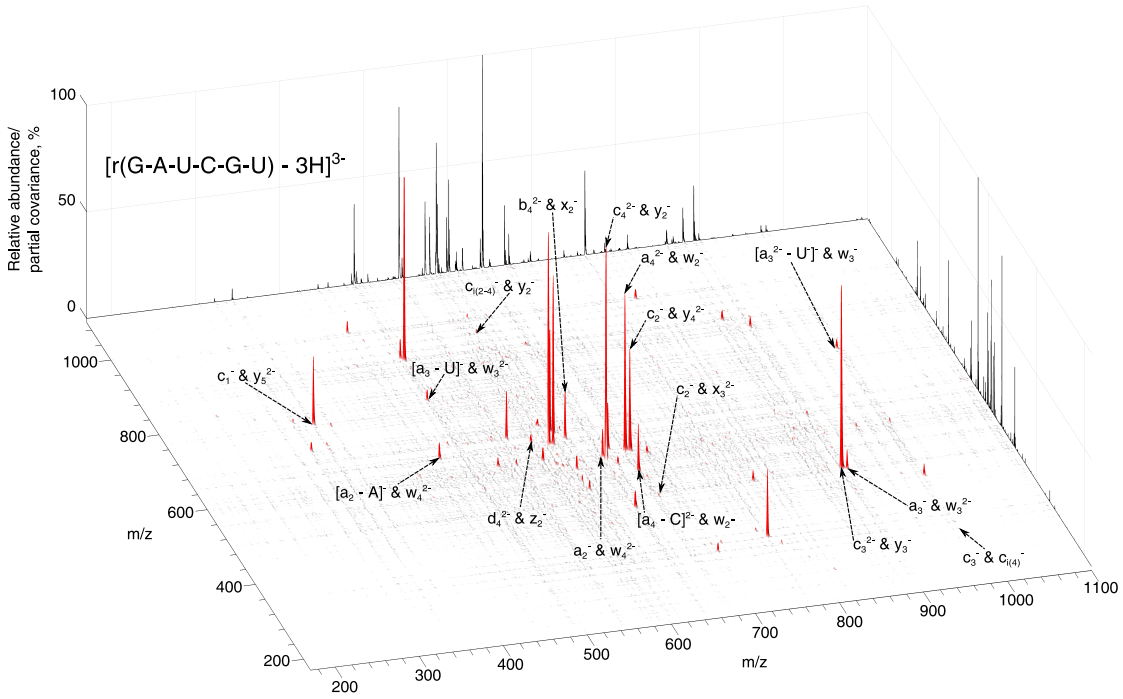


Figure 17.2: pC-2DMS map of the triply deprotonated negative ion of RNA oligonucleotide [r(GAUCGU)-3H]$^{3-}$. Fragment ion nomenclature is the same as in Fig. 17.1. The autocorrelation line has been removed for visual clarity, and the 1D MS/MS spectrum is plotted against the back walls. Interpretation of all signals from Ruth Ayers.

the terminology employed for peptide fragments, such as a subscripted number denoting the number of monomeric units in a particular fragment. The terminals of the oligonucleotide ion are named the 5' ("five-prime") and 3' termini ("three-prime"), instead of the N-terminus and C-terminus. Backbone fragmentation is often accompanied by the loss of one of the neutral or charged bases [9], indicated by e.g. '[a_2-A]' (loss of adenine from an a_2 fragment). The pC-2DMS map reveals a number of correlations induced by fragmentations across the oligonucleotide backbone. Figure 17.2 shows the pC-2DMS map of the fragmentation of the triply deprotonated RNA oligonucleotide sequence [r(GAUCGU)-3H]$^{3-}$ under CID at NCE 15%.

A publication demonstrating the utility of pC-2DMS for structural analysis of RNA oligonucleotide ions, which includes the application of the high specificity of pC-2DMS correlation signals (see Chapter 10) to the notoriously complex problem of identifying oligonucleotide internal ions, is currently in preparation by Ruth Ayers.

Chapter 18

Current shortcomings of pC-2DMS

This chapter briefly reviews the main disadvantages currently associated with pC-2DMS. The primary shortcoming of the technique stems from the need to acquire multiple repeated spectra – typically around 1000 individual scans per 2D map (see section 6.2). This requirement leads to much longer acquisition times compared to standard 1D MS/MS scans, as is the case for any other analytical 2D technique (2D infrared spectroscopy [204], 2D nuclear magnetic resonance spectroscopy [205], 2D Fourier transform ion cyclotron resonance MS [206]).

The vast majority of the work described in this thesis has been performed using a basic linear ion trap mass spectrometer (Thermo Fisher Scientific LTQ XL, first produced over 15 years ago) operating at the fastest acquisition rate it provides, which is 125 000 Da/sec at the fragment ejection step (see section 5.2) resulting in a scan frequency of $\sim$6 Hz. The increase in peak width associated with the use of a higher scan rate is compensated by spectral decongestion along the second dimension (see e.g. Fig. 14.5), high specificity of ion pair correlations even at relatively high mass tolerance (see Chapter 10, $\pm$ 0.8 Da is used for all work in this thesis) and the ability within pC-2DMS to determine fragment ion charge states without resolution of the isotopic envelope (see Chapter 13). Operation at the fastest possible scan rate for the LTQ XL still necessitates acquisition times of a few minutes for each precursor ion to record 1000 scans. This poses a challenge for the standard bottom-up proteomics workflow, which relies on chromatographic separation [1].

However, with the use of a quadrupole mass filter for precursor ion isolation, increased ion trapping efficiency, etc., the fastest instruments currently reach scan frequencies of ∼40 Hz for MS/MS linear ion trap measurements at the fragment ejection rate of 125 000 Da/sec [207]. Data acquisition times for pC-2DMS measurements can therefore be reduced to ∼25 s per precursor ion, bringing pC-2DMS into compatibility with the timescale of the standard bottom-up workflow. Additionally, it was demonstrated in Chapter 14 that pC-2DMS can be successfully applied to top-down experiments on intact proteins, and furthermore the potential for multiplex pC-2DMS measurements on either oligomers or intact proteins (as demonstrated in section 14.2) is vast, both of which considerably ease the requirements on chromatographic separation.

Chapter 19

Future work

19.1 Femtosecond laser-induced ionisation / dissociation (fs-LID)

A significant direction of research for the mass spectrometric community is the development of novel fragmentation methods for biopolymeric ions [32–34], to produce different fragment ions from which new and complementary information can be extracted. The Dantus group have demonstrated the coupling of a 3D quadrupole ion trap[1] to a femtosecond, infrared laser system producing < 35 fs laser pulses at a central wavelength of 800 nm [209]. In the proposed mechanism, tunnel ionisation of the most weakly bound electron is followed by the rapid non-ergodic dissociation of the resultant distonic ion [210], and this fragmentation method was shown to result in pervasive backbone cleavage across the protonated peptides under analysis, providing enhanced sequence coverage compared to the traditional CID activation [209]. Under this suggested mechanism the ultrafast molecular dissociation, which takes place on a femtosecond timescale [210], occurs before the molecule under analysis is able to undergo a redistribution of internal energy and so labile species such as protein phosphorylation are retained on the fragment ions. This is in contrast to traditional

[1]Where, in comparison to the linear ion trap described in section 5.2, trapping is achieved by applying an alternating RF field to a central ring-shaped electrode situated between two grounded hyperbolic 'end-cap' electrodes [208].

activation methods such as CID, where labile species such as protein phosphorylation are typically lost following internal energy redistribution before backbone fragmentation occurs [108]. This limits the possibility for localisation of the modification to a particular amino acid. Using femtosecond laser-induced ionisation / dissociation (fs-LID), the Dantus group has demonstrated extensive sequence coverage combined with preservation of serine (S) and threonine (T) phosphorylation on the originally modified fragment ion in measurements of protonated [211] and deprotonated [212] phosphorylated peptides.

Future work in laboratory H005 in the Blackett Laboratory Laser Consortium[2] aims at performing fs-LID in a linear ion trap system for the first time, by coupling a linear ion trap mass spectrometer to a Coherent *Legend* femtosecond laser system, which produces ∼30 fs pulses at a wavelength centred around ∼800 nm and pulse energy ∼1.5 mJ at a repetition rate of 1 kHz. The linear ion trap mass spectrometer has a number of key advantages over the 3D quadrupole ion trap for mass spectral analysis, such as a higher injection efficiency and greater trap capacity (meaning improved ability to detect fragment ions produced at lower relative abundance) [81]. For the implementation of fs-LID, greater spatial overlap with the laser pulse can be expected for the cylindrical ion cloud in a linear ion trap compared with the spherical ion cloud in a 3D ion trap [55]. Moreover, the application of pC-2DMS analysis to fs-LID spectra is also be expected to provide a number of critical advantages: the non-resonant character of the laser activation is suitable for multiplex measurements such as that demonstrated in section 14.2, extended irradiation times will produce secondary dissociation assisting in the production of e.g. internal ion correlations as required to resolve combinatorial post-translational modifications (see Chapter 9), pC-2DMS is insensitive to the non-fragmented parent ion in cases of limited parent ion fragmentation [209], and generally in the case of complex, pervasive peptide fragmentation, 2D correlation signals will prove useful for fragment ion identification (by virtue of their dramatically enhanced specificity, see Chapter 10) and mechanistic elucidation (see Chapter 12).

[2]Where the pC-2DMS experiments described in this thesis took place

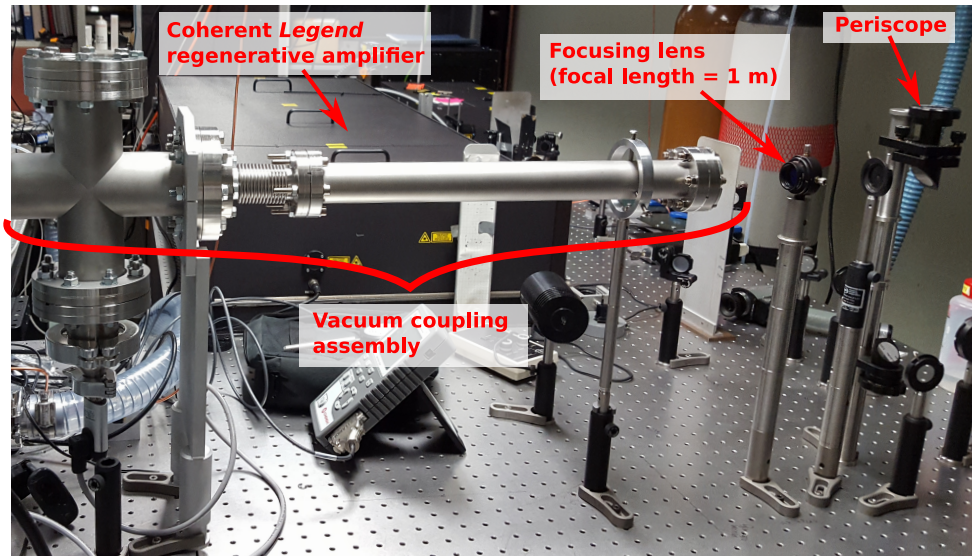


Figure 19.1: Photograph of the constructed setup for implementation of fs-LID in a linear ion trap in laboratory H005, Imperial College London. The long vacuum arm and relatively long (1 m) focal length of the lens which focuses the beam into the trap, are to ensure the beam passes into the components under vacuum before its focused intensity becomes high enough to burn the fused silica window through which it enters. The periscope, which will be enclosed in a protective housing during operation, is to raise the beam from its height at the output of the *Legend* system to the height of the linear ion trap in the mass spectrometer.

To this end, the *Legend* laser system has undergone restoration over the course of the past four years and the initial work to realise fs-LID on a linear ion trap has been performed, in collaboration with Amelle Zaïr, Thomas Barillot and Ruth Ayers. Both an optical and a vacuum setup[3] to couple the laser system to the mass spectrometer have been designed, constructed[4], and initially aligned. The current state of the setup is shown in Fig. 19.1. Further work in the laboratory will bring the laser to its specified operating parameters before attempting to measure the first fs-LID spectra of trapped biomolecules in a linear ion trap.

[3]Vacuum setup designed by Leszek Frasinski
[4]With much help from Andy Gregory

19.2 Computer vision for identification of true correlation signals

The pC-2DMS correlation score (see section 6.2) performs extremely well at distinguishing true correlation features from high-volume features due to statistical noise. The calculation of this score is, however, an inherently computationally heavy task requiring significant resources and time. For this reason, other methods for distinguishing true correlation signals from statistical noise have also been investigated. One promising indicator of statistical noise is the three-dimensional shape of a feature on a pC-2DMS map; it is typically straightforward to identify a small true correlation island from a larger, false correlation island due to statistical noise by visual inspection. Figure 19.2 compares the shapes of a representative example of true correlation islands (panel **a**), identified by the pC-2DMS correlation score and further manual verification, with high-volume false correlation islands determined by the same correlation score (panel **b**), from the pC-2DMS map of the triply protonated ion $[\mathrm{VTIMPK_{Ac}DIQLAR}]^{3+}$ (peptide **1**) which is shown in full on page 71. All features have been individually normalised to their highest value for optimal visualisation of their shape. Distinguishing characteristics of false correlation islands include dispersed, highly non-uniform shape, surrounding regions of negative partial covariance and significant substructure. Design of a standard single parameter which encapsulates all these characteristics is challenging, but thanks to their combined effect the identification of statistical noise by eye is straightforward. The application of neural networks (NNs) and, particularly, convolutional neural networks (CNNs) to such image classification problems has proven astonishingly successful across a huge number of different disciplines in recent years [213]. The automated identification of true and false correlation islands by such an algorithm would be orders of magnitude less computationally heavy than calculation of the pC-2DMS correlation score.

To this end, a simple neural network-based peak classification system for identification of true correlation signals *vs.* statistical noise was first built and tested. The input was the value of the partial covariance in each of the 32×32 2D spectral bins de-

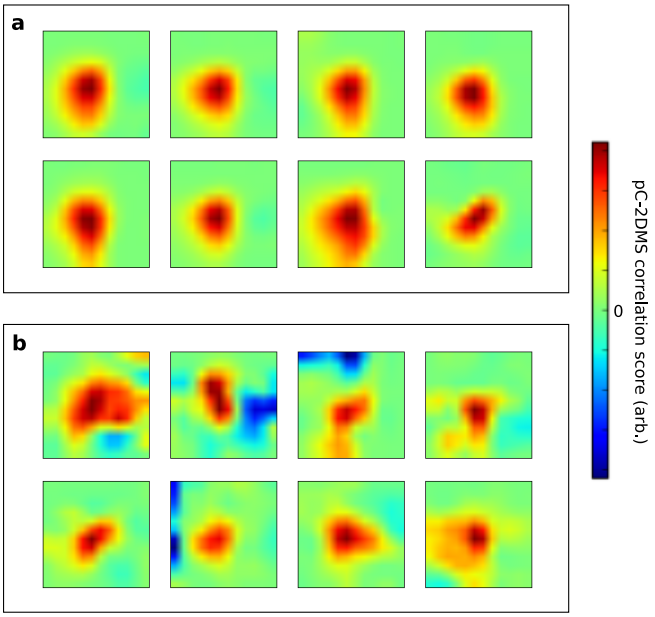


Figure 19.2: Three-dimensional shape of **a** true correlation features identified by the pC-2DMS correlation score and verified by manual analysis and **b** high volume correlation features due to statistical noise (determined by the pC-2DMS correlation score) for the pC-2DMS map of the triply protonated ion of peptide **1**, as shown in full in Fig. 6.1. The colour scale, denoting the relative value of the partial covariance, is normalised to the maximum value of each feature in turn for optimal illustration of the feature shapes. Inspection by eye shows a clear difference in the 3D shapes of true correlation signals *vs.* statistical noise, suggesting classification by machine vision is a possibility.

fined around the highest point of each feature (as plotted in Fig. 19.2). A training set was constructed using the pC-2DMS measurements of ions given in Table 7.1, using high pC-2DMS correlation scoring features subjected to further manual verification to identify true correlations islands, and high volume features with a significantly low pC-2DMS correlation score (ranked below 100) to determine statistical noise (see e.g. Fig 6.4 for examples of such features). All algorithms were implemented by manual construction of a 'graph' using the open source TensorFlow software [214]. The computationally efficient, stochastic Adam algorithm [215] was used for gradient descent, and all training was performed over 2000 epochs. Pre-processing involved normalisation of all pixel values to the maximum value of the most voluminous true correlation feature on the pC-2DMS map from which it was taken, and the pixel values were truncated at the negative value of this maximum. Mean subtraction was not performed, which preserved the sign of the partial covariance values.

A simple, single-layer neural network returned a test accuracy[5] of 90.6% for identification of true correlation signals *vs.* statistical noise, which was modestly improved to 91.8% by a deep neural network (DNN) including two fully-connected hidden layers with 1024 nodes each. A simple convolutional neural network was then constructed and trained, for which each peak extracted from the map was treated as a monochromatic image with pixel values corresponding to the value of partial covariance in the 2D spectral bins (which, unlike a standard image, can also assume negative values). The simple model featured two convolutional layers, each with kernel size 5×5 (stride size 1) and the ReLU activation function followed by a pooling layer which performed 2×2 max pooling (stride size 2). This simple architecture showed a promising increase to 95.3% test accuracy on the same data used for the simple NN. Optimisation of the convolutional neural network architecture, along with further development of feature pre-processing, should provide significantly more accurate feature classification.

[5]Both categories evenly sampled for the test set.

19.3 Transient ionic species

The ability of pC-2DMS to identify true correlation signals between ions where one or both is at extremely low abundance has been demonstrated throughout this thesis (see e.g. Figs. 6.4, 16.1, 6.1, 7.6 and appendix A). In pC-2DMS measurements of a small group of peptides, one or more high pC-2DMS-scoring correlations which suggest the transient migration of a small neutral molecule from one fragment to the other have been observed. The proposed mechanism resulting in this observation is transient hydrogen bonding[6] between the neutral and a fragment produced in the same molecular decomposition as the fragment to which the neutral was originally attached, producing transient species of sufficiently low relative abundance that they are not reliably detected in standard 1D MS/MS.

