DEVELOPING COMPUTATIONAL METHODS FOR FUNDAMENTALS AND METROLOGY OF MASS SPECTROMETRY IMAGING

by

ALEX DEXTER

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Chemistry
College of Engineering and Physical Sciences
The University of Birmingham
January 2018

Abstract

MSI is a suite of powerful imaging tools that can be used to perform untargeted unlabelled analysis into the distribution of a wide range of molecules from a variety of different sample types. Despite widespread use in numerous different research areas, many aspects of MSI fundamentals remain unknown. Not only are experimental aspects such as desorption and ionisation not always fully understood, but the success (or failure) of many of the computational methods used to mine these data cannot yet be easily evaluated. In this thesis, multivariate analysis methods are used to investigate fundamentals of laser parameters in raster mode MALDI imaging, and DF and CF variables in LESA coupled to FAIMS. Following this, novel methods to evaluate clustering agorithms are described, including multivariate normality testing for distance metric evaluation, and means to generate synthetic data based on multivariate normal distribution sampling. These synthetic data are then used to evaluate a variety of different clustering algorithms used previously in MSI and other fields, and a new, more efficient algorithm using graph based clustering and a two phase subset sampling approach is described. This is then demonstrated on large synthetic and real MSI datasets producing extremely accurate and informative segmentation.

ACKNOWLEDGEMENTS

First I would like to thank my supervisors; Josephine Bunch, Iain Styles and Helen Cooper for all their support and guidance throughout my PhD both in and out of work. I would also like to than all those who provided data for this thesis; Andy Creese for the ubiquitin FAIMS data, Naresh Kumar and Jean Luc Vörng for their work on the mixed polymer samples, Rory Steven for his data from the variable fluence images and transverse mouse brain, as well as help with the work in laser fundamentals, and Richard Goodwin, Heather Hulme and Jennifer Barnes for their contributions on the large colon dataset.

This work could not have been undertaken without the funding received from the EPSRC through the Physical Sciences of Imaging in the Biomedical Sciences (PSIBS) Doctoral Training Centre (grant code: EP/F50053X/1), and the NPL Strategic Capability programme 'AIMSHIGHER'.

I would like to thank everyone in the Bunch group at NPL, and in the PSIBS DTC and Cooper group in Birmingham, in particular Alan Race, Rory Steven, Elizabeth Randall, Spencer Thomas, Kenny Robinson, Adam Taylor, and Rian Griffiths. You have all made my PhD both fun and interesting. A big thanks to Vanessa fo all her help while writing up, particularly for keeping me fed while I write.

Finally, a huge thank you to all of my family for all their support. Without them I would never have gotten this far.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

m/z mass-to-charge ratio.

AGC automatic gain control.

CF compensation field.

CHCA α-cyano-4-hydroxycinnamic acid.

CID collision induced dissociation.

DAPNe direct analyte-probed nanoextraction.

DBSCAN density-based spatial clustering of applications with noise.

DC direct current.

DESI desorption electrospray ionisation.

DF dispersion field.

DMF dimethylformamide.

DPSS diode-pumped solid-state.

ECD electron capture dissociation.

EI electron impact.

ESI electrospray ionisation.

FAIMS high-field asymmetric waveform ion mobility spectrometry.

FFT fast Fourier transform.

FID free induction decay.

FLAME fuzzy clustering by local approximation of memberships.

FT Fourier transform.

FTICR Fourier transform ion cyclotron resonance.

FWHM full width half maximum.

GPU graphics processing unit.

H&E hematoxylin and eosin.

ISODATA Iterative Self-Organizing Data Analysis Technique Algorithm.

ITO indium tin oxide.

LA-ICP laser ablation - inductively coupled plasma.

LAESI laser ablation electrospray ionisation.

LC liquid chromatography.

LDA linear discriminant analysis.

LDI laser desorption/ionisation.

LESA liquid extraction surface analysis.

LMJ liquid microjunction.

MALDESI matrix-assisted laser desorption electrospray ionisation.

MALDI matrix assisted laser desorption/ionisation.

MANOVA multivariate analysis of variance.

MS mass spectrometry.

MSI mass spectrometry imaging.

Nd:YAG neodynium-doped yttrium aluminium garnet.

Nd:YLF neodynium-doped yttrium lithium fluoride.

Nd:YVO_4 neodynium-doped yttrium orthovanadate.

NMF non-negative matrix factorisation.

NMR nuclear magnetic resonance.

OPO optical parametric oscillator.

PCA principal component analysis.

pLSA probabilistic latent semantic analysis.

PMMA poly(methyl methacrylate).

PVDF polyvinylidene fluoride.

Q-TOF quadrupole time-of-flight.

RAM random-access memory.

RF radio frequency.

S/N signal-to-noise ratio.

SIMS secondary ion mass spectrometry.

SIT single ion transmission.

SOM self-organising map.

t-SNE t-distributed stochastic neighbour embedding.

TFA trifluoroacetic acid.

THRASH "Thorough High Resolution Analysis of Spectra by Horn".

TIC total ion current.

TIT total ion transmission.

TOF time-of-flight.

CHAPTER 1

INTRODUCTION

Mass spectrometry (MS) involves detecting gas phase ions which have been separated based on their mass-to-charge ratio (m/z). Extension of this to mass spectrometry imaging (MSI), mass spectra are acquired at spatially discrete locations across a sample to produce an image where each pixel contains a mass spectrum [1]. It is an unlabelled approach to map the spatial distribution of hundreds, if not thousands of molecules within a sample (often thin tissue sections), and is used to answer a huge variety of questions from drug pharmakinetics to protein and lipid pathways [2, 3]. Despite producing a huge wealth of information, typical MSI data analysis and review is often very simple, with either spectra from single pixels or regions of interest being investigated, or ion images for a given m/z analysed. MSI workflows can be extremely varied depending on the sample type, molecular class of interest, instrumentation, and desired final output. Generally however it can be divided into four stages; sample preparation, desorption and ionisation, mass analysis, and data processing and analysis. This introduction will provides a brief overview of each of these four steps.