Figure 19.3 shows a scatter plot of the top 50 pC-2DMS correlation-score ordered signals from the dissociation under CID of the doubly protonated peptide ion of the sequence GGNFSGR(Me)GGFGGSR (peptide **8**). Two of these high-scoring features correlate standardly observed H_2O-loss features from canonical b- and y-ions ($[y_5\text{-}H_2O]^+$ and $[y_6\text{-}H_2O]^+$) with unexpected ions demonstrating the transfer of the lost H_2O to their complementary fragments ($[b_9+H_2O]^+$ and $[b_8+H_2O]^+$, respectively). The high-scoring correlations of the same complementary ions without H_2O transfer are also labelled on the scatter plot, indicating the relevant backbone fragmentations occur with a high branching ratio. The insets of the 1D spectrum demonstrate the extremely low abundance at which the transient species featuring addition of H_2O are detected in the 1D CID spectrum, entirely precluding their detection with standard 1D MS.

pC-2DMS measurements on the doubly protonated peptide ion of RDY(SO_3H)TGW-Nle[7]-DF(NH_2) (peptide **2**, see Fig. 6.4) also produced high-scoring correlation features suggesting the transient, hydrogen-bonding mediated migration of the labile SO_3 (mass=80 Da) group from its original b-ion to the complementary y-ion: $[b_3\text{-}SO_3]^+$ (expected at 435.2 Da) & $[y_6+SO_3]^+$ (expected at 817.3 Da), $[b_4\text{-}SO_3]^+$ (expected

[6]Hydrogen bonds are formed on a significantly shorter timescale than covalent bonds [216].
[7]Nle = norleucine

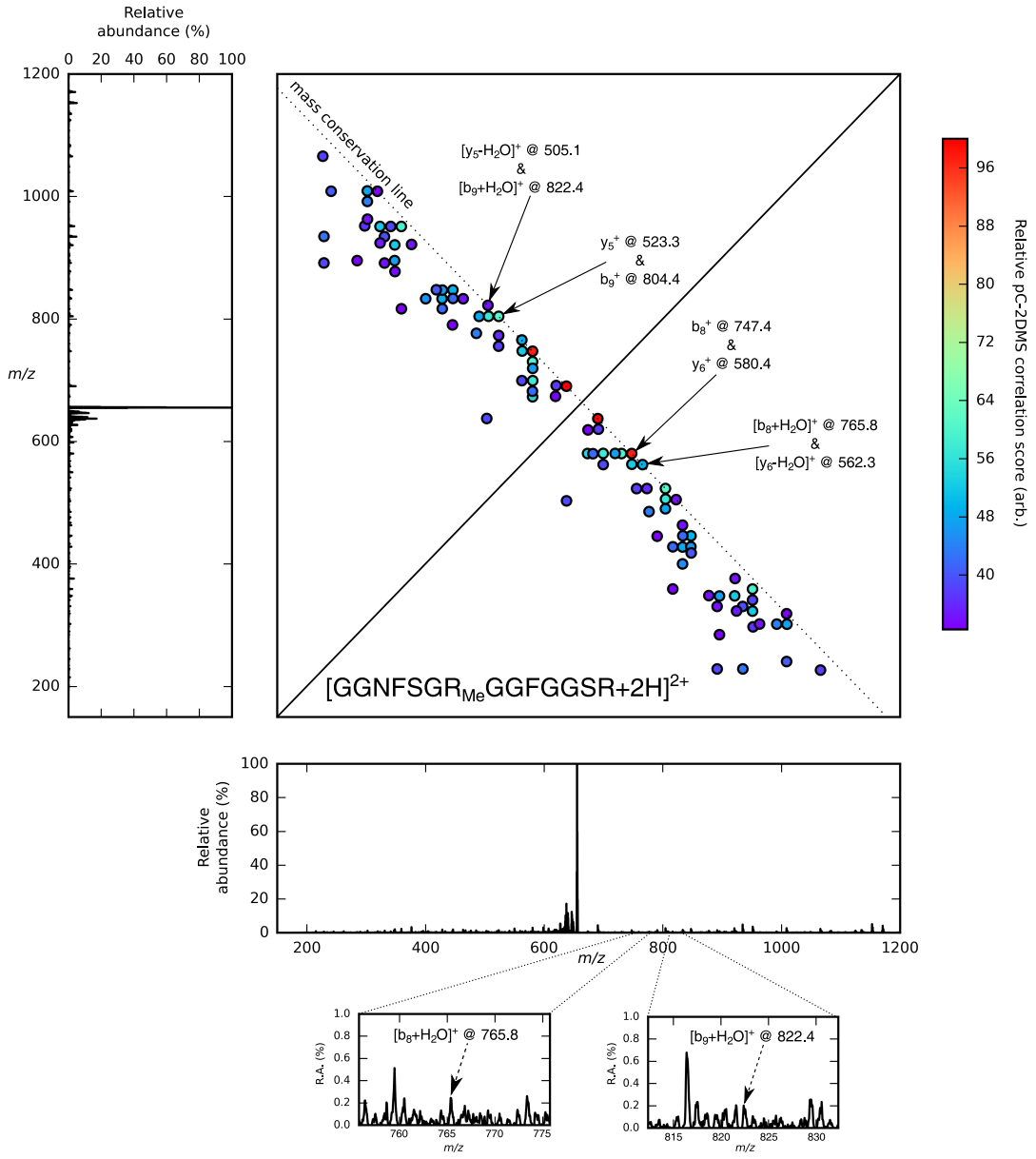


Figure 19.3: Observation of complementary ions showing transfer of H_2O from one complementary b/y fragment to the other in the pC-2DMS measurement of doubly protonated peptide **8**, [GGNFSGR(Me)GGFGGSR+2H]$^{2+}$. This is proposed to be the transient addition of H_2O, eliminated by one complementary fragment, to the other complementary fragment *via* hydrogen bonding.

at 536.2 Da) & $[y_5+SO_3]^+$ (expected at 716.3 Da) and $[b_6\text{-}SO_3]^+$ (expected at 779.3 Da) & $[y_3+SO_3]^+$ (expected at 473.2 Da). Migration of the phosphate group from a phosphorylated residue to an unmodified residue has been observed under CID using 1D MS/MS, causing ambiguity in the assignment of the original modification state [162, 217]. The transfer of a glycine residue from the C-terminus to the N-terminus in pC-2DMS measurements on the triply protonated ion of peptide **8** ($[y_{11}\text{-}G]^{2+}$ & $[b_3+G]^+$) has also been observed, but the transferred glycine is expected to be covalently bound to the b_3^+ fragment which would occur on a significantly longer timescale than transfer mediated by hydrogen bonding.

These interesting observations require further experimental work and detailed analysis to confirm their validity, explore the proposed mechanism and test their generality. This work is planned for the near future.

Chapter 20

Synchrotron and XFEL studies of molecules

This chapter describes two large collaborative experiments at international facilities which the author participated in during his PhD.

Much of the research currently carried out in the Laser Consortium focuses on the ultrafast nuclear and electronic dynamics of photoexcited atoms and molecules. In the H012 and H007 laboratories, ultrafast dynamics have been probed using a number of different techniques such as high harmonic generation spectroscopy [218, 219] and transient absorption [220]. Much work has also been performed on the development of different experimental capabilities for pump-probe experiments on gaseous (and, more recently, condensed phase) molecular species. The general procedure for such experiments is to use two different ultrashort[1] laser pulses, separated by a variable delay, to examine the dynamics of the system in question - one to initiate the ultrafast dynamics (the 'pump' pulse) followed by another which creates the observable providing information on the state of the system after a delay (the 'probe' pulse). These observables can be e.g. ion time-of-flight spectra, photoelectron spectra, photon spectra or Auger electron spectra. Calegari et al. [221] measured the yield of the immonium ion produced by a 4 fs, 1.77 eV pulse after a variable time delay following photoexcitation by an isolated sub-300 as pulse in the extreme ultraviolet regime

[1] i.e. few femtosecond (fs) to attosecond (as) regime

(pulses between 15 eV and 35 eV). They observed an exponential decay of this yield over several tens of femtoseconds, which was modulated by oscillations with a period of 4.3 fs. This period is shorter than the highest possible frequency vibrational response of the molecule, so these oscillations were attributed to so-called charge migration effects in the cation, a purely electron-driven phenomenon first predicted by Cederbaum and Zobeley [222]. Briefly, sudden ionisation is expected to create a coherent superposition of several cationic eigenstates, so that initially localised hole density undergoes rapid evolution driven by the quantum interference of these different electronic states as they evolve over time. Such dynamics are broadly referred to as being driven by 'electron correlation'.

It should be noted that the theoretical framework first set out in [222] only considers electronic motion, with the molecular nuclei treated as fixed points at all times. Recent work from e.g. co-workers at the Department of Chemistry at Imperial College London suggests that the coupling of nuclear and electronic motion is very important for the observation of the so-called hole migration. Nuclear motion was shown to cause a decoherence between the coherently populated cationic states, resulting in a damping of oscillatory motion [223]. Moreover, a simple model including the inherent width of the nuclear wavepacket at the time of initial ionisation, leading to a spread in the energy difference between populated excited states, was shown to lead to a dephasing of the initial coherent superposition and therefore lack of long-lived oscillations for the molecules paraxylene and polycyclic norbornadiene [224]. The effect of all coupling between nuclear and electronic behaviour was found to be highly molecule-dependent, and measurement of the existence and persistence of these hole migration effects remains a frontier challenge for understanding the motion of electrons in the first few moments of a chemical reaction.

The isopropanol molecule, shown in Fig. 20.1, is a good candidate for experiments measuring charge migration in photoexcited polyatomic molecules. Thanks to its room temperature vapour pressure of 52 mbar (298 K, corresponding to a density of $\sim 5 \times 10^{17}$ molecules/cm^3) it can be straightforwardly delivered to an experimental chamber by a hypodermic needle, and its simple structure makes in-depth theoretical

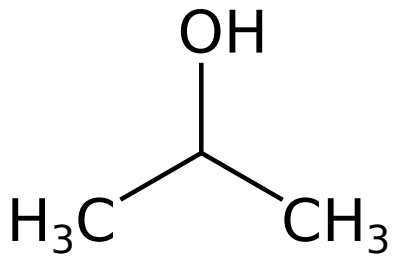


Figure 20.1: Stick diagram of the isopropanol molecule

analysis of the molecular response comparatively straightforward. Furthermore, isopropanol has a large photoionisation cross-section of 120 Mb [225] at 15 eV meaning it will be efficiently pumped by ultrashort pulses in the vacuum ultraviolet spectral region, which can be routinely produced by high harmonic generation (HHG) [226]. The H007 laboratory at Imperial College has recently demonstrated the high harmonic generation of synchronised attosecond pulses at 20 eV (576 ± 16 as) and 90 eV (257 ± 21 as) [227]. The ability to produce two such synchronised attosecond pulses allows for the possibility of a so-called 'attosecond pump - attosecond probe' experiment. In addition to the improved time resolution, this name fails to capture the desirable quality of such experiments that in using a high harmonic pulse to probe the system, strong field dressing of the system is avoided. The high electric field necessitated by the multiphoton processes required at lower photon energies for a visible/near-infrared probe pulse strongly distorts the electric field felt by the system under analysis, obscuring the dynamic effects under investigation [228]. X-ray pump, X-ray probe experiments are now also a possibility at X-ray free electron laser (XFEL) facilities such as the Linac Coherent Light Source at the SLAC National Laboratory, California, U.S.A., at which the experiment in section 20.2 was performed.

20.1 Synchrotron SOLEIL beamtime Barillot-20160424, Nov 2016. Measurement of isopropanol photoelectrons from a monochromatic source.

The beamtime at SOLEIL was proposed in order to aid the interpretation of future pump-probe experiments in isopropanol and glycine, which is the simplest amino acid with a single hydrogen as the functional group (see section 5.3). The more challenging experimental requirements for the delivery of gaseous glycine to the interaction chamber (in laboratory H007 this is performed using an 'oven' which directly sublimes crystalline glycine) meant that the beamtime was eventually restricted to the measurement of isopropanol. The aim of the beamtime was to obtain a series of photoion-photoelectron coincidence measurements of isopropanol under irradiation with the highly monochromatic light afforded by synchrotron radiation. At the SOLEIL synchrotron, bunches of electrons are accelerated first in a linear accelerator and then in a second, circular accelerator to energies of 2.75 GeV. They are then sent to a 113 m diameter storage ring where they are made to move in a circular orbit for several hours thanks to a series of magnetic dipoles. These magnetic dipoles, along with an array of undulators, cause the electrons to accelerate and emit synchrotron radiation. The energy lost by the electrons in radiating is compensated for by radiofrequency cavities. The PLEIADES (Polarized Light source for Electron and Ion Analysis from Diluted Excited Species) beamline provides low divergence radiation at photon energies 10 eV - 1 keV with a resolution of $\sim 1 \times 10^5$ at 50 eV photon energy, at any polarisation. We performed photoion-photoelectron coincidence measurements of isopropanol at photon energies 20 eV and 90 eV, which reflects the experimental capabilities for high harmonic pulses detailed in reference [227].

The photoion-photoelectron measurements were performed using the EPICEA coincidence setup developed by the group at the PLEIADES beamline [230], which couples a three dimensional focusing ion time-of-flight mass spectrometer to a double toroidal electron analyser (DTA) [229, 231]. Following retardation or acceleration by

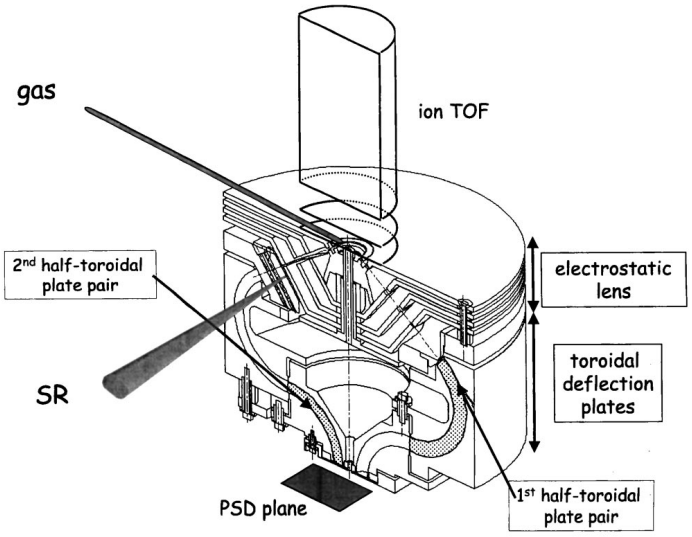


Figure 20.2: Mechanical drawing of the EPICEA detector, prior to adjustment of the electrostatic lens to loosen the focusing for higher energy electrons. Electrons produced by the synchrotron radiation (SR) are guided through the toroidal deflection plates onto a position-sensitive detector (PSD). Reproduced with permission from reference [229].

a user-defined electric field which is applied by an electrostatic lens focusing the electrons on an entrance slit, the DTA employs two toroidal deflecting plates to radially disperse electrons according to their kinetic energy on emission. The electrons are measured at a position-sensitive detector (PSD), as seen in the mechanical illustration of the detector setup in Fig 20.2[2]. The central energy of the electrons which pass through the DTA without being deflected onto one of the toroidal plates, and are therefore detected, is determined by the so-called 'pass energy' which is defined by the voltages applied to the toroidal deflecting plates. The DTA provides a detection bandwidth of 12% of the central electron pass energy and a resolving power of $\sim 1\%$ of the pass energy, at an acceptance angle of roughly 5% of the full 4π steradians [229]. The design of a series of so-called 'lens tables' (optimised combinations of voltages applied to different electron optics in the setup for particular pass energies) through computer simulations and laboratory-based tuning has been able to push the resolution down to between 0.6% and 0.8% of the pass energy whilst also increasing the

[2]The electrostatic lens in the design used for the experiment was modified from that shown in the figure in order to loosen the focusing for higher energy electrons on the analyser slit and so reduce divergence of the electron beam entering the DTA.

overall electron transmission [232]. The retarding or accelerating potential applied to the electrons by the electrostatic lens determines the initial kinetic energy of the electrons which are eventually detected (detection requires the sum of the initial kinetic energy and the retarding/accelerating voltage to fall within the detection bandwidth of the given pass energy).

Clearly, the design of the DTA means it can resolve the angular azimuthal distribution of the detected electrons. The position of the opening of the toroidal deflection plates means that the detected electrons are those emitted at the so-called 'magic angle', $\theta_m =\sim 57.4^\circ$ measured from the drift axis of the ion TOF. According to the Cooper-Zare formula, the differential photoionisation cross section for linearly polarised light is dependent on the second order Legendre polynomial with argument as the angle between the incident electric field and photoemitted electrons. The magic angle is a root of the second order Legendre polynomial, which means that for electrons emitted at $\theta =\sim 57.4^\circ$ with respect to the polarisation of the incoming electric field, the differential cross section for each process is independent of θ and the measured electron yield is directly proportional to the photoionisation cross-section in question.

Ions are measured in the EPICEA by applying extraction voltages to the interaction chamber to send the ions into a 3D focusing ion time-of-flight (TOF) mass spectrometer. The ion TOF spectrometer was introduced to provide the capability for electron-photoion coincidence measurements, but the small acceptance angle of the DTA makes it challenging to accurately correlate detected electrons with the large number of ions which are created and measured at each shot. In order to overcome this issue, the ion TOF operates in so-called pulsed extraction mode, where the extraction voltages are applied only upon the detection of an electron by the DTA. An estimate for the ionic background beneath which the relevant correlated ion or ions are buried at each ion TOF measurement is found by randomly triggering the extraction voltages at additional moments, regardless of whether electrons have been detected or not.

Calibration of the DTA was performed using the valence photoelectrons from

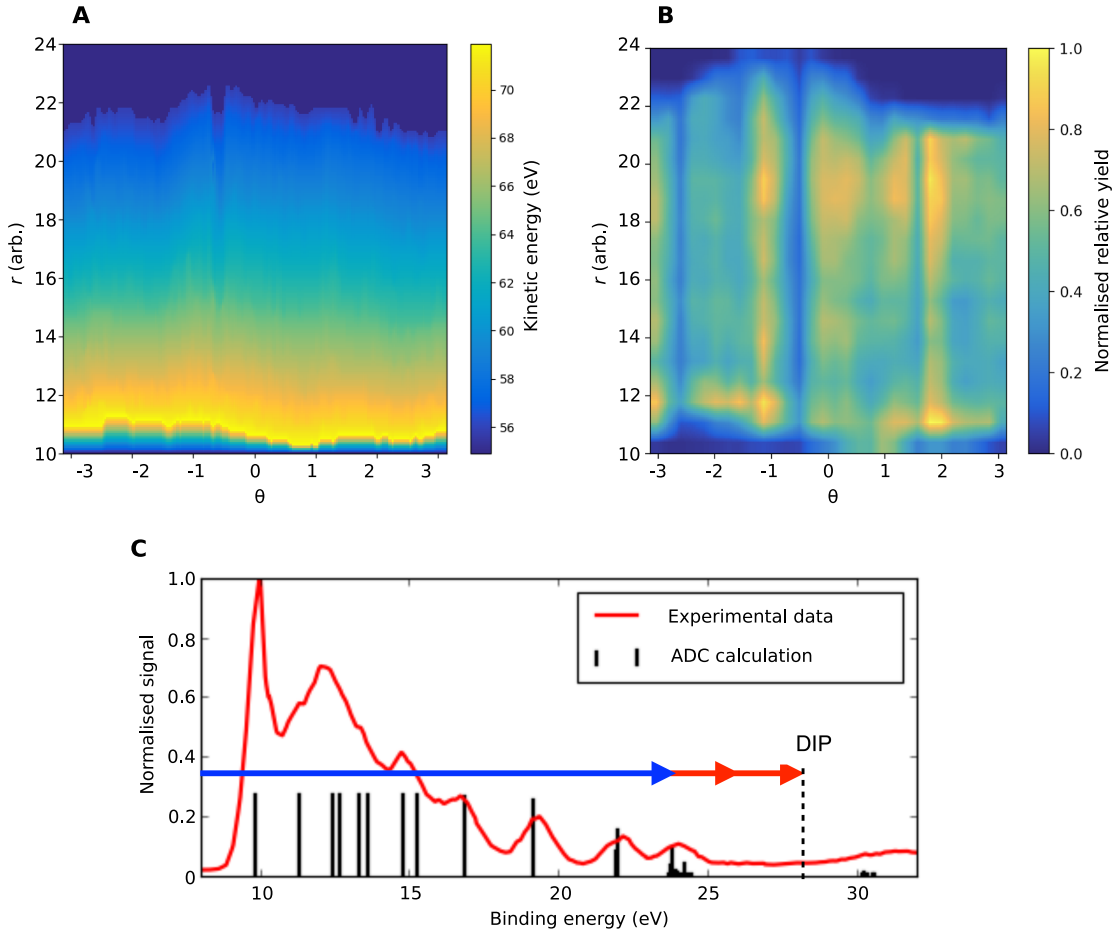


Figure 20.3: Initial experimental results from SOLEIL beamtime (data unpublished). **A**, kinetic energy calibration map for DTA electron detector and **B** detector response calibration map for the same instrument, at 67 eV electron pass energy. r and θ are the radial and angular coordinates of the detected position of electrons on the position-sensitive detector. Both plots courtesy of Aorui Tan. **C**, measured photoelectron spectrum of isopropanol at 90 eV photon energy (red line) shows good agreement with the energetic positions of the molecular states from ADC calculations by Bridgette Cooper (black sticks). The height of the sticks represents the 'one-hole, one-particle' character of the resultant cationic state; normalised to one, the smaller the stick, the greater the multiparticle character of the state. The horizontal arrows represent the photon energies of an originally proposed ultraviolet pump (blue) and infrared probe (red). 'DIP' is the double ionisation potential of the molecule. Plot courtesy of Thomas Barillot, data unpublished.

krypton ($I_p \sim 14$ eV). This was used to construct calibration maps relating polar coordinates to photoelectron energy, one example of which is shown in Fig. 20.3 **A** (kinetic energy of electron detected at a particular point on the PSD denoted by the colour scale). This map has been created by scanning the synchrotron photon energy between 69 eV and 87 eV in steps of 1 eV at a pass energy of 120 eV and accelerating potential set to 53 eV so that the central kinetic energy of the detected electrons is 67 eV. By using the relative valence photoionisation cross-sections extracted from [233] and normalising the scan at each photon energy to the estimated overall photon flux, a map of the detector response (primarily a result of varying transmission efficiencies through the toroidal deflection plates) was also constructed and is shown in Fig. 20.3 **B**. This time, the colour scale denotes the relative detector response.

Following calibration, measurements were performed on isopropanol. The measured photoelectron spectrum of isopropanol at a photon energy of 90 eV, obtained by scanning different accelerating voltages at a set pass energy of 120 eV in order to cover the full bandwidth of the photoelectron spectrum, is shown in Fig. 20.3 **C**. The measured spectrum falls in good agreement with the energetic positions of the expected molecular states determined by Bridgette Cooper using the algebraic diagrammatic construction (ADC) scheme, plotted as black sticks. The data measured at the Synchrotron SOLEIL is currently being used to assist in the ongoing analysis of a number of experimental runs attempting to measure ultrafast dynamics in photoexcited isopropanol.

20.2 LCLS beamtime LR25, Nov-Dec 2017. Measurement of ultrafast hole migration in photoexcited cationic isopropanol.

One further use of the isopropanol photoelectron spectra recorded at the Synchrotron SOLEIL is to assist in the ongoing analysis of the data taken by the LR25 experimental collaboration, which attempted to measure ultrafast electronic motion in

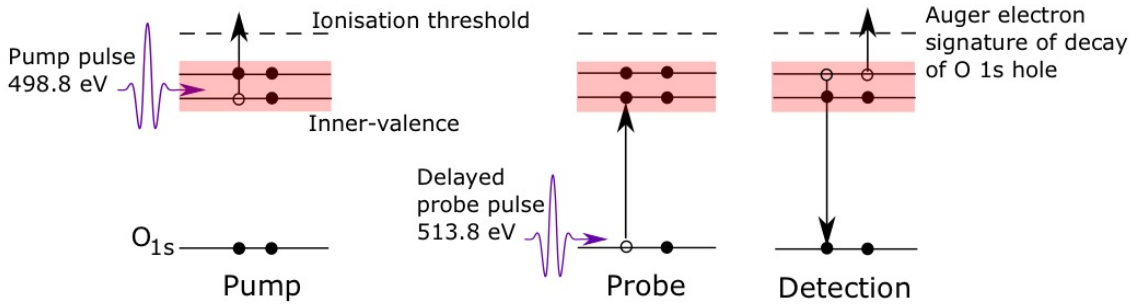


Figure 20.4: Schematic of the proposed two-colour X-ray pump-probe experiment.

valence ionised isopropanol using a two-colour X-ray pump, X-ray probe technique at the LCLS. Beam time was awarded for five shifts of twelve hours between the 30th November and 4th December 2017.

The experiment proposed to measure the ultrafast migration of an electron hole created in an inner valence shell of the following ionisation by an X-ray photon. The general scheme is illustrated in Fig. 20.4, and very similar to the scheme analysed in detail in [234] for the simplest amino acid, glycine[3]. A pump pulse, tuned below the oxygen K-edge, quasi-instantaneously ionises from an inner valence shell of the oxygen (stick diagram of isopropanol in Fig. 20.1). This sudden ionisation is expected to drive ultrafast charge migration, as predicted by calculations performed for isopropanol by Bridgette Cooper at Imperial College London (author of [234]) and shown in Fig. 20.5. The different coloured lines depict the survival probability $S(t)$ for a hole initially created in different single inner valence orbitals of the neutral isopropanol molecule, and the behaviour shown is a general mix of rapid local decay, relative stability and oscillatory dynamics on a timescale of $\sim$ 10 fs. It was proposed to probe such dynamics by a variably delayed pump pulse, resonantly tuned to the transition from the oxygen 1s shell to an inner valence state. This short-lived core hole state would decay *via* the Auger process, with filling of the core hole by an inner valence electron accompanied by the Auger emission of another inner valence electron. The measured yield of these Auger electrons should be directly proportional

[3]An attempt to apply this technique to glycine was made in January 2015 (beam time LF62), but was thwarted by the very unstable operation of the two-bunch, two-colour mode used at that time

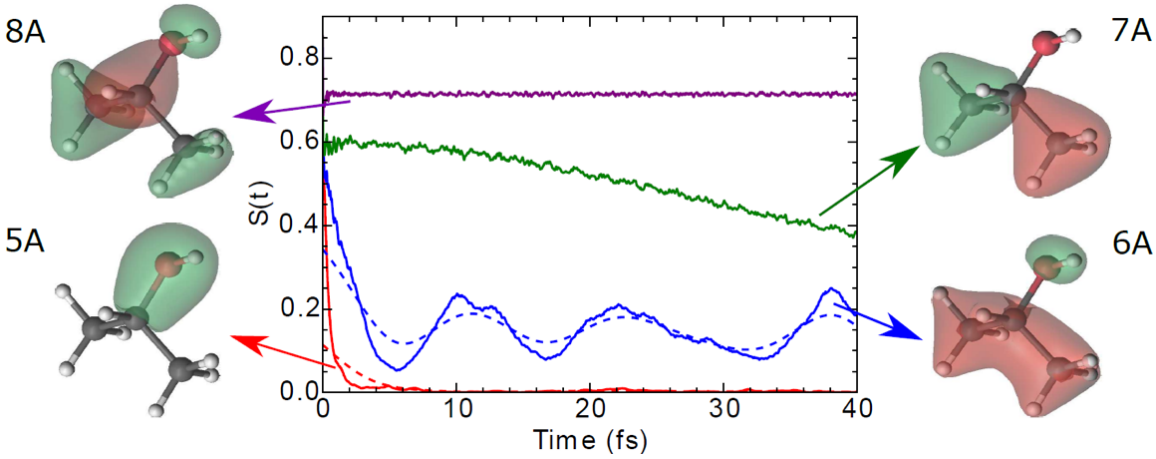


Figure 20.5: Calculated survival probabilities for holes initially created and localised in different inner valence shells of the isopropanol molecule. Dashed lines show the convolution with a 5 fs probe pulse. Calculations are for a 50:50 mix of the two lowest energy conformers, and the shapes of the Hartree-Fock orbitals are shown for the gauche conformer. Figure courtesy of Bridgette Cooper.

to the survival probability of the hole at a given time, and the theoretical Auger spectrum calculated according to the methods described in section 3 of [235] predicted significant discontiguity between the Auger spectrum and the inevitable photoelectron signal [236] from the two pulses. Usefully, the ability to tune the photon energy of the probe pulse (up to a jitter, which is unproblematic because of the shot-to-shot spectral measurement afforded by the X-ray spectrometer provided by Uppsala University) should enable a state-specific probe, which is able to selectively address the holes in different inner valence shells. This is done by tuning the photon energy to resonance with the transition from the oxygen 1s shell to a particular, energetically different, inner valence orbital.

The LCLS is the world's first hard X-ray free electron laser [237], making use of one-third of the linear accelerator at SLAC. The linear accelerator was originally commissioned in 1966 as the world's longest linear accelerator, and has produced a vast body of influential research on sub-atomic structure. For the LCLS, high-brightness electron bunches (0.25 $s^{-1}mm^{-2}mrad^{-2}$ per 0.1% spectral bandwidth) are accelerated using radiofrequency fields to energies of up to ∼15 GeV in the final kilometre of the linear accelerator. Lasing is achieved by passing the bunch through

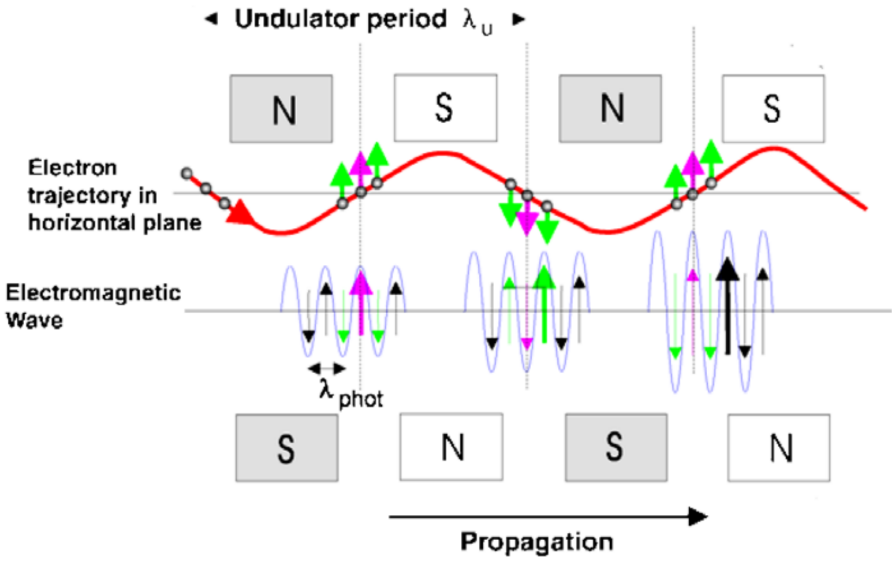


Figure 20.6: Motion of electrons through an undulator with radiation at the resonant wavelength. The radiation 'slips' ahead by one full wavelength in the time it takes the transverse electron motion to complete one full oscillation period. The three different electrons, shown at different stages in their propagation through the undulator, all have transverse velocity locked in different (as indicated by green and purple arrows) but constant phase relations with respect to the radiation field. The interaction of the electrons with the radiation field will eventually cause a spatial modulation of the electron density at the wavelength of the radiation field, meaning increasingly more electrons radiate coherently. Reproduced with permission from [238].

a series of magnetic undulators, where the repetitive structure of regions of high magnetic field with alternating direction cause them to 'wiggle' and so radiate in their direction of travel.

The production of coherent laser light is achieved by the useful collective interaction between the radiating electrons and the radiation they produce. As the electron bunch travels through the undulator there is a coherent build-up of radiation at the resonant wavelength λ_{phot}, which is dictated by the undulator period λ_u and the longitudinal velocity of the electrons as illustrated in Fig 20.6. For electrons travelling at a given velocity through an undulator, radiation is coherently produced at a resonant wavelength λ_{phot} such that in the time it takes for an electron to travel one full undulator period λ_u, the electron has 'slipped' back one full wavelength of the radiation field (it is inevitably travelling at a speed slower than the speed of light). For the radiation

field at this wavelength, the transverse oscillations of the electron are in phase with the radiation and so the radiation produced as the electron wiggles coherently adds to the field. Additionally, the electrons in the electron bunch feel the radiation field they produce as a spatially periodic acceleration or deceleration along their direction of travel, at the wavelength of the radiation. Over time, this modulation in longitudinal electron velocity becomes a modulation in longitudinal density of the electron bunch with the same periodicity as the radiation field (so-called 'micro-bunching'), so that more and more electrons radiate coherently together at the resonant wavelength. This 'positive feedback' process results in an exponential increase in the intensity of the radiation field until a saturation regime is reached, where the electrons begin to re-gain energy from the radiation field [51, 238].

Fresh-slice mode In order to produce two different X-ray pulses at two different wavelengths, the newly developed 'fresh-slice mode' of operation [239] was employed for the experiment. This is a method to produce two (or three) X-ray pulses at variable delay and controlled wavelength by causing different parts of the same electron bunch to lase to saturation in the different undulator sections available at the LCLS. The general setup for two-colour two-pulse generation, as used in this experiment, is shown in Fig. 20.7. The fresh-slice technique exploits the fact that only electrons travelling in a straight line through the oscillator can efficiently lase [240]. A dechirper, situated before the undulators at the LCLS and originally designed to compensate for the deleterious energy chirp of an electron bunch induced by bunch compression [241], is able to passively chirp the accelerated electron bunch by causing the energy of the tail to be reduced *via* interaction with its own wakefield. If an electron bunch enters this dechirper off-axis it also experiences a time-varying transverse kick, as illustrated by the yellow arrows on the electron bunch in the dechirper. The kick experienced by the tail is greater than that experienced by the head. Following the dechirper, the electron bunch is steered so that the tail travels in a straight line through the first undulator, meaning the head is undergoing large betatron oscillations as it travels.The tail can lase to saturation, producing the pump pulse (orange oval), whilst the head

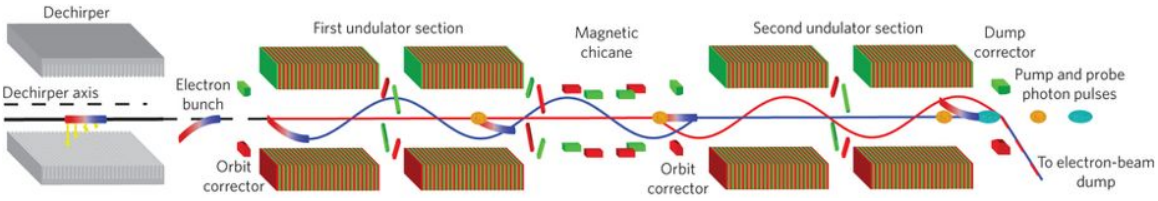


Figure 20.7: Illustration of generation of two pulses at two different colours using fresh-slice mode, reproduced with permission from [239]. The electron bunch tail (red) lases in the first undulator to produce the first pulse (orange oval), and the head (blue) lases in the second undulator to produce the second pulse (light blue oval).

does not lase because it is not travelling in a straight line. After the magnetic chicane, which imparts a variable delay on the electron bunch and therefore the pump-probe separation, a set of magnetic dipoles are used to steer the bunch so that now the head is travelling in a straight line and the tail experiences large betatron oscillations. Therefore, in the second undulator the 'fresh slice' of the head can lase to saturation to produce the pump pulse (blue oval), whilst the tail does not. By changing the delay imparted to the electron bunch in the magnetic chicane, the pump and probe pulses can be temporally scanned through each other. The electrons were measured with a semi-hemispherical electron spectrometer built by Scienta, whose general operating principle is described in detail in [242].

Preliminary analysis Following instrumental testing and calibration and preliminary scans, a series of X-ray pump, X-ray probe scans were performed on the isopropanol molecule at varying nominal[4] time delays between -15 fs and 30 fs. Analysis of the experimental data is ongoing.

The time-dependent observable in this experiment is the yield of resonant Auger electrons resulting from the decay of the oxygen 1s hole which is emptied by the probe pulse, as illustrated in Fig. 20.4. However, the full measured pump-probe spectrum is inevitably highly congested, consisting of the overlapping photoelectrons emitted in numerous different possible interactions of the molecule with the pump pulse, the

[4]Determination of the exact time delay induced by a particular set of magnetic chicane parameters in fresh-slice mode is a part of the continuing analysis.

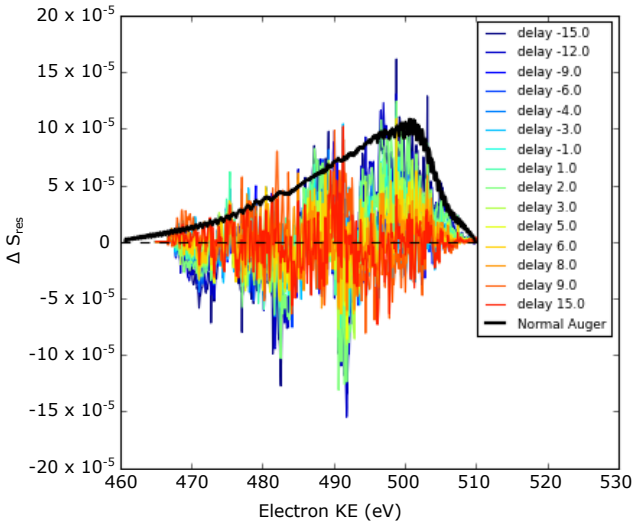


Figure 20.8: Measured isopropanol Auger spectrum at central photon energy 650 eV (black). The coloured lines are the change in residual signal (i.e. following removal of retrieved pump-only and probe-only spectra) at different nominal time delays.

probe pulse or both pulses together. The normal Auger spectrum of isopropanol was measured during the LR25 experimental run at a central photon energy of 650 eV, and is plotted (normalised) in black in Fig. 20.8, in which the coloured lines denote the time dependence of an extracted residual signal from the electron spectrum. This residual signal was determined by first identifying a 'pump-only' and 'probe-only' spectrum for the experiment with a procedure which is now described for extraction of the pump-only spectrum, but whose description for extraction of the probe-only spectrum is identical, except with 'pump' and 'probe' reversed. To extract the pump-only spectrum at different pump photon energies, scans across all time delays were interrogated to find those at low pump pulse energy and those at high pump pulse energy, as determined by the shot-to-shot X-ray photoelectron spectrum provided by the X-ray spectrometer from Uppsala University. The retrieved pump-only signal at a given photon energy was the difference between the averaged spectrum at high pump pulse energy and the averaged spectrum at low pump pulse energy, for the given pump photon energy. This method was assessed by comparing the retrieved pump-only signal from measurements where the probe was intentionally suppressed (this couldn't be done for the pump for technical reasons), with the two spectra

showing good agreement. Once the retrieved pump- and probe-only spectra had been extracted for all pump and probe photon energies, the appropriate pump- and probe-only spectra were removed from each shot (following normalisation according to the measured shot-to-shot pulse energies) at each time delay, with the average of the resultant spectra plotted as the coloured lines in Fig. 20.8. The spectral shape of this residual signal appears to match the shape of the measured Auger spectrum, which is promising. Of course, much further analysis is required. Another ongoing attempt to deconvolve the pump-probe spectra is the use of the static photoelectron spectra measured at the Synchtrotron SOLEIL, such as the one in Fig. 20.3, to remove the pump- and probe-only signals, by using a Gaussian basis defined by a fitting of the static spectra to fit pump- and probe-only signals in the pump-probe spectra.

Another potential method to extract the required signal from the overlapping pump-probe spectra is to exploit the expected photon energy dependence of the resonant Auger electron(s). Because the mechanism producing such electrons is a resonant[5] Auger process, their signal should only appear across a very narrow bandwidth of probe photon energies (width determined by the very short lifetime of the excited state which directly precedes the Auger emission), across which a linear dependence of Auger electron kinetic energy on probe photon energy is expected[6]. To attempt to identify a resonant Auger signal by making use of this photon energy dependence, partial covariance maps were constructed between the photon spectrum and the electron spectrum. The partial covariance map plotted in Fig 20.9, panel **a** has been calculated for all shots in which both a pump and probe pulse were produced by the lasing bunch, for seven runs at nominal time delays between -6 fs and 6 fs, in steps of 3 fs. The exact location of t_0 (perfect temporal overlap between the pump and the probe) is not yet fully defined but should fall within this range. In the 1D averaged photon spectrum (showing the pump pulse centred around ∼501 eV and the probe pulse around ∼515 eV), the spectral peaks of the pump and probe sit on significant

[5]Named 'resonant' because the hole which the relaxing electron fills is initially emptied by a resonant transition between two bound states, instead of photoionisation into the continuum as for 'non-resonant' Auger.

[6]The deviation from exact resonance of the probe photon energy, permitted by time-energy uncertainty, is carried away by the departing Auger electron.