1.1 Sample preparation

Sample preparation in MSI can range from almost nothing to complex multi-day processes depending on the experiment that is carried out. For almost any tissue analysis, and many

other samples, cryosectioning of a single organ or whole animal must be performed followed by thaw mounting onto a suitable substrate such as glass slides or stainless steel target plates. The thickness at which the tissue is sectioned during this can drastically affect the resulting spectral intensities when using certain ionisation methods, and so requires careful consideration [4]. In some cases, such as in certain desorption electrospray ionisation (DESI) and liquid extraction surface analysis (LESA) experiments, no further sample preparation is required. In others however, there can then be a number of additional sample preparation steps depending on the nature of the sample, ionisation method, and molecules of interest. For example, a proteomic experiment performed using tissue that has been formalin fixed will require antigen retrieval methods, as well as possible washing, and digestion steps [5]. Within each of these steps there are also a number of variables including different washing solvents [6, 7], and digestion protocols [8, 9]. For lipidomics experiments, direct analysis of formalin fixed tissue can be performed, without the requirement of further tissue preparation steps. The result of the fixation is a shift from potassium to sodium lipid adducts due to the presence of sodium in typical fixation buffers [10]. Finally, for matrix assisted laser desorption/ionisation (MALDI)-MSI, a matrix must be applied in order to facilitate laser energy absorption. The matrix used, along with the amount and the means used to apply the matrix can also significantly affect the results obtained and must be carefully selected based on the experiments carried out [11]. In summary, sample preparation is one of the most varied aspects of MSI and must be carefully selected based on the experiment that is to be performed [12].

1.2 Desorption and ionisation

There are a number of different desorption and ionisation methods, each of which has a number of different variables which can be altered. Broadly they can be grouped into two categories, hard and soft ionisation techniques. Hard ionisation methods including electron impact (EI) and secondary ion mass spectrometry (SIMS) impart a high amount

of internal energy to the molecules, resulting in a high degree of fragmentation. This can be useful in structure elucidation such as commonly used in EI, however it is generally not suitable for direct molecular analysis of complex samples [13]. Soft ionisation conversely tends to yield intact molecular ions and so is often more applicable to direct analysis of complex samples. Examples of soft ionisation methods include MALDI, and electrospray ionisation (ESI).

1.2.1 SIMS

The earliest examples of MSI are SIMS elemental analysis [14]. SIMS directs a beam of primary ions onto a surface, causing energy and charge transfer to molecules on the surface resulting in their desorption and ionisation [15]. Using monoatomic ion beams such as caesium often result in a large amount of fragmentation, and SIMS spectra typically contain very small to no intact species above m/z 500. The higher the mass of the ions used in the primary ion beam, the softer the ionisation, and large polyatomic and gas cluster ion beams can be used to analyse intact larger molecules. However this comes at the cost of a reduction in spatial resolution both in the xy plane and in depth [16]. SIMS offers higher spatial resolution than any other MSI technique, in part due to the ability to focus the primary ions into <100 nm spots, and due to it's very high sensitivity (10^{12} molecules cm^{-2}) [17, 18]. In addition to this, depth profiling by SIMS can be performed by analysing the surface and then sputtering off a layer [19]. This process can then be repeated to generate an image of a 3D volume [20]. These advantages come at the expense of two factors; the spectral resolution (as defined by the smallest m/z difference that can be separated at a given m/z) of SIMS is typically < 5000 at m/z 200 [21], and SIMS is typically unable to analyse intact molecular species above $m/z \sim 500$ as previously discussed. Some large cluster ion sources are capable of analysing intact molecules as large as 14 kDa but this is not routinely performed and has limited spatial resolution and sensitivity [22, 23].

SIMS is used in both molecular and elemental analysis of a huge variety of samples

including tissue sections [24], single cells [25], polymers [26], and electronics [27]. It is usually employed in applications where MALDI and other ionisation techniques do not provide sufficient spatial resolution, or where 3D imaging is required.

1.2.2 LDI and MALDI

In laser desorption/ionisation (LDI), laser energy is directed onto a surface, and then absorbed by molecules, resulting in their subsequent desorption and ionisation [28]. This is limited to molecules that are capable of absorbing the laser wavelength employed, and are volatile enough to be desorbed from the surface. For molecules that do not readily desorb by LDI, a matrix molecule can be co-crystalised with the molecules of interest to facilitate desorption and ionisation. MALDI uses a laser to deliver energy to matrix molecules, typically low molecular weight organic acids, resulting in the generation of a plume of vaporised matrix and co-crystallised analyte molecules (Figure 1.1) [29]. In this plume, charge transfer occurs from matrix to analyte molecules to produce the ionised analyte ions prior to mass detection. While some models have been proposed, the exact processes involved in MALDI plume generation and ionisation are still not fully understood, a potential major hindrance to areas such as quantitation and experimental design [30, 31].

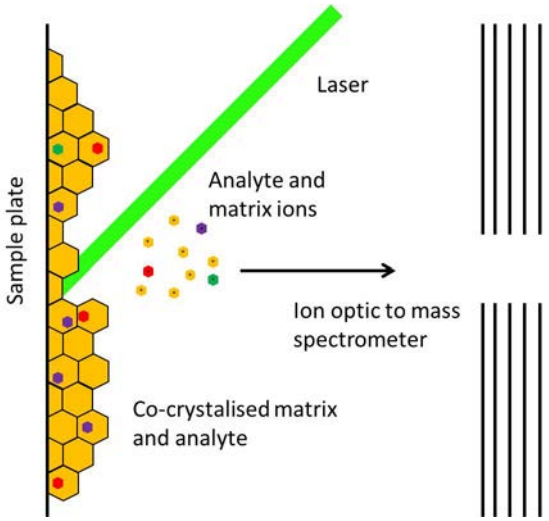

Figure 1.1: Basic principle of the MALDI desorption process. A sample is coated in a matrix compound (orange), and laser energy incident on the sample results in desorption and ionisation of the matrix along with co-crystalised analyte molecules.