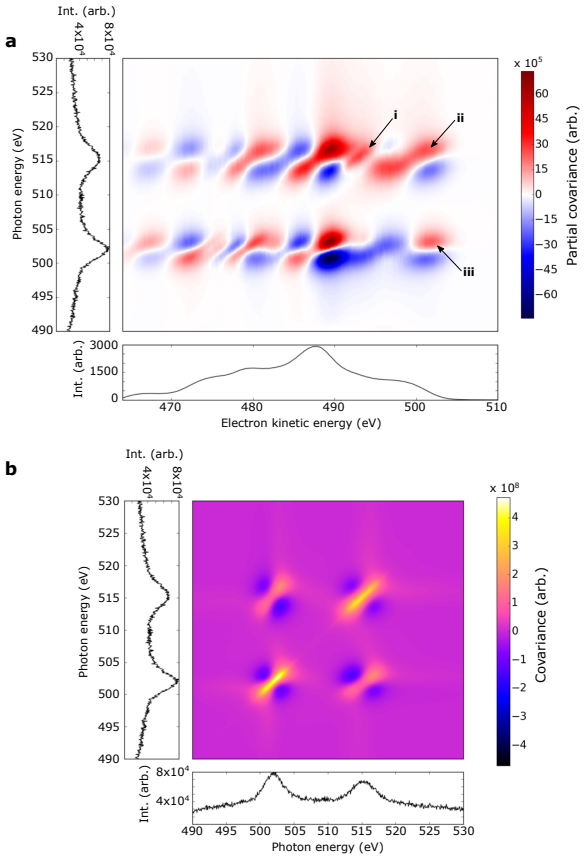


Figure 20.9: **a**, partial covariance map between electron and photon spectrum for five runs at nominal delays −6 fs, −3 fs, 0 fs, 3 fs and 6 fs. The one-dimensional mean averaged spectra are plotted along the side of the map. The pump pulse can be seen centred around ∼501 eV, with the probe pulse centred around ∼515 eV. The partial covariance parameter I was the two-pulse energy measured by the LCLS front-end gas detectors. Features **i**, **ii** and **iii** are discussed in the text. **b**, autocovariance map of the photon spectrum with one-dimensional mean averaged spectra plotted along the side of the map.

background noise, which does not appear on the partial covariance maps because it is very nearly static with any small fluctuations entirely random. The partial covariance parameter is the measured photoluminescence stimulated by the two pulses in nitrogen gas in the front-end gas detectors at the LCLS, which provides a good measure of the highly fluctuating number of photons in each pulse [243]. The inherent photon energy jitter of the LCLS (evident in the peak bandwidth in the averaged 1D photon spectrum) is exploited in this partial covariance map, which provides information on the behaviour of the electron spectrum across a large bandwidth for both pump and probe pulses. As expected, almost all of the partial covariance islands demonstrate a linear dependence of the electron kinetic energy on photon energy, straightforwardly established because the islands of positive partial covariance are tilted at a gradient of $\sim$1. A notable exception to this is the island correlating electron kinetic energy $\sim$500 eV with photon energy $\sim$515 eV (labelled **ii**), which appears to have a gradient of $\sim \frac{1}{2}$. This indicates an electron produced in a two-photon process between two photons of highly correlated photon energy (see below), where the change in the photon energy of each of the photons results in a change of double the size in the resultant electron energy. The regions of significant negative covariance between higher energy electrons and lower energy photons around each correlation island is mirrored in the photon-photon covariance map plotted in panel **b**, which shows regions of strong negative covariance around all the islands between lower pump (probe) photon energies and higher probe (pump) photon energies. Negative covariance between lower and higher energies in the same pulse is straightforwardly explained by the shot-to-shot photon energy jitter (trivially, shots centred around higher central photon energies produce lower numbers of photons at lower photon energies), and negative covariance between lower and higher energies in different pulses is due to the fact that both pulses are produced from the same electron bunch, whose fluctuating characteristics affect the pump and the probe pulse in a similar manner.

Feature **i** in panel **a** of Fig. 20.9 indicates a potential candidate for a resonant Auger signal. This feature correlates an electron peak at $\sim$495 eV (close to the peak of the measured Auger spectrum in Fig. 20.8) with a narrow bandwidth of the probe

spectrum, as expected for an electron produced by a resonant process depending on probe photon energy. The feature shows an approximately linear dependence of electron energy on probe photon energy across this small bandwidth. Analysis is ongoing to determine the physical origin of this correlation signal. Of additional note is feature **iii**, which correlates an electron kinetic energy of ∼501 eV with the pump pulse centred around ∼501 eV. A single pump photon interacting with the neutral molecule should produce electrons at a maximum kinetic energy of ∼491 eV, because the ionisation potential of the most weakly bound electrons in isopropanol is ∼10 eV. An electron produced by the pump pulse and measured at a kinetic energy higher than this, as revealed by the partial covariance map, is therefore expected to have been ejected from a molecule which has already been photoexcited. Three-dimensional covariance [244] between the photon spectrum, photon spectrum and electron spectrum was also calculated in an attempt to isolate electrons produced by interaction between the pump and probe pulse only, however the statistics used were not high enough to produce sufficient signal-to-noise on the 3D covariance map[7].

[7] A rough rule-of-thumb is that a three dimensional covariance map requires an order of magnitude higher number of scans to produce the same signal-to-noise ratio as its 2D counterpart.

Bibliography

[1] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics", *Nature* **422**, 198 (2003).

[2] V. H. Wysocki, K. A. Resing, Q. Zhang, and G. Cheng, "Mass spectrometry of peptides and proteins", *Methods* **35**, 211 (2005).

[3] B. T. Chait, "Mass spectrometry: bottom-up or top-down?", *Science* **314**, 65 (2006).

[4] R. Aebersold and M. Mann, "Mass-spectrometric exploration of proteome structure and function", *Nature* **537**, 347 (2016).

[5] T. K. Toby, L. Fornelli, and N. L. Kelleher, "Progress in top-down proteomics and the analysis of proteoforms", *Annual Review of Analytical Chemistry* **9**, 499 (2016).

[6] B. F. Cravatt, G. M. Simon, and J. R. Yates, "The biological impact of mass-spectrometry-based proteomics", *Nature* **450**, 991 (2007).

[7] S. A. McLuckey, G. J. Van Berkel, and G. L. Glish, "Tandem mass spectrometry of small, multiply charged oligonucleotides", *Journal of the American Society for Mass Spectrometry* **3**, 60 (1992).

[8] E. Nordhoff, F. Kirpekar, and P. Roepstorff, "Mass spectrometry of nucleic acids", *Mass Spectrometry Reviews* **15**, 67 (1996).

[9] J. Wu and S. A. McLuckey, "Gas-phase fragmentation of oligonucleotide ions", *International Journal of Mass Spectrometry* **237**, 197 (2004).

[10] S. Schürch, J. M. Tromp, and S. T. M. Monn, "Mass spectrometry of oligonucleotides", *Nucleosides, Nucleotides and Nucleic Acids* **26**, 1629 (2007).

[11] R. Harkewicz and E. A. Dennis, "Applications of mass spectrometry to lipids and membranes", *Annual Review of Biochemistry* **80**, 301 (2011).

[12] J. Zaia, "Mass spectrometry of oligosaccharides", *Mass Spectrometry Reviews* **23**, 161 (2004).

[13] M. Mann and O. N. Jensen, "Proteomic analysis of post-translational modifications", *Nature Biotechnology* **21**, 255 (2003).

[14] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse, "Electrospray ionization for mass spectrometry of large biomolecules.", *Science* **246**, 64 (1989).

[15] M. Karas and F. Hillenkamp, "Laser desorption ionization of proteins with molecular masses exceeding 10 000 daltons", *Analytical Chemistry* **60**, 2299 (1988).

[16] A. K. Shukla and J. H. Futrell, "Tandem mass spectrometry: Dissociation of ions by collisional activation", *Journal of Mass Spectrometry* **35**, 1069 (2000).

[17] R. G. Sadygov, D. Cociorva, and J. R. Yates, "Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book", *Nature Methods* **1**, 195 (2004).

[18] D. F. Hunt, J. R. Yates, J. Shabanowitz, S. Winston, and C. R. Hauer, "Protein sequencing by tandem mass spectrometry.", *Proceedings of the National Academy of Sciences* **83**, 6233 (1986).

[19] H. Lam, E. W. Deutsch, J. S. Eddes, J. K. Eng, N. King, et al., "Development and validation of a spectral library searching method for peptide identification from MS/MS", *Proteomics* **7**, 655 (2007).

[20] K. F. Medzihradszky and R. J. Chalkley, "Lessons in de novo peptide sequencing by tandem mass spectrometry", *Mass Spectrometry Reviews* **34**, 43 (2015).

[21] E. W. Deutsch, H. Lam, and R. Aebersold, "Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics", *Physiological Genomics* **33**, 18 (2008).

[22] R. S. Johnson, M. T. Davis, J. A. Taylor, and S. D. Patterson, "Informatics for protein identification by mass spectrometry", *Methods* **35**, 223 (2005).

[23] A. Michalski, E. Damoc, J.-P. Hauschild, O. Lange, A. Wieghaus, et al., "Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer", *Molecular & Cellular Proteomics* **10**, M111.011015 (2011).

[24] G. Tsaprailis, H. Nair, Á. Somogyi, V. H. Wysocki, W. Zhong, et al., "Influence of secondary structure on the fragmentation of protonated peptides", *Journal of the American Chemical Society* **121**, 5142 (1999).

[25] B. Paizs and S. Suhal, "Fragmentation pathways of protonated peptides", *Mass Spectrometry Reviews* **24**, 508 (2005).

[26] Y. Huang, J. M. Triscari, G. C. Tseng, L. Pasa-Tolic, M. S. Lipton, et al., "Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns", *Analytical Chemistry* **77**, 5800 (2005).

[27] P. Roepstorff and J. Fohlman, "Proposal for a common nomenclature for sequence ions in mass spectra of peptides.", *Biomedical Mass Spectrometry* **11**, 601 (1984).

[28] B. A. Garcia, "What Does the Future Hold for Top Down Mass Spectrometry?", *Journal of the American Society for Mass Spectrometry* **21**, 193 (2010).

[29] N. L. Young, M. D. Plazas-Mayorca, and B. A. Garcia, "Systems-wide proteomic characterization of combinatorial post-translational modification patterns", *Expert Review of Proteomics* **7**, 79 (2010).

[30] D. Phanstiel, J. Brumbaugh, W. T. Berggren, K. Conard, X. Feng, et al., "Mass spectrometry identifies and quantifies 74 unique histone H4 isoforms in differentiating human embryonic stem cells", *Proceedings of the National Academy of Sciences* **105**, 4093 (2008).

[31] N. Abshiru, O. Caron-Lizotte, R. E. Rajan, A. Jamai, C. Pomies, et al., "Discovery of protein acetylation patterns by deconvolution of peptide isomer mass spectra", *Nature Communications* **6**, 1 (2015).

[32] J. E. P. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt, "Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry.", *Proceedings of the National Academy of Sciences of the United States of America* **101**, 9528 (2004).

[33] R. A. Zubarev, N. L. Kelleher, and F. W. McLafferty, "Electron capture dissociation of multiply charged protein cations. A nonergodic process", *Journal of the American Chemical Society* **120**, 3265 (1998).

[34] J. S. Brodbelt, "Photodissociation mass spectrometry: New tools for characterization of biological molecules", *Chemical Society Reviews* **43**, 2757 (2014).

[35] C. R. Weisbrod, M. R. Hoopmann, M. W. Senko, and J. E. Bruce, "Performance evaluation of a dual linear ion trap-Fourier transform ion cyclotron resonance mass spectrometer for proteomics research", *Journal of Proteomics* **88**, 109 (2013).

[36] M. Scigelova, M. Hornshaw, A. Giannakopulos, and A. Makarov, "Fourier Transform Mass Spectrometry", *Molecular & Cellular Proteomics* **10** (2011).

[37] L. J. Frasinski, A. J. Giles, P. A. Hatherly, J. H. Posthumus, M. R. Thompson, et al., "Covariance mapping and triple coincidence techniques applied to multielectron dissociative ionization", *Journal of Electron Spectroscopy and Related Phenomena* **79**, 367 (1996).

[38] L. J. Frasinski, K. Codling, and P. A. Hatherly, "Covariance mapping: a correlation method applied to multiphoton multiple ionization", *Science* **246**, 1029 (1989).

[39] L. J. Frasinski, K. Codling, and P. A. Hatherly, "Multiple multiphoton ionization of N2 probed by covariance mapping", *Physics Letters A* **142**, 499 (1989).

[40] P. A. Hatherly, M. Stankiewicz, K. Codling, L. J. Frasinski, and G. M. Cross, "The multielectron dissociative ionization of molecular iodine in intense laser fields", *Journal of Physics B: Atomic, Molecular and Optical Physics* **27**, 2993 (1999).

[41] A. E. Boguslavskiy, J. Mikosch, A. Gijsbertsen, M. Spanner, S. Patchkovskii, et al., "The Multielectron Ionization Dynamics Underlying Attosecond Strong-Field Spectroscopies", *Science* **335**, 1336 (2012).

[42] L. J. Frasinski, "Covariance mapping techniques", *Journal of Physics B: Atomic, Molecular and Optical Physics* **49**, 152004 (2016).

[43] O. Kornilov, M. Eckstein, M. Rosenblatt, C. P. Schulz, K. Motomura, et al., "Coulomb explosion of diatomic molecules in intense XUV fields mapped by partial covariance", *Journal of Physics B: Atomic, Molecular and Optical Physics* **46**, 164028 (2013).

[44] L. J. Frasinski, V. Zhaunerchyk, M. Mucke, R. J. Squibb, M. Siano, et al., "Dynamics of hollow atom formation in intense X-ray pulses probed by partial covariance mapping", *Physical Review Letters* **111**, 1 (2013).

[45] P. Pfändler, G. Bodenhausen, J. Rapin, R. Houriet, and T. Gäuman, "Two-dimensional Fourier transform ion cyclotron resonance mass spectrometry", *Chemical Physics Letters* **138**, 195 (1987).

[46] C. W. Ross, S. Guan, P. B. Grosshans, T. L. Ricca, A. G. Marshall, et al., "Two-Dimensional Fourier Transform Ion Cyclotron Resonance Mass Spectrometry/Mass Spectrometry with Stored-Waveform Ion Radius Modulation", *Journal of the American Chemical Society* **115**, 7854 (1993).

[47] C. F. Gauss, in *Werke, vol. 4* (Göttingen, Germany, 1823), p. 1.

[48] L. J. Frasinski, K. Codling, P. Hatherly, J. Barr, I. N. Ross, et al., "Femtosecond dynamics of multielectron dissociative ionization by use of a picosecond laser", *Physical Review Letters* **58**, 2424 (1987).

[49] L. J. Frasinski, "On the absence of negative correlations between products of competing processes", *Journal of Physics B: Atomic, Molecular and Optical Physics* **26**, 2225 (1993).

[50] K. Codling, C. Cornaggia, L. J. Frasinski, P. A. Hatherly, J. Morellec, et al., "Charge-symmetric fragmentation of diatomic molecules in intense picosecond laser fields", *Journal of Physics B: Atomic, Molecular and Optical Physics* **24**, L593 (1999).

[51] B. W. J. McNeil and N. R. Thompson, "X-ray free-electron lasers", *Nature Photonics* **4**, 814 (2010).

[52] W. J. Krzanowski, *Principles of Multivariate Analysis, Revised Edition* (Oxford University Press, Oxford, 2000).

[53] J. Mikosch and S. Patchkovskii, "Coincidence and covariance data acquisition in photoelectron and -ion spectroscopy. I. Formal theory", *Journal of Modern Optics* **60**, 1426 (2013).

[54] T. Pekar Second, J. D. Blethrow, J. C. Schwartz, G. E. Merrihew, M. J. MacCoss, et al., "Dual-pressure linear ion trap mass spectrometer improving the analysis of complex protein mixtures", *Analytical Chemistry* **81**, 7757 (2009).

[55] D. J. Douglas, A. J. Frank, and D. Mao, "Linear ion traps in mass spectrometry", *Mass Spectrometry Reviews* **24**, 1 (2005).

[56] J. Thomson, "XIX. Further experiments on positive rays", *Philosophical Magazine Series 6* **24**, 209 (1912).

[57] S. D. Fuerstenau and W. H. Benner, "Molecular weight determination of megadalton DNA electrospray ions using charge detection time-of-flight mass spectrometry.", *Rapid Communications in Mass Spectrometry* **9**, 1528 (1995).

[58] W. J. H. Van Berkel, R. H. H. Van Den Heuvel, C. Versluis, and A. J. R. Heck, "Detection of intact megadalton protein assemblies of vanillyl-alcohol oxidase by mass spectrometry", *Protein Science* **9**, 435 (2000).

[59] P. Kebarle and U. H. Verkerk, "Electrospray: From ions in solution to ions in the gas phase, what we know now", *Mass Spectrometry Reviews* **28**, 898 (2009).

[60] J. Zeleny, "Instability of electrified liquid surfaces", *Physical Review* **10**, 1 (1917).

[61] M. L. Alexandrov, L. N. Gall, N. V. Krasnov, V. I. Nikolaev, V. A. Panvlenko, et al., "Direct coupling of a microcolumn liquid chromatograph and a mass-spectrometer", *Bioorganicheskaia Khimiia* **10**, 710 (1984).

[62] M. Dole, L. L. Mack, R. L. Hines, R. C. Mobley, L. D. Ferguson, et al., "Molecular Beams of Macroions", *The Journal of Chemical Physics* **49**, 2240 (1968).

[63] E. Ahadi and L. Konermann, "Ejection of solvated ions from electrosprayed methanol/water nanodroplets studied by molecular dynamics simulations", *Journal of the American Chemical Society* **133**, 9354 (2011).

[64] G. I. Taylor and A. D. Mcewan, "The stability of a horizontal fluid interface in a vertical electric field", *Journal of Fluid Mechanics* **22**, 1 (1965).

[65] L. Rayleigh, "XX. On the equilibrium of liquid conducting masses charged with electricity", *Philosophical Magazine Series 5* **14**, 184 (1882).

[66] M. Wilm, "Principles of Electrospray Ionization", *Molecular & Cellular Proteomics* **10**, M111.009407 (2011).

[67] M. S. Wilm and M. Mann, "Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last?", *International Journal of Mass Spectrometry and Ion Processes* **136**, 167 (1994).

[68] A. Gomez and K. Tang, "Charge and fission of droplets in electrostatic sprays", *Physics of Fluids* **6**, 404 (1994).

[69] R. L. Grimm and J. L. Beauchamp, "Evaporation and discharge dynamics of highly charged multicomponent droplets generated by electrospray ionization", *Journal of Physical Chemistry A* **114**, 1411 (2010).

[70] L. Konermann, A. D. Rodriguez, and J. Liu, "On the formation of highly charged gaseous ions from unfolded proteins by electrospray ionization", *Analytical Chemistry* **84**, 6798 (2012).

[71] M. Wilm and M. Mann, "Analytical properties of the nanoelectrospray ion source", *Analytical Chemistry* **68**, 1 (1996).

[72] L. Konermann, E. Ahadi, A. D. Rodriguez, and S. Vahidi, "Unraveling the mechanism of electrospray ionization", *Analytical Chemistry* **85**, 2 (2013).

[73] S. Nguyen and J. B. Fenn, "Gas-phase ions of solute species from charged droplets of solutions.", *Proceedings of the National Academy of Sciences of the United States of America* **104**, 1111 (2007).

[74] J. V. Iribarne and B. A. Thomson, "On the evaporation of small ions from charged droplets", *The Journal of Chemical Physics* **64**, 2287 (1976).

[75] E. Ahadi and L. Konermann, "Surface charge of electrosprayed water nanodroplets: A molecular dynamics study", *Journal of the American Chemical Society* **132**, 11270 (2010).

[76] B. E. Winger, K. J. Light-Wahl, R. R. Ogorzalek Loo, H. R. Udseth, and R. D. Smith, "Observation and implications of high mass-to-charge ratio ions from electrospray ionization mass spectrometry", *Journal of the American Society for Mass Spectrometry* **4**, 536 (1993).

[77] E. Ahadi and L. Konermann, "Modeling the behavior of coarse-grained polymer chains in charged water droplets: Implications for the mechanism of electrospray ionization", *Journal of Physical Chemistry B* **116**, 104 (2012).

[78] M. Oss, A. Kruve, K. Herodes, and I. Leito, "Electrospray ionization efficiency scale of organic compound", *Analytical Chemistry* **82**, 2865 (2010).

[79] A. J. Heck and R. H. Van Den Heuvel, "Investigation of intact protein complexes by mass spectrometry", *Mass Spectrometry Reviews* **23**, 368 (2004).

[80] M. Karas, U. Bahr, and T. Dülcks, "Nano-electrospray ionization mass spectrometry: addressing analytical problems beyond routine", *Fresenius' Journal of Analytical Chemistry* **366**, 669 (2000).

[81] J. C. Schwartz, M. W. Senko, and J. E. P. Syka, "A two-dimensional quadrupole ion trap mass spectrometer", *Journal of the American Society for Mass Spectrometry* **13**, 659 (2002).

[82] J. Mitchell Wells and S. A. McLuckey, "Collision-induced dissociation (CID) of peptides and proteins", *Methods in Enzymology* **402**, 148 (2005).

[83] Thermo Fisher Scientific, *LTQ XL Hardware Manual, Revision A (June 2006).*

[84] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman Lectures on Physics, Vol. II: Mainly Electromagnetism and Matter* (Addison-Wesley, Boston, 1964) Chap. 7.2.

[85] R. I. Thompson, T. J. Harmon, and M. G. Ball, "The rotating-saddle trap: a mechanical analogy to RF-electric-quadrupole ion trapping?", *Canadian Journal of Physics* **80**, 1433 (2002).

[86] É. Mathieu, *Course de physique mathématique* (Paris, 1873).

[87] G. Floquet, "Sur les équations différentielles linéaires à coefficients périodiques", *Annales scientifiques de l'É.N.S.* **2**, 47 (1883).

[88] L. Ruby, "Applications of the Mathieu equation", *American Journal of Physics* **64**, 39 (1996).

[89] H. J. W. Müller-Kirsten, in *Introduction to Quantum Mechanics: Schrödinger Equation and Path Integral, 2nd ed.* (2012), pp. 415–471.

[90] P. H. Dawson, "Ion Optical Properties of Quadrupole Mass Filters", *Advances in Electronics and Electron Physics* **53**, 153 (1980).

[91] G. C. Stafford, "Instrumental Aspects of Positive and Negative Ion Chemical Ionization Mass Spectrometry", *Environmental Health Perspectives* **36**, 85 (1980).

[92] Y. L. Chen, B. A. Collings, and D. J. Douglascor, "Collision cross sections of myoglobin and cytochrome c ions with Ne, Ar, and Kr", *Journal of the American Society for Mass Spectrometry* **8**, 681 (1997).

[93] G. Hopfgartner, E. Varesio, V. Tschäppät, C. Grivet, E. Bourgogne, et al., "Triple quadrupole linear ion trap mass spectrometer for the analysis of small molecules and macromolecules", *Journal of Mass Spectrometry* **39**, 845 (2004).

[94] D. J. Douglas, "Applications of collision dynamics in quadrupole mass spectrometry", *Journal of the American Society for Mass Spectrometry* **9**, 101 (1998).

[95] J. S. Brodbelt, "Ion Activation Methods for Peptides and Proteins", *Analytical Chemistry* **88**, 30 (2016).

[96] S. A. McLuckey, "Principles of collisional activation in analytical mass spectrometry", *Journal of the American Society for Mass Spectrometry* **3**, 599 (1992).

[97] J. V. Olsen, B. Macek, O. Lange, A. Makarov, S. Horning, et al., "Higher-energy C-trap dissociation for peptide modification analysis", *Nature Methods* **4**, 709 (2007).

[98] J. N. Louris, R. G. Cooks, J. E. P. Syka, P. E. Kelley, G. C. Stafford, et al., "Instrumentation, applications, and energy deposition in quadrupole ion-trap tandem mass spectrometry", *Analytical Chemistry* **59**, 1677 (1987).

[99] D. E. Goeringer and S. A. McLuckey, "Kinetics of collision-induced dissociation in the paul trap: A first-order model", *Rapid Communications in Mass Spectrometry* **10**, 328 (1996).

[100] D. E. Goeringer, K. G. Asano, and S. A. Mcluckey, "Ion internal temperature and ion trap collisional activation : protonated leucine enkephalin", *Int. J. Mass Spectrom.* **182/183**, 275 (1999).

[101] O. Meroueh and W. L. Hase, "Collisional Activation of Small Peptides", *The Journal of Physical Chemistry A* **103**, 3981 (1999).

[102] J. Laskin and J. H. Futrell, "On the efficiency of energy transfer in collisional activation of small peptides", *Journal of Chemical Physics* **116**, 4302 (2002).

[103] M. W. MacArthur and J. M. Thornton, "Deviations from planarity of the peptide bond in peptides and proteins", *Journal of Molecular Biology* **264**, 1180 (1996).

[104] G. N. Ramachandran and V. Sasisekharan, "Conformation of Polypeptides and Proteins", *Advances in Protein Chemistry* **23**, 283 (1968).

[105] B. Paizs and S. Suhai, "Spectra of Protonated Oligopeptides . 1 : Mechanism of Amide Bond Cleavage", *Journal of the American Society for Mass Spectrometry* **15**, 103 (2004).

[106] M. M. Cordero, J. J. Houser, and C. Wesdemiotis, "The Neutral Products Formed during Backbone Fragmentations of Protonated Peptides in Tandem Mass Spectrometry", *Analytical Chemistry* **65**, 1594 (1993).

[107] K. Xiao, F. Yu, H. Fang, B. Xue, Y. Liu, et al., "Are neutral loss and internal product ions useful for top-down protein identification?", *Journal of Proteomics* **160**, 21 (2017).

[108] K. F. Medzihradszky, "Peptide sequence analysis", *Methods in Enzymology* **402**, 209 (2005).

[109] B. Paizs, Z. Szlávik, G. Lendvay, K. Vékey, and S. Suhai, "Formation of a2+ ions of protonated peptides. An ab initio study.", *Rapid communications in mass spectrometry : RCM* **14**, 746 (2000).

[110] D. R. Mueller, M. Eckersley, and W. J. Richter, "Hydrogen Transfer Reactions in the Formation of "Y + 2" Sequence Ions from Protonated Peptides", *Organic Mass Spectrometry* **23**, 217 (1988).

[111] O. Burlet, C. Y. Yang, and S. J. Gaskell, "Influence of cysteine to cysteic acid oxidation on the collision-activated decomposition of protonated peptides: Evidence for intraionic interactions", *Journal of the American Society for Mass Spectrometry* **3**, 337 (1992).

[112] V. H. Wysocki, G. Tsaprailis, L. L. Smith, and L. A. Breci, "Mobile and localized protons: A framework for understanding peptide dissociation", *Journal of Mass Spectrometry* **35**, 1399 (2000).

[113] A. R. Dongré, J. L. Jones, Á. Somogyi, and V. H. Wysocki, "Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: Evidence for the mobile proton model", *Journal of the American Chemical Society* **118**, 8365 (1996).

[114] A. L. McCormack, A. R. Dongré, V. H. Wysocki, and A. Somogyi, "Fragmentation of Protonated Peptides: Surface-Induced Dissociation in Conjunction with a Quantum Mechanical Approach", *Analytical Chemistry* **65**, 2859 (1993).

[115] Á. Somogyi, V. H. Wysocki, and I. Mayer, "The effect of protonation site on bond strengths in simple peptides: Application of Ab initio and modified neglect of differential overlap bond orders and modified neglect of differential overlap energy partitioning", *Journal of the American Society for Mass Spectrometry* **5**, 704 (1994).

[116] B. Paizs and S. Suhai, "Combined quantum chemical and RRKM modeling of the main fragmentation pathways of protonated GGG. II. Formation of b2, y1, and y2 ions", *Rapid Communications in Mass Spectrometry* **16**, 375 (2002).

[117] B. Paizs, I. P. Csonka, G. Lendvay, and S. Suhai, "Proton mobility in protonated glycylglycine and N-formylglycylglycinamide: A combined quantum chemical and RKKM study", *Rapid Communications in Mass Spectrometry* **15**, 637 (2001).

[118] K. D. Ballard and S. J. Gaskell, "Sequential mass spectrometry applied to the study of the formation of "internal" fragment ions of protonated peptides", *International Journal of Mass Spectrometry and Ion Processes* **111**, 173 (1991).

[119] A. G. Harrison, "Linear Free Energy Correlations in Mass Spectrometry", *Journal of Mass Spectrometry* **34**, 577 (1999).

[120] Student, "The Probable Error of a Mean", *Biometrika* **6**, 1 (1908).

[121] I. A. Papayannopoulos, "The interpretation of collision-induced dissociation tandem mass spectra of peptides", *Mass Spectrometry Reviews* **14**, 49 (1995).

[122] P. J. Boersema, S. Mohammed, and A. J. Heck, "Phosphopeptide fragmentation and analysis by mass spectrometry", *Journal of Mass Spectrometry* **44**, 861 (2009).

[123] A. Bird, "DNA methylation patterns and epigenetic memory", *Genes Dev* **16**, 6 (2002).

[124] C. Papin, A. Ibrahim, S. Le Gras, A. Velt, I. Stoll, et al., "Combinatorial DNA methylation codes at repetitive elements", *Genome Research* **27**, 934 (2017).

[125] R. Holliday, "The Inheritance of Epigenetic Defects", *Science* **238**, 163 (1987).

[126] J. Taunton, C. A. Hassig, and S. L. Schreiber, "A Mammalian Histone Deacetylase Related to the Yeast Transcriptional Regulator Rpd3p", *Science* **272**, 408 (1996).

[127] J. E. Brownell, J. Zhou, T. Ranalli, R. Kobayashi, D. G. Edmondson, et al., "Tetrahymena histone acetyltransferase A: A homolog to yeast Gcn5p linking histone acetylation to gene activation", *Cell* **84**, 843 (1996).

[128] B. D. Strahl and C. D. Allis, "The language of covalent histone modifications", *Nature* **403**, 41 (2000).

[129] T. Jenuwein and C. D. Allis, "Translating the histone code", *Science* **293**, 1074 (2001).

[130] Y. Wei, L. Yu, J. Bowen, M. A. Gorovsky, and C. D. Allis, "Phosphorylation of histone H3 is required for proper chromosome condensation and segregation.", *Cell* **97**, 99 (1999).

[131] M. J. Hendzel, Y. Wei, M. A. Mancini, A. Van Hooser, T. Ranalli, et al., "Mitosis-specific phosphorylation of histone H3 initiates primarily within pericentromeric heterochromatin during G2 and spreads in an ordered fashion coincident with mitotic chromosome condensation", *Chromosoma* **106**, 348 (1997).

[132] M. J. Barratt, C. A. Hazzalin, E. Cano, and L. C. Mahadevan, "Mitogen-stimulated phosphorylation of histone H3 is targeted to a small hyperacetylation-sensitive fraction.", *Proceedings of the National Academy of Sciences of the United States of America* **91**, 4781 (1994).

[133] S. Thomson, L. C. Mahadevan, and A. L. Clayton, "MAP kinase-mediated signalling to nucleosomes and immediate-early gene induction", *Seminars in Cell and Developmental Biology* **10**, 205 (1999).

[134] P. Filippakopoulos, S. Picaud, M. Mangos, T. Keates, J. P. Lambert, et al., "Histone recognition and large-scale structural analysis of the human bromodomain family", *Cell* **149**, 214 (2012).

[135] S. L. Berger, "The complex language of chromatin regulation during transcription", *Nature* **447**, 407 (2007).

[136] A. Zippo, R. Serafini, M. Rocchigiani, S. Pennacchini, A. Krepelova, et al., "Histone Crosstalk between H3S10ph and H4K16ac Generates a Histone Code that Mediates Transcription Elongation", *Cell* **138**, 1122 (2009).

[137] J. Eswaran, D. Patnaik, P. Filippakopoulos, F. Wang, R. L. Stein, et al., "Structure and functional characterization of the atypical human kinase haspin", *Proceedings of the National Academy of Sciences of the United States of America* **106**, 20198 (2009).

[138] Y. Li, L. Sun, Y. Zhang, D. Wang, F. Wang, et al., "The histone modifications governing TFF1 transcription mediated by estrogen receptor", *Journal of Biological Chemistry* **286**, 13925 (2011).

[139] W. S. Lo, R. C. Trievel, J. R. Rojas, L. Duggan, J. Y. Hsu, et al., "Phosphorylation of serine 10 in histone H3 is functionally linked in vitro and in vivo to Gcn5-mediated acetylation at lysine 14", *Molecular Cell* **5**, 917 (2000).

[140] W.-s. Lo, L. Duggan, N. C. T. Emre, and S. L. Berger, "Snf1 — a Histone Kinase That Works in Concert with the Histone Acetyltransferase Gcn5 to Regulate Transcription", *Science* **293**, 1142 (2001).

[141] R. Y. Tweedie-Cullen, A. M. Brunner, J. Grossmann, S. Mohanna, D. Sichau, et al., "Identification of combinatorial patterns of post-translational modifications on individual histones in the mouse brain", *PLoS ONE* **7** (2012).

[142] A. Moradian, A. Kalli, M. J. Sweredoski, and S. Hess, "The top-down, middle-down, and bottom-up mass spectrometry approaches for characterization of histone variants and their post-translational modifications", *Proteomics* **14**, 489 (2014).

[143] L.-M. P. Britton, M. Gonzales-Cope, B. M. Zee, and B. A. Garcia, "Breaking the histone code with quantitative mass spectrometry", *Expert Review of Proteomics* **8**, 631 (2011).

[144] N. Siuti and N. L. Kelleher, "Decoding protein modifications using top-down mass spectrometry", *Nature Methods* **4**, 817 (2007).

[145] X. Dang, J. Scotcher, S. Wu, R. K. Chu, N. Tolić, et al., "The first pilot project of the consortium for top-down proteomics: A status report", *Proteomics* **14**, 1130 (2014).

[146] C. E. Thomas, N. L. Kelleher, and C. A. Mizzen, "Mass spectrometric characterization of human histone H3: a bird's eye view", *J.Proteome Res.* **5**, 240 (2006).

[147] A. P. Lothrop, M. P. Torres, and S. M. Fuchs, "Deciphering post-translational modification codes", *FEBS Letters* **587**, 1247 (2013).

[148] N. L. Young, P. A. DiMaggio, M. D. Plazas-Mayorca, R. C. Baliban, C. A. Floudas, et al., "High Throughput Characterization of Combinatorial Histone Codes", *Molecular & Cellular Proteomics* **8**, 2266 (2009).

[149] B. A. Garcia, J. J. Pesavento, C. A. Mizzen, and N. L. Kelleher, "Pervasive combinatorial modification of histone H3 in human cells", *Nature Methods* **4**, 487 (2007).

[150] R. S. Wu, H. T. Panusz, C. L. Hatch, and W. M. Bonner, "Histones and their modifications", *Critical Reviews in Biochemistry* **20**, 201 (1986).

[151] V. Zhaunerchyk, L. J. Frasinski, J. H. D. Eland, and R. Feifel, "Theory and simulations of covariance mapping in multiple dimensions for data analysis in high-event-rate experiments", *Physical Review A - Atomic, Molecular, and Optical Physics* **89**, 1 (2014).

[152] A. Bateman, M. J. Martin, C. O'Donovan, M. Magrane, E. Alpi, et al., "UniProt: The universal protein knowledgebase", *Nucleic Acids Research* **45**, D158 (2017).

[153] B. J. Cargile, J. L. Bundy, and J. L. Stephenson, "Potential for false positive identifications from large databases through tandem mass spectrometry", *Journal of Proteome Research* **3**, 1082 (2004).

[154] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data", *Electrophoresis* **20**, 3551 (1999).

[155] Y. Chen, J. Zhang, G. Xing, and Y. Zhao, "Mascot-derived false positive peptide identifications revealed by manual analysis of tandem mass spectra", *Journal of Proteome Research* **8**, 3141 (2009).

[156] J. K. Eng, B. C. Searle, K. R. Clauser, and D. L. Tabb, "A Face in the Crowd: Recognizing Peptides Through Database Search", *Molecular & Cellular Proteomics* **10** (2011).

[157] A. Nesvizhskii, O. Vitek, and R. Aebersold, "Analysis and validation of proteomic data generated by tandem mass spectrometry", *Nature Methods* **4**, 787 (2007).

[158] B. Ma and R. Johnson, "De Novo Sequencing and Homology Searching", *Molecular & Cellular Proteomics* **11** (2012).

[159] J. K. Eng, A. L. Mccormack, and J. R. Yates, "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database", *Journal of the American Society for Mass Spectrometry* **5**, 976 (1994).

[160] J. S. Cottrell, "Protein identification using MS/MS data", *Journal of Proteomics* **74**, 1842 (2011).

[161] S. Houel, R. Abernathy, K. Renganathan, K. Meyer-Arendt, N. Ahn, et al., "Quantifying the impact of chimera MS/MS spectra on peptide identification in large scale proteomic studies", *Journal of Proteome Research* **9**, 4152 (2010).

[162] A. M. Palumbo and G. E. Reid, "Evaluation of Gas-Phase Rearrangement and Competing Fragmentation Reactions on Protein Phosphorylation Site Assignment Using Collision Induced Dissociation-MS/MS and MS3", *Analytical Chemistry* **80**, 9735 (2008).

[163] C. Bleiholder, S. Osburn, T. D. Williams, S. Suhai, M. Van Stipdonk, et al., "Sequence-scrambling fragmentation pathways of protonated peptides", *Journal of the American Chemical Society* **130**, 17774 (2008).

[164] A. Nyakas, R. P. Eberle, S. R. Stucki, and S. Schürch, "More than charged base loss - Revisiting the fragmentation of highly charged oligonucleotides", *Journal of the American Society for Mass Spectrometry* **25**, 1155 (2014).

[165] V. S. Rakov, O. V. Borisov, and C. M. Whitehouse, "Establishing low-energy sequential decomposition pathways of leucine enkephalin and its N- and C-terminus fragments using multiple-resonance CID in quadrupolar ion guide", *Journal of the American Society for Mass Spectrometry* **15**, 1794 (2004).

[166] J. Yaqüe, A. Paradela, M. Ramos, S. Ogueta, A. Marina, et al., "Peptide rearrangement during quadrupole ion trap fragmentation: Added complexity to MS/MS spectra", *Analytical Chemistry* **75**, 1524 (2003).

[167] Y. Zhang, B. R. Fonslow, B. Shan, M.-C. Baek, and J. R. Yates 3rd, "Protein Analysis by Shotgun/Bottom-up Proteomics", *Chemical Reviews* **113**, 2343 (2013).

[168] H. R. Morris, T. Paxton, A. Dell, J. Langhorne, M. Berg, et al., "High Sensitivity Collisionally-activated Decomposition Tandem Mass Spectrometry on a Novel Quadrupole/Orthogonal-acceleration Time-of-flight Mass Spectrometer", *Rapid Communications in Mass Spectrometry* **10**, 889 (1996).

[169] M. K. Bunger, B. J. Cargile, A. Ngunjiri, J. L. Bundy, and J. L. Stephenson, "Automated proteomics of E. coli via top-down electron-transfer dissociation mass spectrometry", *Analytical Chemistry* **80**, 1459 (2008).