The funadmental processes of MALDI are a highly studied area of research with numerous mechanisms for ion formation proposed (table 1.1). A full understanding of the relationship between these mechanisms, and their relative contributions however, remains unknown. Despite this, MALDI is one of the primary ionisation techniques used in MSI as it offers the potential to analyse a wide range of molecular classes, can acquire images at high spatial resolution ($\sim 1\mu m$), and is very sensitive for certain compounds [10, 32–34]. MALDI-MSI has been used to analyse a huge range of samples types including thin tissue sections [35], bacterial colonies [36], and plants [37]. By far the most common of these is the analysis of thin tissue sections, and within this, it has been used to analyse a wide range of different tissue types including a number of cancer types [38–40], different organs [10, 41, 42], and even whole animals [43]. In addition to this, MALDI has been used to analyse a number of different molecular classes from small molecule lipids and metabolites [44, 45], to large proteins [46, 47].

Model	Ionisation process	Reference
Multiphoton ionisation	$M \xrightarrow{n(hv)} M^{+\bullet} + e^-$ $M \xrightarrow{hv} M^* \xrightarrow{m(hv)} M^{+\bullet} + e^-$	[48, 49]
Energy pooling	$MM \xrightarrow{2hv} M^*M^* \rightarrow M + M^{+\bullet} + e^-$ $M^*M^* + A \rightarrow MM + A^{+\bullet} + e^-$	[50, 51]
Excited state proton transfer	$M + hv \rightarrow m^*$ $M^* + A \rightarrow (M - H)^- + AH^+$ $M^* + M \rightarrow (M - H)^- + MH^+$	[52]
Disproportionation reactions	$2M \xrightarrow{n(hv)} (MM)^+ \rightarrow (M - H)^- + MH^+$ and/or $\rightarrow M^- + M^+$	[49]
Desorption of preformed ions	$[M+H^+_{(s)} \rightarrow M+H^+_{(g)}$ $[M+X^+_{(s)} \rightarrow M+X^+_{(g)}$	[53]
Thermal ionisation	$2M \xrightarrow{\Delta H} M^- + M^+$	[49]
Pressure pulses	Rapid mechanical stress	[54]

Table 1.1: Table of the various desorption processes in MALDI. The exact mechanism may involve different proportions of some or all of these reactions which may depend on the sampling conditions employed.

There are a large number of variables that can be altered in an LDI or MALDI experiment, one area of which is the properties of the laser employed. Variable properties of the laser include the number of laser pulses delivered per second (repetition rate), the time over which a single laser pulse occurs (pulse width), the energy delivered in a single pulse, the wavelength employed, and the laser spot size employed. In many cases there is little to no control or selection of many of these parameters on commercial instruments. The effect of, and relationship between some of these parameters is complex and is only partially understood [49]. For example, ion intensity can be related to laser fluence (a function of energy per pulse and spot size), by a power law [55]. However the threshold at

which the ionisation occurs is also dependent on the spot size, and also spot shape [56]. Further complicating this, MALDI often employs oversampling, where the laser spots overlap. Since it is assumed all the matrix is ablated in each analysis this means that the removed material can be considered to be only from the previously unsampled region. A more detailed understanding of the effect of these variables would allow a better understanding of the MALDI process and would allow more efficient optimisation of these parameters.

The wavelength of the laser employed is typically defined by the lasing medium employed and by any harmonic generation. Initial MALDI studies employed nitrogen lasers which typically operate at 337 nm, repetition rates in the 10s of Hz, and 0.5-3 ns pulse widths [57]. Recently, neodynium based neodynium-doped yttrium aluminium garnet (Nd:YAG), neodynium-doped yttrium orthovanadate (Nd:YVO_4), and neodynium-doped yttrium lithium fluoride (Nd:YLF) lasers capable of operating in the kHz range have been applied to MALDI experiments, increasing throughput and sensitivity [58, 59]. A summary of some of these characteristics can be seen in table 1.2. Variable wavelengths can also be achieved through the use of an optical parametric oscillator (OPO), and investigations into the effect of wavelength have shown improved ion yields corresponding to wavelengths at which the matrices have high optical absorption [60].

The minimum laser spot size is ultimately diffraction limited and will be limited by the wavelength employed [61]. This is not typically an issue however as laser spot sizes in MALDI vary from around 5-100 μm. This is mainly due to sensitivity limitations when only analysing a small portion of a sample. Focusing the laser into a smaller spot size at the same energy will result in a higher fluence, however, reducing the spot size has also been shown to increase the threshold fluence at which ionisation occurs [56]. The total energy delivered in a single pulse, along with laser spot size determines the fluence incident on the sample. Below a given threshold fluence, little to no desorption or ionisation occurs. Above this threshold, a rapid increase in the detected ion intensity is observed which has been described by a power law [49]. Use of even higher fluences then

results in an even greater amount of internal energy delivered to the molecules present, and causes a greater amount of fragmentation [62]. Typically, this parameter is manually optimised before any experiment to give a compromise between high intact molecular ion intensities and fewer fragment ions. With such a complex interdependence between these variables, it is crucial to more fully understand the relationship between them in order to acquire robust and reliable MALDI-MSI data. In order to be able to fully understand the effects of changing these variables, more complex multivariate data analysis methods are also required to efficiently mine and analyse these data.

Laser	Fundamental wavelength	Typical operating wavelengths	Repetition rates	Pulse width
Nitrogen	337 nm	337 nm	10's of Hz	0.5-3 ns [57]
Nd:YAG	1064 nm	355, 532 and 266 nm	10 Hz - 10 kHz	5 ns [13]
Nd:YVO_4	1064 nm	355, 532 and 266 nm	10 Hz - 20 kHz [58]	1.4 ns [63]
Nd:YLF	1048 nm	349 and 262 nm [64, 65],	up to 5 kHz	5 ns [66]

Table 1.2: Table of common lasers employed in UV-MALDI, and their properties

1.2.3 DESI

DESI has some similarity with SIMS, in that primary ions, in this case solvent clusters generated via ESI, are directed onto the target surface. These then desorb and ionise analyte molecules from the surface allowing their subsequent mass analysis [67]. Unlike SIMS, DESI MS is performed under ambient sampling conditions, and there are two possible routes for the desorption and ionisation process. Direct desorption and ionisation similar to the SIMS ionisation process can occur, resulting in spectra containing primarily singly charged species. Alternatively, analyte molecules can be extracted into the solvent droplets. This, followed by their desolvation until charge repulsion overcomes the surface tension, as occurs in classical ESI, gives rise to multiply charged species [68, 69]. While

DESI is sometimes considered non-destructive [70], there will be some loss of material in order to form ions at the detector, along with possible modifications on the surface due to the presence of solvent. This has implications on spatial resolution, as unlike MALDI, true oversampling cannot be achieved. Kertesz *et al.* however, showed spatial resolutions smaller than the size of the impact plume on printed ink standards [71]. This was attributed to the effective desorption and ionisation region of the impact plume being significantly smaller than the size of the plume itself, therefore decreasing the effective spot size interogated. DESI is capable of directly analysing proteins [72], however from a tissue imaging perspective, it has so far been limited to small molecule lipid, drug, and metabolite analysis [73, 74]. As with MALDI there are a number of different parameters within a DESI experiment that can affect the resulting spectra, such as the solvent composition, spray angle, and the distance of the spray head to sample [75–77].

1.2.4 LMJ based surface sampling

Liquid microjunction (LMJ) based surface sampling techniques involve performing localised extractions of molecules at a surface, then subjecting the resulting analyte containing solution to ESI. There are currently four platforms that have been used to perform liquid microjunction based sampling; LESA, commercialised by Advion in the Nanomate (Figure 1.2), continuous flow LMJ sampling, commercialised by Prosolia in the Flowprobe (Figure 1.3), nano-DESI (Figure 1.4), and direct analyte-probed nanoextraction (DAPNe) [78–82]. The continuous flow LMJ surface sampling in the flowprobe and nano-DESI are essentially the same as one another with the exception that in nano-DESI the two capillaries are separate from one another as opposed to one inside the other for the flowprobe. DAPNe is very similar to the nanomate, except the solvent volumes are considerably lower, and there is a much greater degree of control in the sampling tip. Since the method of ionisation is ESI, multiply charged ions are produced, allowing intact proteins to be analysed using high resolution mass analysers [83].

LMJ surface sampling coupled to ESI MS has been demonstrated to be capable of

simultaneously analysing a number of molecular classes such as drugs, lipids and proteins, and analysing a wide variety of different samples types such as thin tissue sections, dried blood spots, and bacterial colonies [84–89]. The relatively fast acquisition for single location analysis, high sensitivity, and minimal sample preparation means that these techniques are ideally suited towards pharmaceutical applications [90], and proteomics [89]. From a proteomics perspective, LESA has a number of significant advantages over traditional MSI methods such as MALDI and DESI. The extraction and ionisation processes are decoupled from one another, thereby allowing modifications such as enzymatic digestion to be performed prior to ionisation and mass analysis [83, 89]. This does however introduce additional analysis time, which limits its viability in larger scale imaging studies. Also, since a relatively large volume of solvent can be used to perform the extraction (2-5 μL), this means that for a single pixel, a long analysis can be performed, allowing for detailed analysis and tandem mass spectrometry from a single pixel [91]. The LESA and flowprobe systems have also been coupled to high-field asymmetric waveform ion mobility spectrometry (FAIMS) offering a further orthogonal mode of separation to enhance the capabilities of these techniques [92, 93].

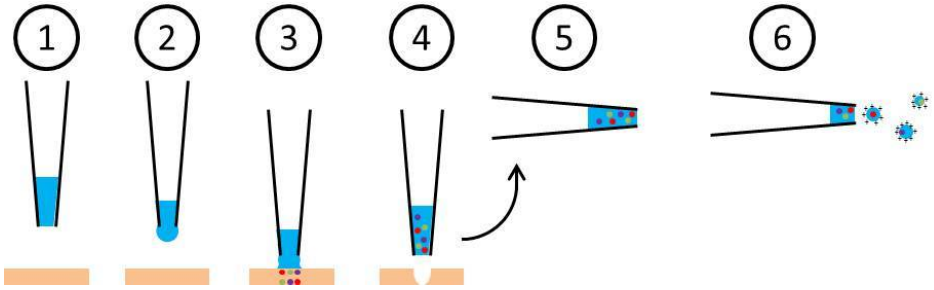

Figure 1.2: Basic principle of LESA. Solvent is aspirated into a conductive pipette tip (1) before being dispensed to create a microjunction with the surface (3) extracting the analyte molecules of interest before being reaspirated into the pipette (4) and subsequent mass spectrometric analysis by ESI (6).

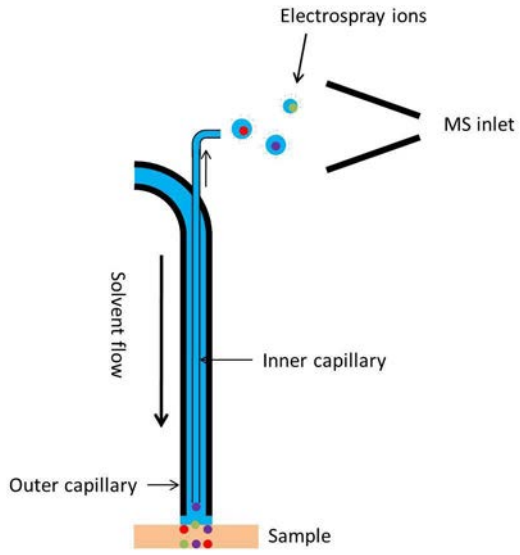

Figure 1.3: Basic principle of the continuous flow surface sampling system. Like LESA, a solvent microjunction is held on the surface, but unlike LESA, solvent continuously flows from the outer capillary to the sample and back up through the inner capillary to the ESI source.