[170] T.-Y. Kim, M. S. Thompson, and J. P. Reilly, "Peptide photodissociation at 157 nm in a linear ion trap mass spectrometer", *Rapid Communications in Mass Spectrometry* **19**, 1657 (2005).

[171] J. K. Diedrich and R. R. Julian, "Site selective fragmentation of peptides and proteins at quinone modified cysteine residues investigated by ESI-MS", *Analytical Chemistry* **82**, 4006 (2011).

[172] R. Luethy, D. E. Kessner, J. E. Katz, B. MacLean, R. Grothe, et al., "Precursor-ion mass re-estimation improves peptide identification on hybrid instruments", *Journal of Proteome Research* **7**, 4031 (2008).

[173] C. Masselon, G. A. Anderson, R. Harkewicz, J. E. Bruce, L. Pasa-Tolic, et al., "Accurate mass multiplexed tandem mass spectrometry for high-throughput polypeptide identification from mixtures", *Analytical Chemistry* **72**, 1918 (2000).

[174] J. D. Chapman, D. R. Goodlett, and C. D. Masselon, "Multiplexed and data-independent tandem mass spectrometry for global proteome profiling", *Mass Spectrometry Reviews* **33**, 452 (2014).

[175] J. Wilson and R. W. Vachet, "Multiplexed MS/MS in a quadrupole ion trap mass spectrometer", *Analytical Chemistry* **76**, 7346 (2004).

[176] A. M. Graichen and R. W. Vachet, "Multiplexed MS/MS in a miniature rectilinear ion trap", *Journal of the American Society for Mass Spectrometry* **22**, 683 (2011).

[177] P. V. C. Hough, "Machine analysis of bubble chamber pictures", *2nd International Conference on High-Energy Accelerators and Instrumentation* **73**, 554 (1959).

[178] R. O. Duda and P. E. Hart, "Use of the Hough transformation to detect lines and curves in pictures", *Communications of the ACM* **15**, 11 (1972).

[179] D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes", *Pattern Recognition* **13**, 111 (1981).

[180] Matrix Science, *Mascot Top Down AKA "Big Mascot"*, 2011.

[181] Matrix Science, *Mascot Top Down Help*, 2016.

[182] B. Chen, Y. Peng, S. G. Valeja, L. Xiu, A. J. Alpert, et al., "Online Hydrophobic Interaction Chromatography-Mass Spectrometry for Top-Down Proteomics", *Analytical Chemistry* **88**, 1885 (2016).

[183] F. Kryuchkov, T. Verano-Braga, T. A. Hansen, R. R. Sprenger, and F. Kjeldsen, "Deconvolution of mixture spectra and increased throughput of peptide identification by utilization of intensified complementary ions formed in tandem mass spectrometry", *Journal of Proteome Research* **12**, 3362 (2013).

[184] A. J. R. Heck, L. J. de Koning, F. A. Pinkse, and N. M. M. Nibbering, "Mass-specific selection of ions in Fourier-transform ion cyclotron resonance mass spectrometry. Unintentional off-resonance cyclotron excitation of selected ions", *Rapid Communications in Mass Spectrometry* **5**, 406 (1991).

[185] J. W. Gauthier, T. R. Trautman, and D. B. Jacobson, "Sustained off-resonance irradiation for collision-activated dissociation involving Fourier transform mass spectrometry. Collision-activated dissociation technique that emulates infrared multiphoton dissociation", *Analytica Chimica Acta* **246**, 211 (1991).

[186] A. Michalski, J. Cox, and M. Mann, "More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS", *Journal of Proteome Research* **10**, 1785 (2011).

[187] N. Zhang, X. J. Li, M. Ye, S. Pan, B. Schwikowski, et al., "ProbIDtree: An automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer", *Proteomics* **5**, 4096 (2005).

[188] M. R. Hoopmann, G. L. Finney, and M. J. MacCoss, "High-speed data reduction, feature detection and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass Spectrometry", *Analytical Chemistry* **79**, 5620 (2007).

[189] A. T. Kong, F. V. Leprevost, D. M. Avtonomov, D. Mellacheruvu, and A. I. Nesvizhskii, "MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics", *Nature Methods* **14**, 513 (2017).

[190] R. S. Johnson, S. A. Martin, K. Biemann, J. T. Stults, and J. T. Watson, "Novel Fragmentation Process of Peptides by Collision-Induced Decomposition in a Tandem Mass Spectrometer: Differentiation of Leucine and Isoleucine", *Analytical Chemistry* **59**, 2621 (1987).

[191] M. Edelson-Averbukh, R. Pipkorn, and W. D. Lehmann, "Analysis of protein phosphorylation in the regions of consecutive serine/threonine residues by negative ion electrospray collision-induced dissociation. Approach to pinpointing of phosphorylation sites", *Analytical Chemistry* **79**, 3476 (2007).

[192] J. H. Bowie, C. S. Brinkworth, and S. Dua, "Collision-induced fragmentations of the (M-H)-parent anions of underivatized peptides: An aid to structure determination and some unusual negative ion cleavages", *Mass Spectrometry Reviews* **21**, 87 (2002).

[193] M. Edelson-Averbukh, A. Shevchenko, R. Pipkorn, and W. D. Lehmann, "Discrimination between peptide O-sulfo-and O-phosphotyrosine residues by negative ion mode electrospray tandem mass spectrometry", *Journal of the American Society for Mass Spectrometry* **22**, 2256 (2011).

[194] N. P. Ewing and C. J. Cassady, "Dissociation of Multiply Charged Negative Ions for Hirudin (54–65), Fibrinopeptide B, and Insulin A (Oxidized)", *Journal of the American Society for Mass Spectrometry* **12**, 105 (2000).

[195] J. Jai-nhuknan and C. J. Cassady, "Negative ion postsource decay time-of-flight mass spectrometry of peptides containing acidic amino acid residues", *Analytical Chemistry* **70**, 5122 (1998).

[196] S. T. Steinborner and J. H. Bowie, "A comparison of the positive- and negative-ion mass spectra of bio- active peptides from the dorsal secretion of the Australian red tree frog, Litoria rubella", *Rapid Communications in Mass Spectrometry* **10**, 1243 (1996).

[197] N. M. Riley, M. J. P. Rush, C. M. Rose, A. L. Richards, N. W. Kwiecien, et al., "The Negative Mode Proteome with Activated Ion Negative Electron Transfer Dissociation (AI-NETD)", *Molecular & Cellular Proteomics* **14**, 2644 (2015).

[198] J. Shendure, S. Balasubramanian, G. M. Church, W. Gilbert, J. Rogers, et al., "DNA sequencing at 40: Past, present and future", *Nature* **550**, 345 (2017).

[199] F. Ozsolak and P. M. Milos, "RNA sequencing: advances, challenges and opportunities", *Nature Reviews Genetics* **12**, 87 (2011).

[200] J. C. Venter, M. D. M. Adams, E. E. W. Myers, P. W. Li, R. J. Mural, et al., "The sequence of the human genome", *Science* **291**, 1304 (2001).

[201] E. Dudley and L. Bond, "Mass spectrometry analysis of nucleosides and nucleotides", *Mass Spectrometry Reviews* **33**, 302 (2014).

[202] K. D. Robertson, "DNA methylation and human disease", *Nature Reviews Genetics* **6**, 597 (2005).

[203] B. M. Turner, "Defining an epigenetic code", *Nature Cell Biology* **9**, 2 (2007).

[204] J. A. Dunbar, D. G. Osborne, J. M. Anna, and K. J. Kubarych, "Accelerated 2D-IR using compressed sensing", *Journal of Physical Chemistry Letters* **4**, 2489 (2013).

[205] P. Schanda, "Fast-pulsing longitudinal relaxation optimized techniques: Enriching the toolbox of fast biomolecular NMR spectroscopy", *Progress in Nuclear Magnetic Resonance Spectroscopy* **55**, 238 (2009).

[206] F. Bray, J. Bouclon, L. Chiron, M. Witt, M. A. Delsuc, et al., "Nonuniform sampling acquisition of two-dimensional Fourier transform ion cyclotron resonance mass spectrometry for increased mass resolution of tandem mass spectrometry precursor ions", *Analytical Chemistry* **89**, 8589 (2017).

[207] Thermo Fisher Scientific, *Orbitrap Fusion Lumos Tribrid Mass Spectrometer Specifications Sheet*, 2017.

[208] R. E. March, "An introduction to quadrupole ion trap mass spectrometry", *Journal of Mass Spectrometry* **32**, 351 (1997).

[209] C. L. Kalcic, T. C. Gunaratne, A. D. Jones, M. Dantus, and G. E. Reid, "Femtosecond laser-induced ionization/dissociation of protonated peptides", *Journal of the American Chemical Society* **131**, 940 (2009).

[210] C. L. Kalcic, G. E. Reid, V. V. Lozovoy, and M. Dantus, "Mechanism elucidation for nonstochastic femtosecond laser-induced ionization / dissociation: from amino acids to peptides", *The Journal of Physical Chemistry A* **116**, 2764 (2012).

[211] S. A. Smith, C. L. Kalcic, K. A. Safran, P. M. Stemmer, M. Dantus, et al., "Enhanced characterization of singly protonated phosphopeptide ions by femtosecond laser-induced ionization/dissociation tandem mass spectrometry (fs-LID-MS/MS)", *Journal of the American Society for Mass Spectrometry* **21**, 2031 (2010).

[212] S. A. Smith, C. L. Kalcic, L. Cui, and G. E. Reid, "Femtosecond laser-induced ionization/dissociation tandem mass spectrometry (fsLID-MS/MS) of deprotonated phosphopeptide anions", *Rapid Communications in Mass Spectrometry* **27**, 2807 (2013).

[213] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: a comprehensive review", *Neural Computation* **29**, 2352 (2017).

[214] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, et al., "TensorFlow: large-scale machine learning on heterogeneous distributed systems", *arXiv:1603.04467* (2016).

[215] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization", *arXiv:1412.6980* (2014).

[216] S. Woutersen, Y. Mu, G. Stock, and P. Hamm, "Hydrogen-bond lifetime measured by time-resolved 2D-IR spectroscopy: N-methylacetamide in methanol", *Chemical Physics* **266**, 137 (2001).

[217] N. Mischerikow, A. F. M. Altelaar, J. D. Navarro, S. Mohammed, and A. J. R. Heck, "Comparative assessment of site assignments in CID and Electron Transfer Dissociation spectra of phosphopeptides discloses limited relocation of phosphate groups", *Molecular & Cellular Proteomics* **9**, 2140 (2010).

[218] S. Baker, J. S. Robinson, C. A. Haworth, H. Teng, R. A. Smith, et al., "Probing Proton Dynamics in Molecules on an Attosecond Time Scale", *Science* **312**, 424 (2006).

[219] F. McGrath, A. S. Johnson, D. R. Austin, P. Hawkins, D. Wood, et al., "An apparatus for quantitative high-harmonic generation spectroscopy in molecular vapours", *Review of Scientific Instruments* **88** (2017).

[220] E. R. Simpson, A. Sanchez-Gonzalez, D. R. Austin, Z. Diveki, S. E. Hutchinson, et al., "Polarisation response of delay dependent absorption modulation in strong field dressed helium atoms probed near threshold", *New Journal of Physics* **18** (2016).

[221] F. Calegari, D. Ayuso, A. Trabattoni, L. Belshaw, S. De Camillis, et al., "Ultrafast electron dynamics in phenylalanine initiated by attosecond pulses", *Science* **346**, 336 (2014).

[222] L. S. Cederbaum and J. Zobeley, "Ultrafast charge migration by electron correlation", *Chemical Physics Letters* **307**, 205 (1999).

[223] D. Mendive-Tapia, M. Vacher, M. J. Bearpark, and M. A. Robb, "Coupled electron-nuclear dynamics: Charge migration and charge transfer initiated near a conical intersection", *Journal of Chemical Physics* **139** (2013).

[224] M. Vacher, L. Steinberg, A. J. Jenkins, M. J. Bearpark, and M. A. Robb, "Electron dynamics following photoionization: Decoherence due to the nuclear-wavepacket width", *Physical Review A - Atomic, Molecular, and Optical Physics* **92**, 1 (2015).

[225] H. Koizumi, K. Hironaka, K. Shinsaka, S. Arai, H. Nakazawa, et al., "VUV-optical oscillator strength distributions of C2H6O and C3H8O isomers", *The Journal of Chemical Physics* **85**, 4276 (1986).

[226] M. Lewenstein, P. Balcou, M. Y. Ivanov, A. L'Huillier, and P. B. Corkum, "Theory of high-harmonic generation by low-frequency laser fields", *Physical Review A* **49**, 2117 (1994).

[227] D. Fabris, T. Witting, W. A. Okell, D. J. Walke, P. Matia-Hernando, et al., "Synchronized pulses generated at 20 eV and 90 eV for attosecond pump–probe experiments", *Nature Photonics* **9**, 383 (2015).

[228] M. Lezius, V. Blanchet, M. Y. Ivanov, and A. Stolow, "Polyatomic molecules in strong laser fields: Nonadiabatic multielectron dynamics", *Journal of Chemical Physics* **117**, 1575 (2002).

[229] K. Le Guen, D. Céolin, R. Guillemin, C. Miron, N. Leclercq, et al., "Development of a four-element conical electron lens dedicated to high resolution Auger electron-ion(s) coincidence experiments", *Review of Scientific Instruments* **73**, 3885 (2002).

[230] X. J. Liu, C. Nicolas, E. Robert, and C. Miron, "EPICEA: Probing High-Energy Electron Emission in the Molecular Frame", *Journal of Physics: Conference Series* **488**, 13 (2014).

[231] C. Miron, M. Simon, N. Leclercq, and P. Morin, "New high luminosity "double toroidal" electron spectrometer", *Review of Scientific Instruments* **68**, 3728 (1997).

[232] X. J. Liu, C. Nicolas, and C. Miron, "Design of a lens table for a double toroidal electron spectrometer", *Review of Scientific Instruments* **84** (2013).

[233] G. V. Marr and J. B. West, "Absolute photoionization cross-section tables for helium, neon, argon, and krypton in the VUV spectral regions", *Atomic Data and Nuclear Data Tables* **18**, 497 (1976).

[234] B. Cooper, P. Kolorenč, L. J. Frasinski, V. Averbukh, and J. P. Marangos, "Analysis of a measurement scheme for ultrafast hole dynamics by few femtosecond resolution X-ray pump-probe Auger spectroscopy", *Faraday Discussions* **171**, 93 (2014).

[235] A. Sanchez-Gonzalez, T. R. Barillot, R. J. Squibb, P. Kolorenč, M. Agaker, et al., "Auger electron and photoabsorption spectra of glycine in the vicinity of the oxygen K-edge measured with an X-FEL", *Journal of Physics B: Atomic, Molecular and Optical Physics* **48**, 234004 (2015).

[236] D. Nordfors, A. Nilsson, N. Mårtensson, S. Svensson, U. Gelius, et al., "X-ray excited photoelectron spectra of free molecules containing oxygen", *Journal of Electron Spectroscopy and Related Phenomena* **56**, 117 (1991).

[237] P. Emma, R. Akre, J. Arthur, R. Bionta, C. Bostedt, et al., "First lasing and operation of an ångstrom-wavelength free-electron laser", *Nature Photonics* **4**, 641 (2010).

[238] J. Feldhaus, J. Arthur, and J. B. Hastings, "X-ray free-electron lasers", *Journal of Physics B: Atomic, Molecular and Optical Physics* **38**, S799 (2005).

[239] A. A. Lutman, T. J. Maxwell, J. P. MacArthur, M. W. Guetg, N. Berrah, et al., "Fresh-slice multicolour X-ray free-electron lasers", *Nature Photonics* **10**, 745 (2016).

[240] H.-D. Nuhn, S. Marks, and J. Wu, *LCLS-II Undulator Tolerance Analysis SLAC-109 PUB-15062*, tech. rep. (2012).

[241] Z. Zhang, K. Bane, Y. Ding, Z. Huang, R. Iverson, et al., "Electron beam energy chirp control with a rectangular corrugated structure at the Linac Coherent Light Source", *Physical Review Special Topics - Accelerators and Beams* **18**, 1 (2015).

[242] O. Kugeler, S. Marburger, and U. Hergenhahn, "Calculation and measurement of the time-of-flight spread in a hemispherical electron energy analyzer", *Review of Scientific Instruments* **74**, 3955 (2003).

[243] S. P. Hau-Riege, R. M. Bionta, D. D. Ryutov, and J. Krzywinski, "Measurement of x-ray free-electron-laser pulse energies by photoluminescence in nitrogen gas", *Journal of Applied Physics* **103** (2008).

[244] L. J. Frasinski, P. A. Hatherly, and K. Codling, "Multiphoton multiple ionisation of N2O probed by three-dimensional covariance mapping", *Physics Letters A* **156**, 227 (1991).

[245] B. Godugu, P. Neta, Y. Simón-Manso, and S. E. Stein, "Effect of N-terminal glutamic acid and glutamine on fragmentation of peptide ions", *Journal of the American Society for Mass Spectrometry* **21**, 1169 (2010).

Appendices

Appendix A

Selected annotated partial covariance maps for species in Table 7.1

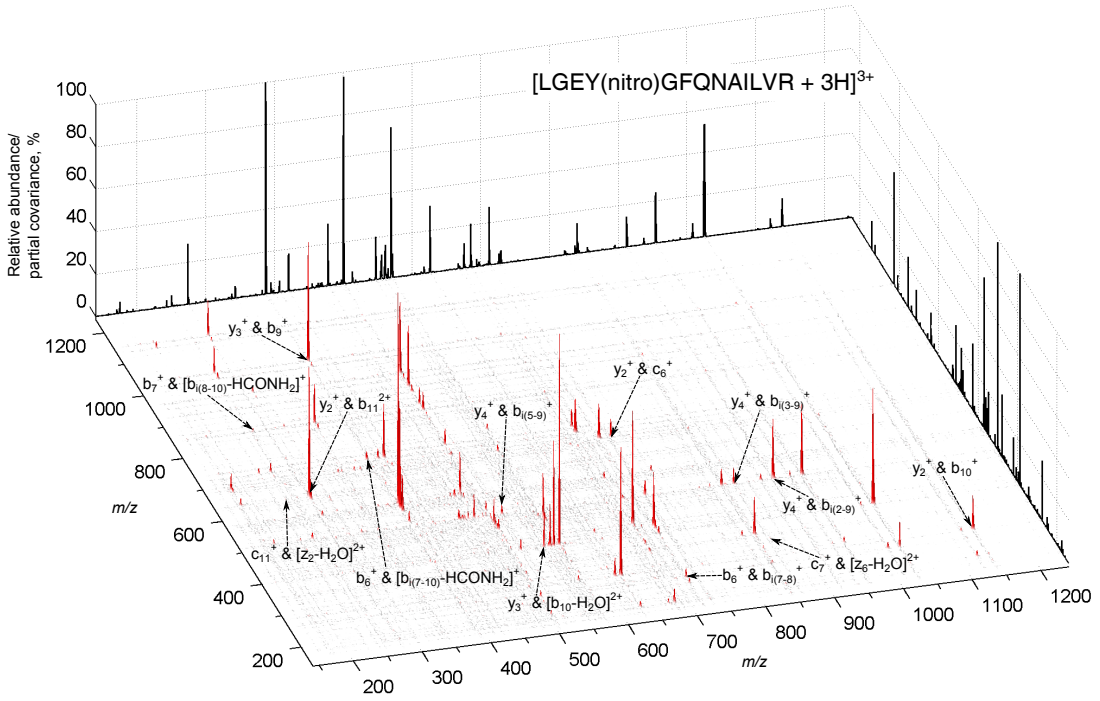


Figure A.1: pC-2DMS map of the triply protonated nitrated peptide LGEY(nitro)GFQNAILVR, Y(nitro)=3-nitrotyrosine, measured at normalised collision energy 35% and AGC MS^n target value 100. The m/z values of the correlating fragments are plotted along the x- and y-axes and the surface represents the partial covariance, calculated using Eq. 3.1, normalised to the highest off-diagonal pC-2DMS peak. The autocorrelation line, which trivially correlates each spectral signal to itself along the diagonal ($x = y$), has been removed and a standard 8×8 median filtering procedure has been applied to the map for visual clarity. The black line graph plotted against the back walls of the pC-2DMS map is the averaged 1D MS/MS mass spectrum. A non-exhaustive selection of top scoring pC-2DMS peaks is labelled, including very weak signals corresponding to the true (intrinsic) fragment ion correlations identified using the pC-2DMS correlation score as described in section 6.2.