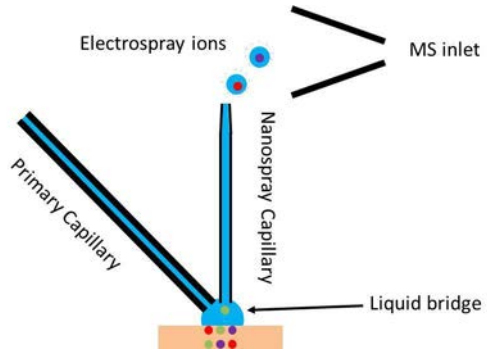

Figure 1.4: Basic principle of nano-DESI. This is very similar to the continuous flow sampling, except the two capillaries are separate from one another rather than coaxial.

FAIMS

Performing LESA MS from biological substrates extracts a large number of different biomolecules, resulting in very complex spectra. Typically when performing MS experiments on these complex biological samples, an orthogonal mode of separation, usually

liquid chromatography (LC) is applied prior to ESI and subsequent mass analysis. While this has been demostrated within a LESA workflow [83], the timeframes involved in LC experiments make it unsuitable for high throughput or imaging experiments. An alternative to this is the use of FAIMS, which has been shown to separate different species in much more amenable timeframes for imaging experiments, as well as having the potential to separate isomeric species [94], and increasing signal-to-noise ratio (S/N) [93].

In FAIMS, ions are separated based on their changes in mobility in the presence of high and low electric fields. Ions travel between two parallel electrodes by means of a carrier gas. During this time, a high electric field is applied for a short time, followed by a lower field of opposite polarity for a longer time, such that the integral of these is equal to one another (Figure 1.5), this is known as the dispersion field (DF). If an ion's mobility doesn't change between the high and low field, then there will be no change in its trajectory through the cell (blue ions in figure 1.5). Other ions will have increased mobility in high field, and as such will be drawn towards the top electrode (green ions in figure 1.5), whilst ions which have decreased mobility will be drawn towards the bottom electrode (red ions in figure 1.5). By applying a static field, known as the compensation field (CF), different ions can be selectively transmitted through the FAIMS cell (Figure 1.6). Combining LESA with FAIMS, improvements have been made in S/N since much of the chemical noise does not pass through the mobility cell and is therefore filtered out. This reduces analysis time as for many ions of interest, a single or few scans can be used, rather than having to combine multiple scans to achieve sufficient S/N [93]. By using FAIMS and tailoring CF and DF conditions, both lipids and proteins have been detected from a single LESA extraction spot which cannot be achieved by other methods such as MALDI or DESI [91].

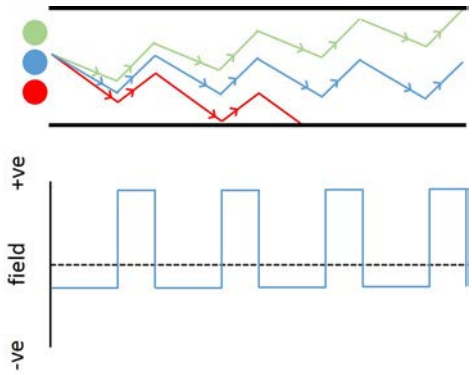

Figure 1.5: Principle behind FAIMS separation. The three ions (red, green and blue) have different mobilities in the high and low positive and negative fields. The red ions move faster in the low field and so hit the lower electrode, whereas the green have a higher mobility in the high field and so hit the top electrode. The blue ions have the same mobility in high and low fields and so pass through the FAIMS cell.

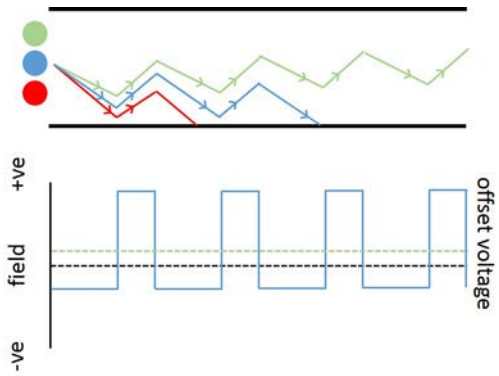

Figure 1.6: FAIMS separation with a DC offset. This then results in a greater low field mobility and a smaller high field mobility. This allows transmission of the ions whose mobilities are greater in high fields (green ions). A similar offset for a greater high field could be applied to transmit the ions whose mobilities are greater in low fields thus transitting the red ions.

For a comparison of the common ionisation sources in MSI see table 1.3.

Technique	Fastest speed	Smallest pixel size reported	Molecular classes
MALDI	50 pixels / s [95]	1 μm [96]	Drugs, metabolites, lipids, proteins and more
DESI	50 pixels / s [97]	40 μm [71]	Lipids, drugs and metabolites
TOF-SIMS	100 pixels / s [98]	390 nm [34]	Small molecules and fragments
LESA	1 min per pixel [84]	1 mm [84]	Drugs, metabolites, lipids, proteins and more

Table 1.3: Comparison of common ionisation sources used in MSI.

1.2.5 Additional ionisation sources

There are now upwards of thirty additional ionisation methods, and a full review of these methods can be found elsewhere [99, 100]. These can be classified into a number of groups; laser, plasma, and solvent based methods. Some of these techniques such as laser ablation electrospray ionisation (LAESI) and matrix-assisted laser desorption electrospray ionisation (MALDESI) use combinations of these by using lasers to desorb ions and solvent to provide charged ions. Some of these methods such as LAESI have been applied to imaging. The plasma based methods have not seen imaging applications yet however as there are few ways to control the footprint of plasma sampling.

1.3 Mass analysis

Following ionisation, the gas phase ions must be separated according to their mass to charge ratio by means of a mass analyser. This can be achieved in a number of ways, these are all based around accelaration of ions using electric or magnetic fields. The mass analyser used in any MS experiment will affect performance aspects such as the mass accuracy, resolving power, and m/z range, as well as have an effect on sensitivity.

1.3.1 TOF

Time-of-flight (TOF) mass analysers accelerate ions down a long flight tube using applied voltages such that the resulting kinetic energy for any ion is proportional to its charge. For an ion of mass m and charge z, the kinetic energy $E_k = \frac{1}{2}mv^2$, thus $v^2 \propto \frac{z}{m}$. By accelerating these ions down a long flight tube length d, they will separate based on their differential velocities, and thus will be detected at differential time points where the time of detection $t \propto \frac{d}{v}$ (depending on flight path length l). Combining this with $v^2 \propto \frac{z}{m}$ and $E_k = \frac{1}{2}mv^2$ gives $t \propto \sqrt{\frac{m}{z}}$, and with knowledge of the length of the flight tube, and the electric field applied, the time of detection can then be converted back into a mass to charge ratio. Since $t \propto \sqrt{\frac{m}{z}}$, mass resolution in TOF analysers becomes increasingly poor as m/z increases. This can be improved through the use of longer flight tubes but becomes unrealistic above a certain size. The alternative is to redirect the motion of the ions via an electrostatic field known as a reflectron generating either a v or w shaped flight path (Figure 1.7) [101]. As well as increasing the effective flight path, this also minimises any initial spread in ions velocity caused by inhomogeneous acceleration voltages because higher energy ions will travel further into the reflectron field than low energy ions [101].

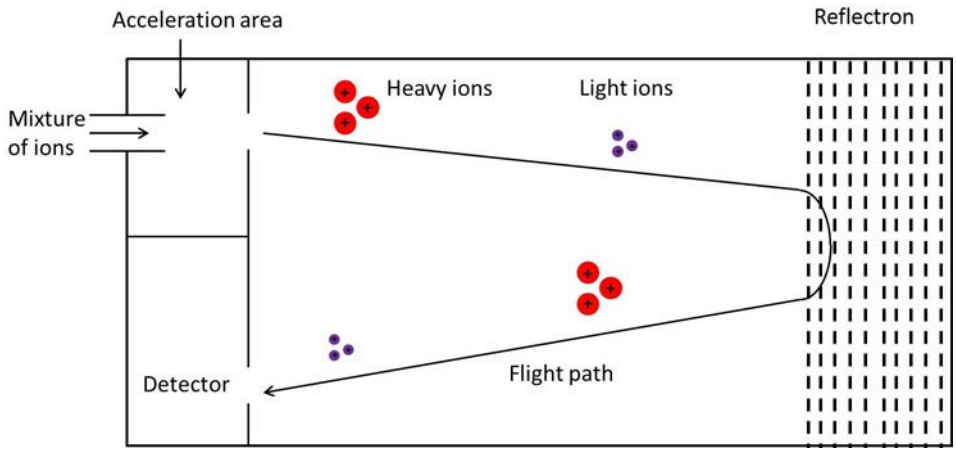

Figure 1.7: Basic principle of a TOF mass analyser. Ions are accelarated into the flight tube, deflected back along the flight tube by the reflectron before reaching the detector. The heavier ions will take longer to reach the detector than the lighter ions, and thus the ions will separate by time.

TOF mass analysers are one of the most commonly used mass analysers in MSI, particularly MALDI and SIMS, as they can offer fast scan times per pixel, and a theoretically unlimited m/z range. They are the only mass analyser that can be used when performing intact protein analysis by MALDI due to their theoretically unlimited mass range. In SIMS, TOF mass analysers are almost exclusively used as they are very fast in the lower mass range [59]. When using TOF mass analysers in MALDI and SIMS experiments however, ions will take different times to reach the acceleration fields depending on the topography of the sample leading to topographically induced mass shifts in the spectra [102]. These shifts can be corrected for if peaks of known m/z are present within the data or can be used to determine the topography of the sample itself [102] but otherwise lead to inaccurate and imprecise measurements of m/z. One way to overcome this is by orienting the TOF orthogonal to the inlet of the mass spectrometer [103]. Since the direction of motion down the flight tube is orthogonal to the initial direction of motion of the ions, this no longer influences their arrival time at the detector and so will not affect their resultant m/z measurement [104]. Orthogonal TOF mass analysers are usually preceeded by a quadrupole mass filter placed before the flight tube.

1.3.2 Quadrupole

Quadrupole mass analysers consist of four parallel cylindrical rods, to which radio frequency (RF) and direct current (DC) potentials are applied [105]. Depending on the potentials that are applied, quadrupole mass analysers filter ions based on their m/z, as only ions of a given m/z will have a stable trajectory (Figure 1.8) [105]. By scanning through the potentials, a mass spectrum can be generated, but this will result in a low ion transmission yield as all ions outside of the selected m/z are lost and so not detected when scanning through the full m/z range. The m/z range and resolution of quadrupole mass analysers are dependent on the properties of the rods employed e.g. length and diameter, as well as the maximum applicable RF voltage and frequency [105]. Typically, quadrupoles are operated at an m/z of up to a few thousand and with a resolutions of up

to one thousand [105], but a resolution of up to 3000 full width half maximum (FWHM) and mass range up to 70 kDa can be achieved [106]. Single quadrupoles are fairly limited on their own due to this limited mass resolution and range, and low ion transmission yield for acquisition of a full spectrum. They are usually employed in tandem mass spectrometry experiments to filter ions of a given m/z prior to fragmentation.