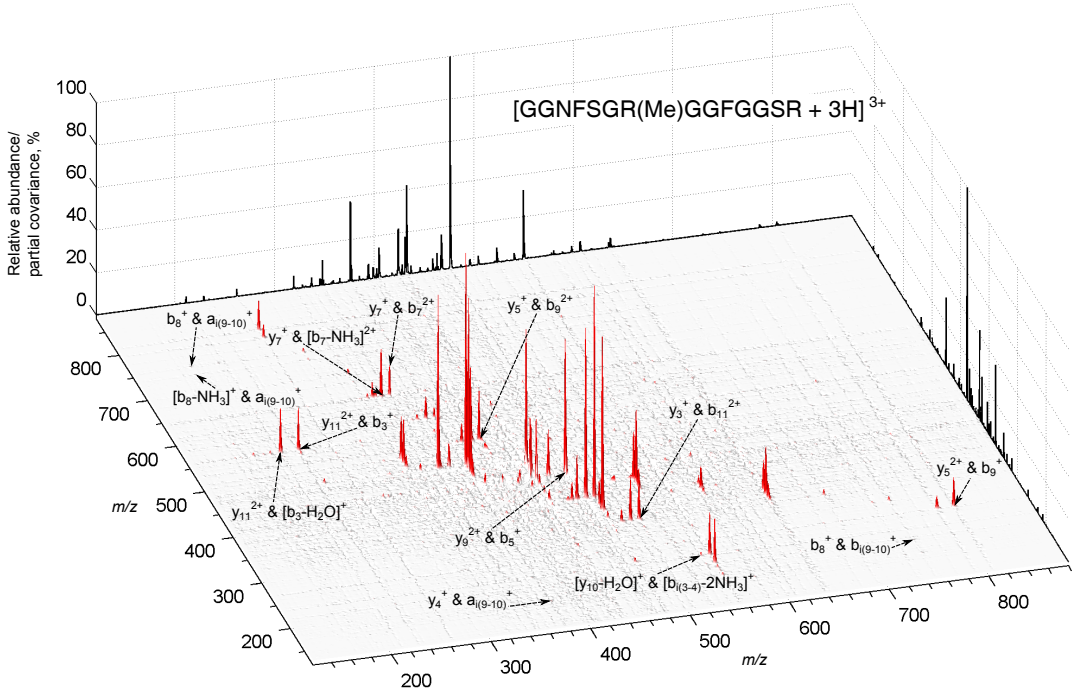


Figure A.2: pC-2DMS map of the triply protonated methylated peptide GGNFSGR(Me)GGFGGSR, R(Me)=N^G-monomethylarginine. All other details are the same as for Fig. A.1

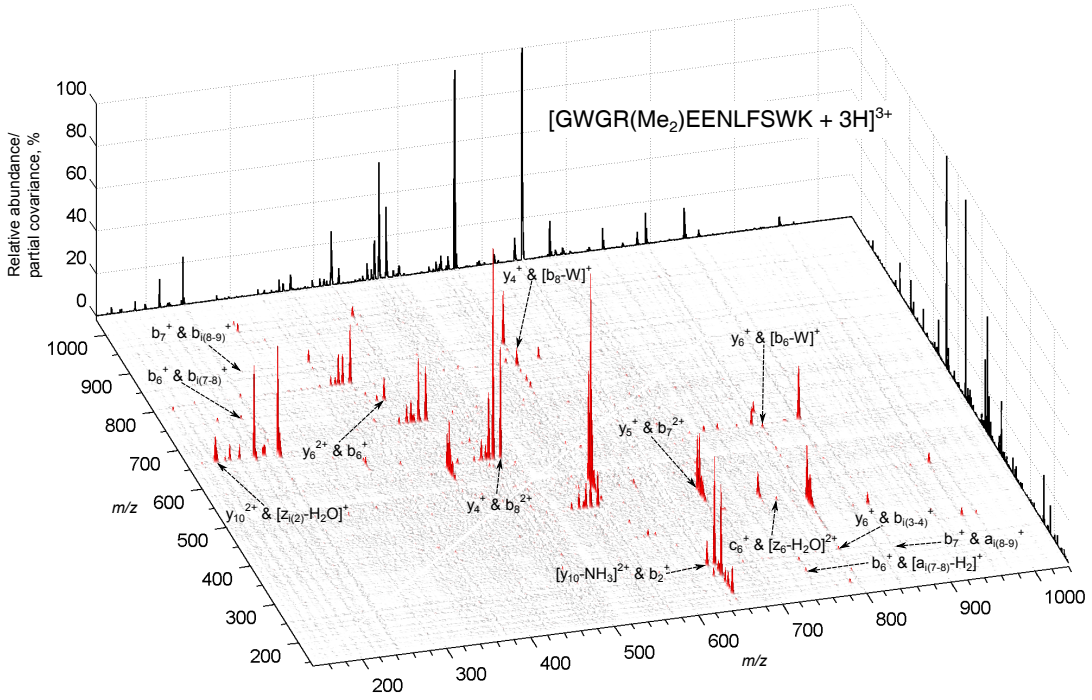


Figure A.3: pC-2DMS map of the triply protonated methylated peptide GWGR(Me$_2$)EENLFSWK, R(Me$_2$)=N$^{\mathrm{G}}$,N$^{\mathrm{G}}$-dimethylarginine. All other details are the same as for Fig. A.1

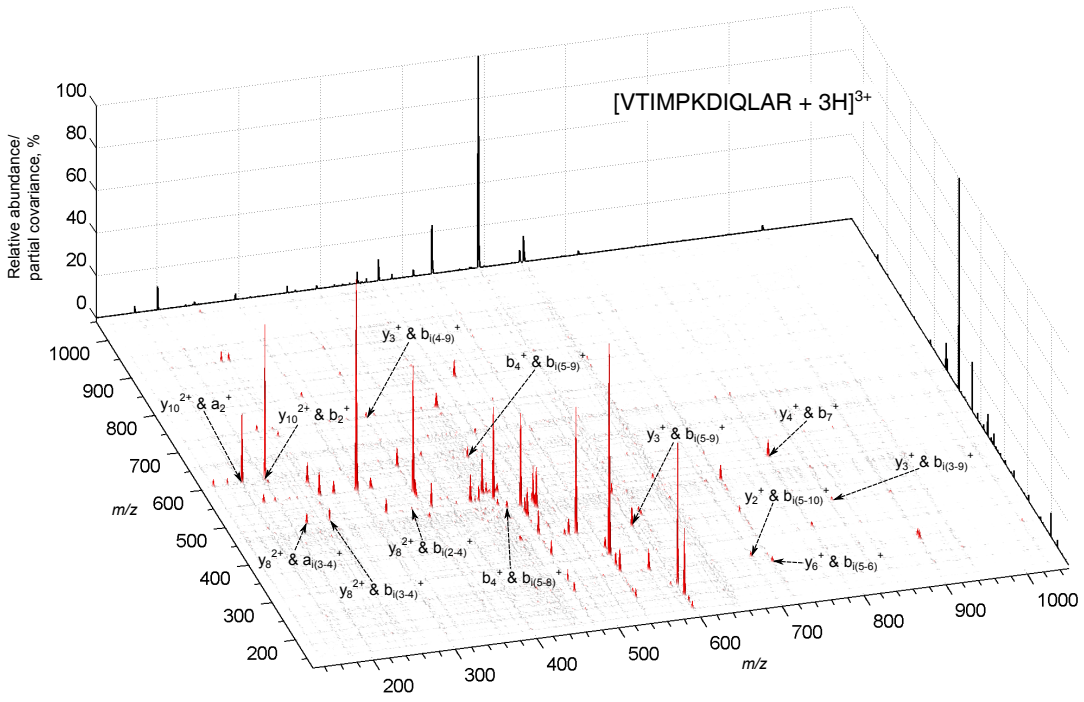


Figure A.4: pC-2DMS map of the triply protonated unmodified peptide VTIMPKDIQLAR. All other details are the same as for Fig. A.1

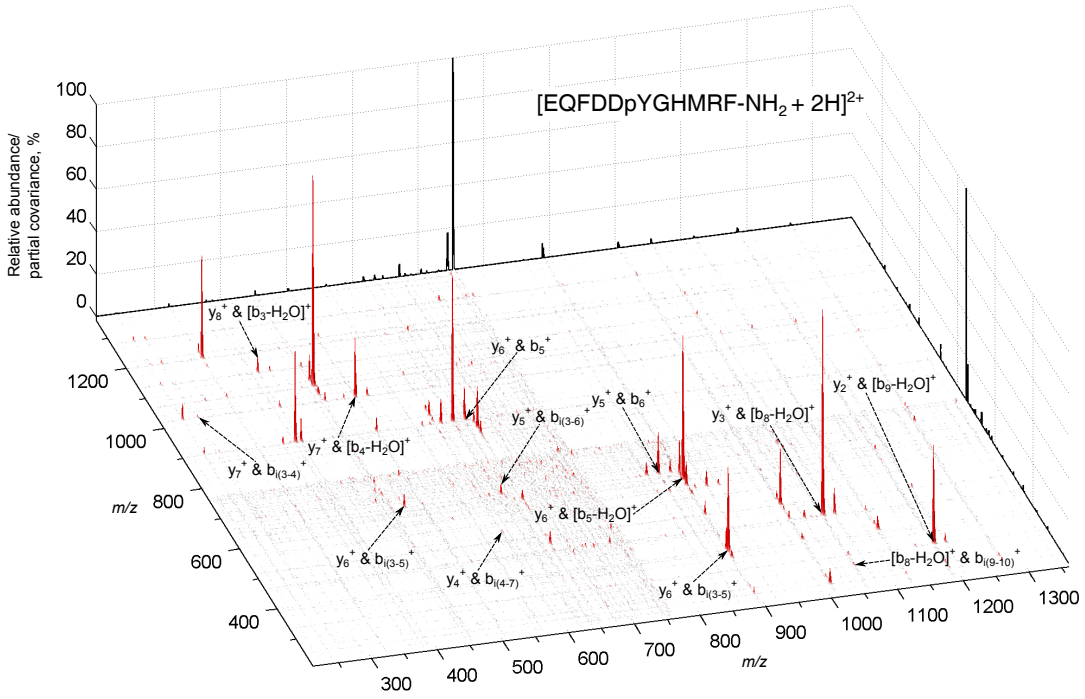


Figure A.5: pC-2DMS map of the doubly protonated phosphorylated peptide EQFDDpYGHMRF-NH_2, pY=phosphotyrosine. All other details are the same as for Fig. A.1. Note that for this peptide ion, the 1D averaged MS/MS spectrum is dominated by the structurally uninformative neutral loss of water, which is facilitated by the N-terminal glutamic acid residue [245]. In contrast, it is a general feature of pC-2DMS maps (see also e.g. Fig. 6.4) that such uninformative neutral losses, which can also be caused by the presence of a labile post-translational modification, are not permitted to dominate the 2D spectrum. This is because, unlike for the majority of reactions producing informative sequence ions, neutral loss fragmentations do not produce two detectable ions to correlate. This allows pC-2DMS maps to reveal structure-informative information even in cases where the 1D spectrum is unable to do so as a result of dominant neutral loss from the parent molecule.

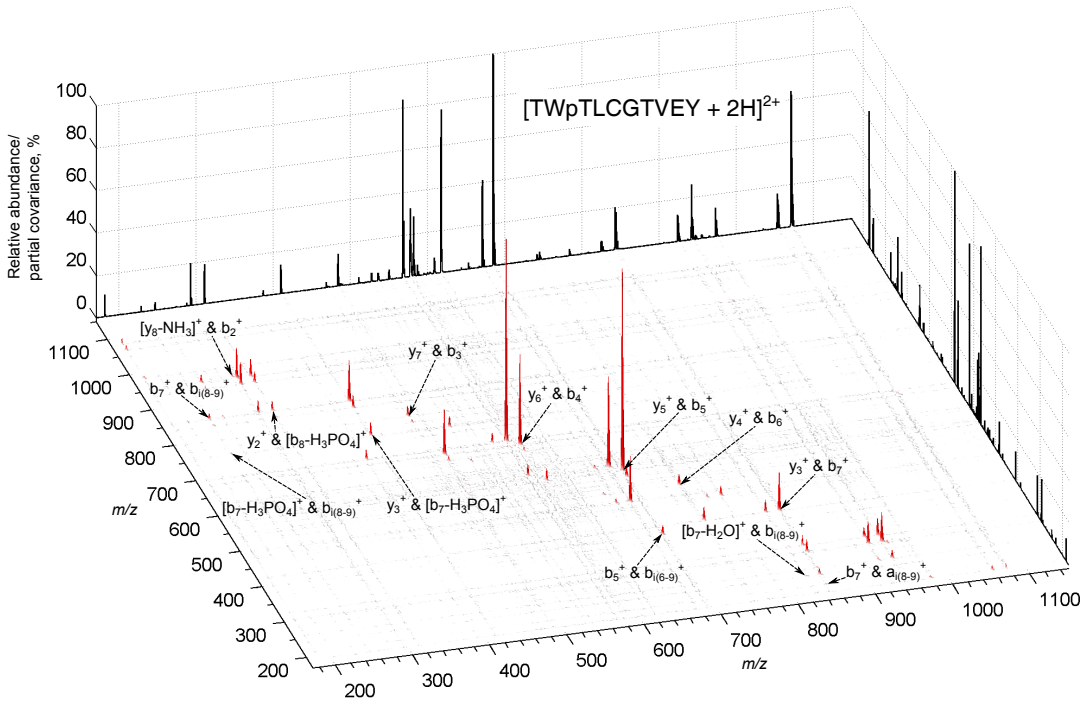


Figure A.6: pC-2DMS map of the doubly protonated phosphorylated peptide TWpTLCGTVEY, pT=phosphothreonine. All other details are the same as for Fig. A.1

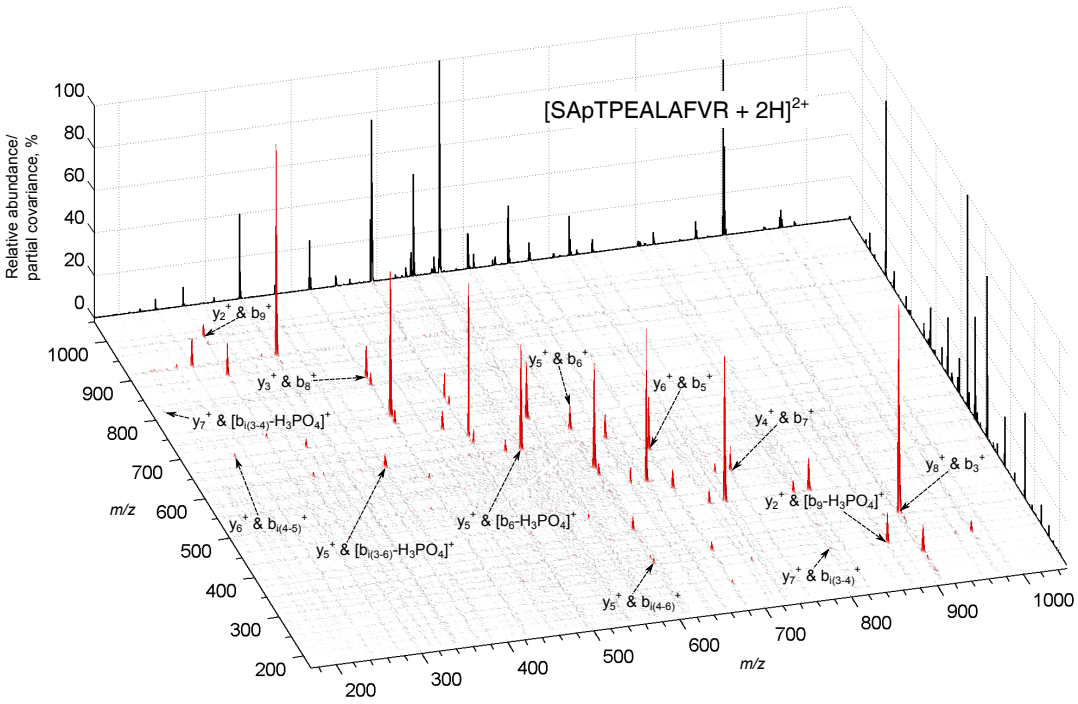


Figure A.7: pC-2DMS map of the doubly protonated phosphorylated peptide SApT-PEALAFVR, pT=phosphothreonine. All other details are the same as for Fig. A.1

Appendix B

Selected MS^3 scans

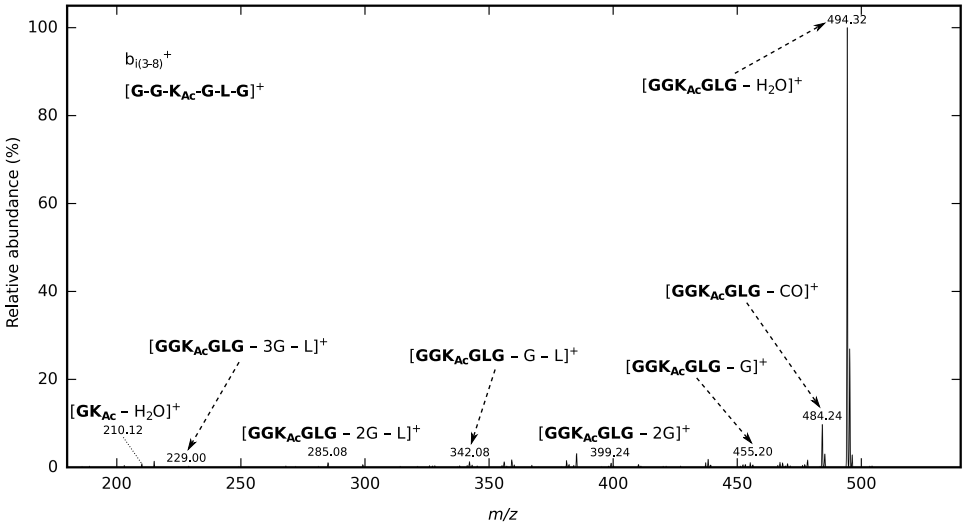


Figure B.1: MS/MS scan at normalised collision energy 35% on the m/z 512.28 fragment of the triply protonated peptide GKGGK$_{\mathrm{Ac}}$GLGKGGAK$_{\mathrm{Ac}}$R (K$_{5,Ac}$K$_{16,Ac}$ isomer of histone H4 fragment 4–17), confirming its identity as the internal b-ion $\mathrm{b}^{+}_{i(3-8)}$. See Chapter 9.