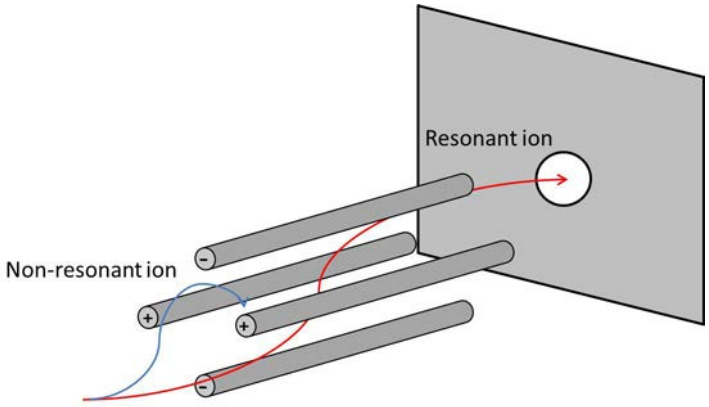

Figure 1.8: Schematic of a quadrupole mass analyser. Potentials are applied to the four rods, resulting in stable trajectories for resonant ions of a given m/z and causing non-resonant ions to discharge on the rods.

Tandem mass spectrometry

In mass spectrometry, there are a multitude of different possible isomeric structures that could give rise to a peak at a single m/z. Added to this, any uncertainty in the m/z will further add to the number of possible molecules that any given peak could be derived from. In a number of areas such as drug discovery and proteomics, it is vital to be able to elucidate the exact structure of the ion that has produced a peak to prevent any possible false positives for experiments attempting to determine drug or protein localisation. In MSI this is achieved through tandem mass spectrometry experiments. Tandem mass spectrometry involves the combination of multiple mass analysers with some form of fragmentation in between [107]. A parent mass is first selected using a quadrupole, and this is then fragmented in a number of possible ways. Any given molecule will fragment

into a unique set of ions, which can then be analysed using a second mass spectrometer, producing a characteristic fragmentation spectrum. By analysing the mass of the parent molecule, along with it's fragmentation, a more confident assignment of the parent molecule can be made. One of the most common methods to perform fragmentation is collision induced dissociation (CID). This involves selecting a parent m/z of interest using a quadrupole, and accelerating the ions within this m/z window rapidly in a second RF only quadrupole, followed by collision with a neutral collision gas, usually nitrogen or argon [108]. This causes the ions to fragment, and the resultant fragment ions can be analysed by a final mass spectrometer. The most common setup for tandem mass spectrometry MSI experiments are quadrupole-TOF instruments where parents are selected in the quadrupole, and a TOF mass analyser is used to separate the fragment ions. Alternatively, ion traps can be used, which provide greater S/N in full scan modes, but lower precision [109]. Another possible means to impart more confidence in a molecular assignment, is through the use of high mass resolution MS using Fourier transform (FT) based mass analysers.

1.3.3 FTICR

Ions are accelerated by means of an RF electric field, and diverted into a cyclic orbit by a high magnetic field generated using superconducting magnets [110]. The frequency at which ions orbit is dictated by an ion's mass, charge, and the strength of the magnet employed [110]. In a Fourier transform ion cyclotron resonance (FTICR) mass analyser, ions are detected by their induction of a current on electrodes due to cyclotron motion as they pass near to them (Figure 1.9). This produces a free induction decay (FID), similar to that observed in nuclear magnetic resonance (NMR), also known as the transient, which is converted into a mass spectrum via Fourier transform. Since a single ion can induce a current multiple times, FTICR instruments have extremely high sensitivity. FTICR mass spectrometers also offer the highest possible mass resolution of up to 39,000,000 at m/z 609 [111]. This resolution is determined by the strength of the superconducting magnet

used, and duration of detection, thus FTICR instruments are much slower at acquiring data than TOF instruments, and are less widely used in imaging applications [112]. These mass analysers are used in applications that require the very high mass resolution, such as resolving isotopic fine structure in proteomics imaging [113].

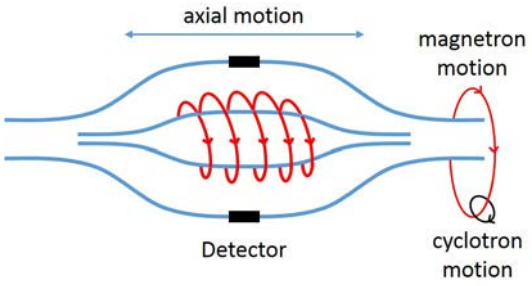

Figure 1.9: Principle behind FTICR and orbitrap mass analysers. In Orbitraps the frequency of axial motion is detected, and in FTICRs the frequency of cyclotron motion is detected.

1.3.4 Orbitrap

Orbitrap mass analysers contain two coaxial electrodes(Figure 1.9 blue components) which redirects ions into orbital trajectories, thus trapping them within the cell [114]. Unlike FTICR, orbitrap mass analysers distinguish the m/z of ions based on the frequency of axial motion along the spindle rather than their cyclotron frequency around it. One advantage of this is that this is not dependent on the initial velocity of the ions and thus resolution is limited by charge interaction and inhomogeneity of the electric fields [115]. As with FTICR, detection of these ions occurs via the induced charge resulting from the ions passing electrodes, which is then converted into spectra using Fourier transform resulting in similarly high sensitivities. The main advantage of orbitraps over FTICR are that they do not require superconducting magnets, and so are significatly smaller and cheaper. They are however limited to a mass range of up to 6,000 Da typically, and thus cannot be used for intact protein analysis unless multiply charged ions can be generated. Orbitrap mass analysers are used in similar applications to FTICR, such as proteomic and metabolomic

studies where high mass resolution is required to resolve isotopic strucutre [116, 117]. There are a number of different Orbitrap mass analyser types which will differ in their scan speeds, maximum available resolution, and MS/MS capabilities.