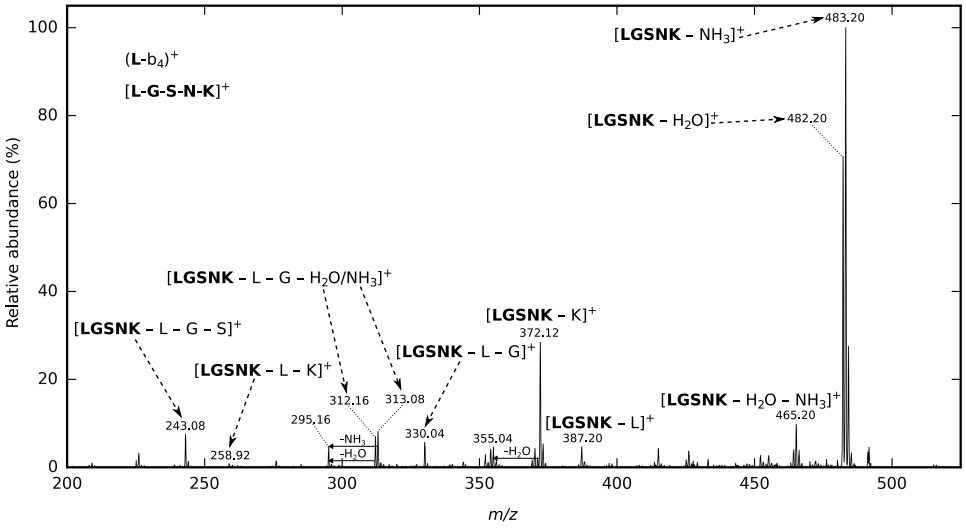


Figure B.2: MS/MS scan at normalised collision energy 30% on the m/z 500.28 fragment of the doubly protonated peptide GSNKGAIIGLM, confirming its identity as the sequence scrambling product (L–b_4^{+}) (L=leucine). See Chapter 12.

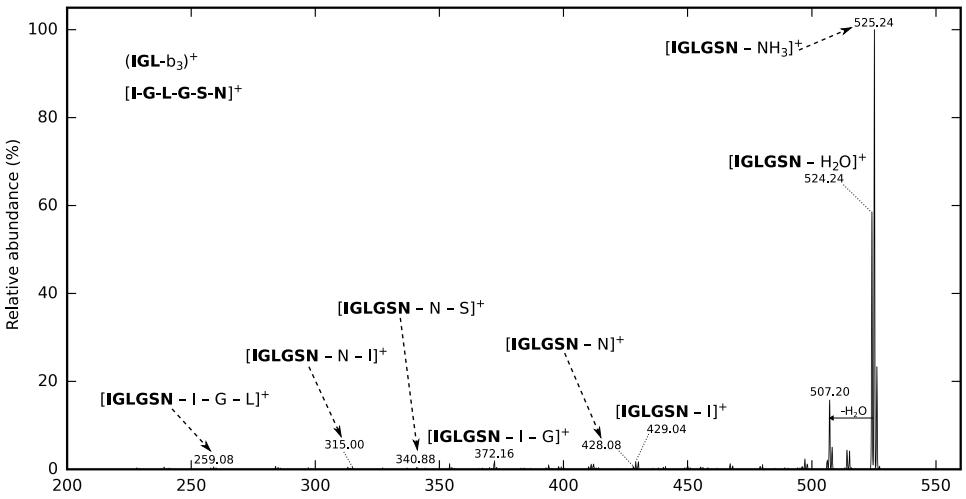


Figure B.3: MS/MS scan at normalised collision energy 35% on the m/z 542.29 fragment of the doubly protonated peptide GSNKGAIIGLM, confirming its identity as the sequence scrambling product (IGL–b_3^+) (I=isoleucine, G=glycine, L=leucine). See Chapter 12.

Appendix C

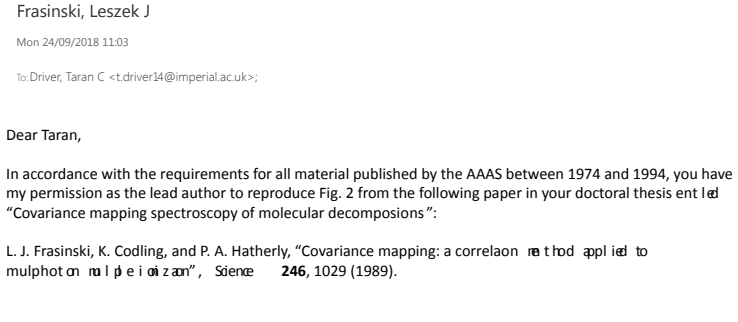
Permissions to reproduce figures

Figure 3.2

Reproduced under the terms of the Creative Commons Attribution 3.0 licence:

https://creativecommons.org/licenses/by/3.0/

Figure 2.1



Frasinski, Leszek J

Mon 24/09/2018 11:03

To: Driver, Taran C <t.driver14@imperial.ac.uk>;

Dear Taran,

In accordance with the requirements for all material published by the AAAS between 1974 and 1994, you have my permission as the lead author to reproduce Fig. 2 from the following paper in your doctoral thesis entitled "Covariance mapping spectroscopy of molecular decomposions":

L. J. Frasinski, K. Codling, and P. A. Hatherly, "Covariance mapping: a correlaon method applied to mulphoon mulple ionizaon", Science **246**, 1029 (1989).

Figure 2.1

THE AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE LICENSE TERMS AND CONDITIONS

Sep 25, 2018

This Agreement between Mr. Taran Driver ("You") and The American Association for the Advancement of Science ("The American Association for the Advancement of Science") consists of your license details and the terms and conditions provided by The American Association for the Advancement of Science and Copyright Clearance Center.

License Number	4434790152817
License date	Sep 23, 2018
Licensed Content Publisher	The American Association for the Advancement of Science
Licensed Content Publication	Science
Licensed Content Title	Covariance Mapping: A Correlation Method Applied to Multiphoton Multiple Ionization
Licensed Content Author	L. J. FRASINSKI,K. CODLING,P. A. HATHERLY
Licensed Content Date	Nov 24, 1989
Licensed Content Volume	246
Licensed Content Issue	4933
Volume number	246
Issue number	4933
Type of Use	Thesis / Dissertation
Requestor type	Scientist/individual at a research institution
Format	Print and electronic
Portion	Figure
Number of figures/tables	1
Order reference number	
Title of your thesis / dissertation	Covariance mapping spectroscopy of molecular decompositions
Expected completion date	Oct 2018
Estimated size(pages)	1
Requestor Location	Mr. Taran Driver 38 Walsingham Road St. Andrews Bristol, BS6 5BT United Kingdom Attn: Mr. Taran Driver
Billing Type	Invoice
Billing Address	Mr. Taran Driver 38 Walsingham Road St. Andrews Bristol, United Kingdom BS6 5BT Attn: Mr. Taran Driver
Total	0.00 USD

Figure 5.1

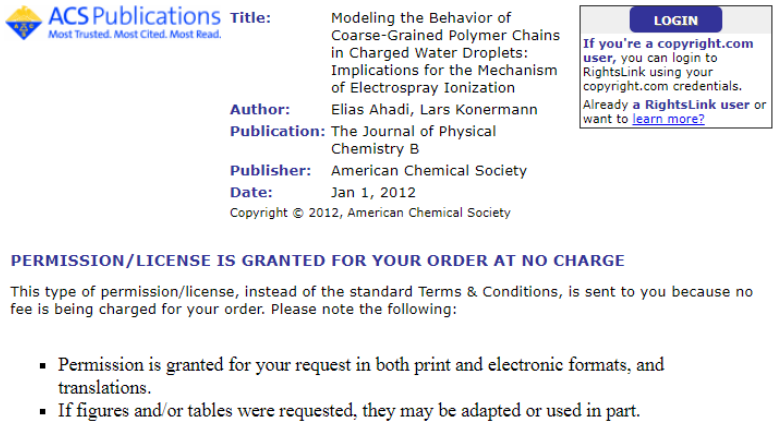
JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Sep 14, 2018

This Agreement between Mr. Taran Driver ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4423671223430
License date	Sep 07, 2018
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Mass Spectrometry Reviews
Licensed Content Title	Electrospray: From ions in solution to ions in the gas phase, what we know now
Licensed Content Author	Paul Kebarle, Udo H. Verkerk
Licensed Content Date	Jun 23, 2009
Licensed Content Volume	28
Licensed Content Issue	6
Licensed Content Pages	20
Type of use	Dissertation/Thesis
Requestor type	University/Academic
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	1
Original Wiley figure/table number(s)	Figure 1
Will you be translating?	No
Title of your thesis / dissertation	Covariance mapping spectroscopy of molecular decompositions
Expected completion date	Oct 2018
Expected size (number of pages)	1
Requestor Location	Mr. Taran Driver 38 Walsingham Road St. Andrews Bristol, BS6 5BT United Kingdom Attn: Mr. Taran Driver
Publisher Tax ID	EU826007151
Total	0.00 USD

Figure 5.2



ACS Publications
Most Trusted. Most Cited. Most Read.

Title: Modeling the Behavior of Coarse-Grained Polymer Chains in Charged Water Droplets: Implications for the Mechanism of Electrospray Ionization

Author: Elias Ahadi, Lars Konermann

Publication: The Journal of Physical Chemistry B

Publisher: American Chemical Society

Date: Jan 1, 2012

Copyright © 2012, American Chemical Society

LOGIN

If you're a copyright.com user, you can login to RightsLink using your copyright.com credentials.

Already a **RightsLink user** or want to learn more?

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.

Figures 5.3 & 5.7

Permission to reprint figures from LTQ XL manual in PhD thesis, Imperial College London

Driver, Taran C

Tue 25/09/2018 13:40

Sent Items

To: unity.uk@thermofisher.com <unity.uk@thermofisher.com>;

Dear Sir/Madam,

I am currently completing my PhD thesis at Imperial College London.

I am writing to seek your permission to reprint in my thesis two figures from the LTQ XL Hardware Manual (Revision A, printed June 2006). These two figures are:

- Figure 16. Internal (under vacuum) components of the MS detector

- Figure 19. Cross sectional view of the ion detection system showing the electron multiplier and the conversion dynode

With your permission, I would like to include these figures in my thesis. The thesis will be added to Spiral, Imperial's institutional repository (http://spiral.imperial.ac.uk/), and made available to the public, for non-commercial purposes, under a Creative Commons Attribution-NonCommercial-NoDerivs licence. Of course, the figures will be fully credited.

If you are happy to grant me all the permissions requested, I would greatly appreciate it if you would please let me know by reply to this email address.

Thank you very much.

Yours faithfully,

Taran Driver

Figure 5.4

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Sep 14, 2018

This Agreement between Mr. Taran Driver ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4418890943012
License date	Aug 30, 2018
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Journal of The American Society for Mass Spectrometry
Licensed Content Title	A two-dimensional quadrupole ion trap mass spectrometer
Licensed Content Author	Jae C. Schwartz, Michael W. Senko, John E. P. Syka
Licensed Content Date	Jan 1, 2002
Licensed Content Volume	13
Licensed Content Issue	6
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Will you be translating?	no
Circulation/distribution	<501
Author of this Springer Nature content	no
Title	Covariance mapping spectroscopy of molecular decompositions
Instructor name	Leszek Frasinski
Institution name	Imperial College London
Expected presentation date	Oct 2018
Portions	Figure 1
Requestor Location	Mr. Taran Driver 38 Walsingham Road St. Andrews Bristol, BS6 5BT United Kingdom Attn: Mr. Taran Driver
Billing Type	Invoice
Billing Address	Mr. Taran Driver 38 Walsingham Road St. Andrews Bristol, United Kingdom BS6 5BT Attn: Mr. Taran Driver
Total	0.00 USD

Figure 5.5

Order Details

Canadian journal of physics

Order detail ID: 71575334
Order License Id: 4437730005933
ISSN: 0008-4204
Publication Type: Journal
Volume:
Issue:
Start page:
Publisher: N R C RESEARCH PRESS

Permission Status: Granted
Permission type: Republish or display content
Type of use: Republish in a thesis/dissertation

Hide details

Requestor type	Academic institution
Format	Print, Electronic
Portion	chart/graph/table/figure
Number of charts/graphs/tables/figures	1
The requesting person/organization	Taran Driver
Title or numeric reference of the portion(s)	Fig. 2
Title of the article or chapter the portion is from	The rotating-saddle trap: a mechanical analogy to RF-electric-quadrupole ion trapping?
Editor of portion(s)	N/A
Author of portion(s)	Ball, M G ; Harmon, T J ; Thompson, R I
Volume of serial or monograph	80
Issue, if republishing an article from a serial	12
Page range of portion	1435
Publication date of portion	Dec 1, 2002
Rights for	Main product
Duration of use	Life of current edition
Creation of copies for the disabled	no
With minor editing privileges	no
For distribution to	Worldwide
In the following language(s)	Original language of publication
With incidental promotional use	no
Lifetime unit quantity of new product	Up to 499
Title	Covariance mapping spectroscopy of molecular decompositions
Instructor name	Leszek Frasinski
Institution name	Imperial College London
Expected presentation date	Oct 2018

Note: This item will be invoiced or charged separately through CCC's **RightsLink** service. More info **$ 0.00**

Figure 5.6

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Sep 14, 2018

This Agreement between Mr. Taran Driver ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4427170997929
License date	Sep 13, 2018
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Mass Spectrometry Reviews
Licensed Content Title	Linear ion traps in mass spectrometry
Licensed Content Author	Donald J. Douglas, Aaron J. Frank, Dunmin Mao
Licensed Content Date	Apr 6, 2004
Licensed Content Volume	24
Licensed Content Issue	1
Licensed Content Pages	29
Type of use	Dissertation/Thesis
Requestor type	University/Academic
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	1
Original Wiley figure/table number(s)	Figure 1
Will you be translating?	No
Title of your thesis / dissertation	Covariance mapping spectroscopy of molecular decompositions
Expected completion date	Oct 2018
Expected size (number of pages)	1
Requestor Location	Mr. Taran Driver 38 Walsingham Road St. Andrews Bristol, BS6 5BT United Kingdom Attn: Mr. Taran Driver
Publisher Tax ID	EU826007151
Total	0.00 USD

Figure 5.10

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Sep 14, 2018

This Agreement between Mr. Taran Driver ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4423170308542
License date	Sep 06, 2018
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Rapid Communications in Mass Spectrometry
Licensed Content Title	Combined quantum chemical and RRKM modeling of the main fragmentation pathways of protonated GGG. II. Formation of b2, y1, and y2 ions
Licensed Content Author	Béla Paizs, Sándor Suhai
Licensed Content Date	Jan 28, 2002
Licensed Content Volume	16
Licensed Content Issue	5
Licensed Content Pages	15
Type of use	Dissertation/Thesis
Requestor type	University/Academic
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	1
Original Wiley figure/table number(s)	Scheme 2, panel (b)
Will you be translating?	No
Title of your thesis / dissertation	Covariance mapping spectroscopy of molecular decompositions
Expected completion date	Oct 2018
Expected size (number of pages)	1
Requestor Location	Mr. Taran Driver 38 Walsingham Road St. Andrews Bristol, BS6 5BT United Kingdom Attn: Mr. Taran Driver
Publisher Tax ID	EU826007151
Total	0.00 USD

Figure 11.1

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Sep 14, 2018

This Agreement between Mr. Taran Driver ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4410150821271
License date	Aug 15, 2018
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Methods
Licensed Content Title	Analysis and validation of proteomic data generated by tandem mass spectrometry
Licensed Content Author	Alexey I Nesvizhskii, Olga Vitek, Ruedi Aebersold
Licensed Content Date	Sep 27, 2007
Licensed Content Volume	4
Licensed Content Issue	10
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Will you be translating?	no
Circulation/distribution	<501
Author of this Springer Nature content	no
Title	Covariance mapping spectroscopy of molecular decompositions
Instructor name	Leszek Frasinski
Institution name	Imperial College London
Expected presentation date	Oct 2018
Portions	Figure 1
Requestor Location	Mr. Taran Driver 38 Walsingham Road St. Andrews Bristol, BS6 5BT United Kingdom Attn: Mr. Taran Driver
Billing Type	Invoice
Billing Address	Mr. Taran Driver 38 Walsingham Road St. Andrews Bristol, United Kingdom BS6 5BT Attn: Mr. Taran Driver
Total	0.00 GBP

Figure 20.2

AIP PUBLISHING LICENSE
TERMS AND CONDITIONS

Sep 14, 2018

This Agreement between Mr. Taran Driver ("You") and AIP Publishing ("AIP Publishing") consists of your license details and the terms and conditions provided by AIP Publishing and Copyright Clearance Center.

License Number	4410151431771
License date	Aug 15, 2018
Licensed Content Publisher	AIP Publishing
Licensed Content Publication	Review of Scientific Instruments
Licensed Content Title	Development of a four-element conical electron lens dedicated to high resolution Auger electron–ion(s) coincidence experiments
Licensed Content Author	K. Le Guen, D. Céolin, R. Guillemin, et al
Licensed Content Date	Nov 1, 2002
Licensed Content Volume	73
Licensed Content Issue	11
Type of Use	Thesis/Dissertation
Requestor type	Student
Format	Print and electronic
Portion	Figure/Table
Number of figures/tables	1
Title of your thesis / dissertation	Covariance mapping spectroscopy of molecular decompositions
Expected completion date	Oct 2018
Estimated size (number of pages)	1
Requestor Location	Mr. Taran Driver 38 Walsingham Road St. Andrews Bristol, BS6 5BT United Kingdom Attn: Mr. Taran Driver
Billing Type	Invoice
Billing Address	Mr. Taran Driver 38 Walsingham Road St. Andrews Bristol, United Kingdom BS6 5BT Attn: Mr. Taran Driver
Total	0.00 USD

Figure 20.6

Order Details

Journal of Physics B : Atomic, Molecular and Optical Physics

Order detail ID:	71575328
Order License Id:	4437721191755
ISSN:	0953-4075
Publication Type:	Journal
Volume:	
Issue:	
Start page:	
Publisher:	IOP Publishing
Author/Editor:	American Institute of Physics ; Institute of Physics (Great Britain)

Permission Status: Granted

Permission type: Republish or display content

Type of use: Thesis/Dissertation

Hide details

Requestor type	Academic institution
Format	Print, Electronic
Portion	chart/graph/table/figure
Number of charts/graphs/tables/figures	1
The requesting person/organization	Taran Driver
Title or numeric reference of the portion(s)	Figure 2
Title of the article or chapter the portion is from	X-ray free-electron lasers
Editor of portion(s)	N/A
Author of portion(s)	J. Feldhaus, J. Arthur, J.B. Hastings
Volume of serial or monograph	N/A
Page range of portion	S802
Publication date of portion	2005
Rights for	Main product
Duration of use	Life of current edition
Creation of copies for the disabled	no
With minor editing privileges	no
For distribution to	Worldwide
In the following language(s)	Original language of publication
With incidental promotional use	no
Lifetime unit quantity of new product	Up to 499
Title	Covariance mapping spectroscopy of molecular decompositions
Instructor name	Leszek Frasinski
Institution name	Imperial College London
Expected presentation date	Oct 2018

Note: This item will be invoiced or charged separately through CCC's **RightsLink** service. More info **$ 0.00**

Figure 20.7

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Sep 21, 2018

This Agreement between Mr. Taran Driver ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4433620614933
License date	Sep 21, 2018
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Photonics
Licensed Content Title	Fresh-slice multicolour X-ray free-electron lasers
Licensed Content Author	Alberto A. Lutman, Timothy J. Maxwell, James P. MacArthur, Marc W. Guetg, Nora Berrah et al.
Licensed Content Date	Oct 24, 2016
Licensed Content Volume	10
Licensed Content Issue	11
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Will you be translating?	no
Circulation/distribution	<501
Author of this Springer Nature content	no
Title	Covariance mapping spectroscopy of molecular decompositions
Instructor name	Leszek Frasinski
Institution name	Imperial College London
Expected presentation date	Oct 2018
Portions	Figure 1
Requestor Location	Mr. Taran Driver 38 Walsingham Road St. Andrews Bristol, BS6 5BT United Kingdom Attn: Mr. Taran Driver
Billing Type	Invoice
Billing Address	Mr. Taran Driver 38 Walsingham Road St. Andrews Bristol, United Kingdom BS6 5BT Attn: Mr. Taran Driver
Total	0.00 USD