1.4 Data analysis

MSI data contains a spectrum per pixel with m/z intensity pairs and a pixel co-ordinate location in x and y. As such, MSI data contains a huge wealth of information, analogous to hyperspectral imaging data. One of the main challenges in MSI data analysis is the data size, as it is high dimensional data, many algorithms are not suitable or not scalable to MSI data. In addition to being large, MSI data are also highly heterogeneous with a large amount of pixel to pixel variability further adding to data analysis challenges. Typical data processing workflows in MSI involve a number of steps; If necessary, data must be converted from proprietary format into the required format for the software used to analyse it [118]. Following data conversion, a number of preprocessing steps can be applied, including zero filling, smoothing, baseline correction, normalisation and peak detection. For a full review of data conversion and preprocessing methods see [119]. These processed data can be represented by a datacube with pixels in x and y and spectra in the z dimension (Figure 1.10). Following data preprocessing, in many cases, simple univariate analysis of individual mass spectra or ion images can be performed, however this is very labour intensive, is prone to user bias, and does not always fully capture the full information available. In most examples of multivariate analysis of MSI data, the x and y dimensions are combined to give a single matrix of pixel by m/z, removing the spatial neighbourhood information in the data. The exception to this is the spatially aware denoising and clustering methods described by Alexandrov et al. [120, 121].

Data mining is a valuable tool in MSI, where even a single image can contain more information than can be feasibly interpreted by a single person in a realistic timeframe. It is becoming increasingly clear however that simple univariate analysis is both impractical

and does not take full advantage of the rich content of the data, and that multivariate analysis methods are increasingly important to effectively mine these data [122]. There are a number of different ways in which data mining is used, such as correlation analyses, image segmentation, and feature extraction. For each one of these goals there are a number of different multivariate analysis methods that can achieve this, which can be broken down into three categories; clustering algorithms, dimensionality reduction techniques, and classifiers.

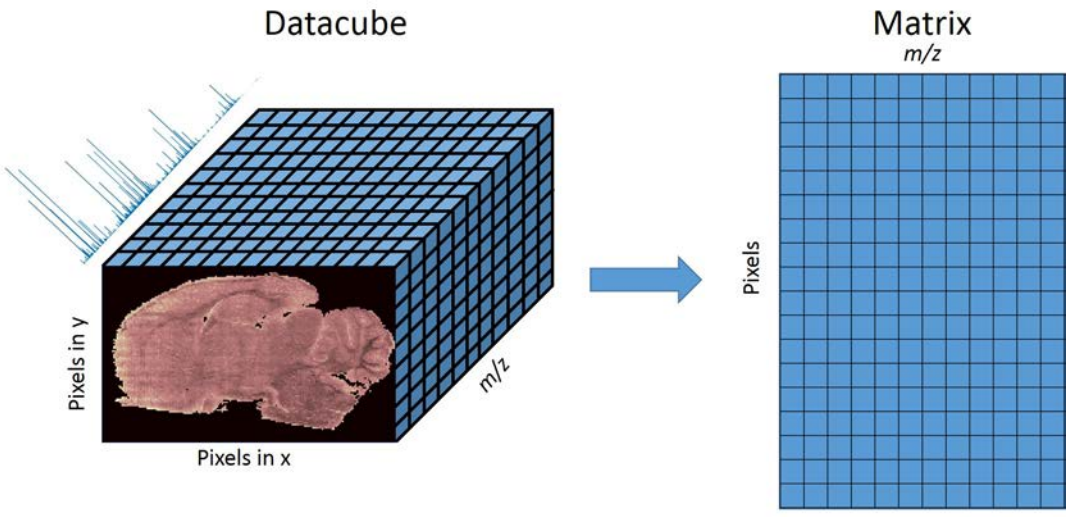

Figure 1.10: Datacube representation of MSI data. Each pixel has a co-ordinate location in x and y, and the spectrum occupies the z dimension. For most multivariate analysis, the x and y dimensions are combined to give a 2D matrix of pixels by spectra.

1.4.1 Computational complexity

Computational complexity is a means to assess the resources (time and memory) required to undertake a given algorithm [123]. Rather than denoting the exact measures of efficiency, which can depend on a large number of variables, the complexity of an algorithm is estimated based on the limiting factors of an arbitrarily large dataset. There are two mains forms of algorithm complexity; time complexity, the number of calculations required for the algorithm, and space complexity, the amount of memory required by the

algorithm. For an MSI dataset with n pixels, d spectral channels, clustered into k clusters, the complexity of different algorithms is summarised in table 1.4. In the case of iterative algorithms, i is the number of iterations taken to converge, and for spatially aware clustering q is the number of dimensions the data is reduced to during the FastMap step, and r is the size of the neighbourhood.

Algorithm	Time complexity	Space complexity
PCA	$O(n.d^2 + d^3)$	$n.d + d^2$
NMF	$O(n.d.k.i)$	$n.d$ [124]
t-SNE	$O(n^2.d)$	n^2
k-means	$O(n.d.k.i)$	$n.d$ [125]
bisecting k-means	$O(n.d.\log_2 k.i)$	$n.d$
Agglomerative hierarchical	$O(n^2.d + n^2 \log n)$	$n^2 + n.d$
Spatially aware and structurally adaptive	$O(n.d.q + n.q.i.k.r^2)$	$n(2q+1) + n.d(2r+1)^2$
Spectral clustrering	$O(n^2.d + n^{2.373})$ [126, 127]	$n^2 + n.d$
DBSCAN	$O((n \log n).d)$	$n.d$
FLAME	$O(n^2.d)$ [128]	$n.d$
Affinity propogation	$O(n^2.d + n^2.i)$	$n^2 + n.d$
Two phase k-means	$O(n.d.k.i)$	$(\sqrt{n.k}).d$

Table 1.4: Complexities and memory requirements for different clustering algorithms and dimensionality reduction techniques.

1.4.2 Dimensionality reduction

Dimensionality reduction techniques seek to generate lower dimensional representations of the data that still retain the majority of the information. This can either be used as a precursor step to other analyses such as PCA being used prior to clustering [129], or can

be used to directly interpret the data itself [130]. Examples of dimensionality reduction techniques that have been applied to MSI data include PCA [131], non-negative matrix factorisation (NMF) [132], random projection [133], probabilistic latent semantic analysis (pLSA) [134], t-SNE [122], self-organising map (SOM) [122], and autoencoders [135]. Reduction of the data by these means allows users to quickly determine particular ions, or groups of ions that show interesting features, or are correlated with one another. This removes much of the need to manually analyse all of the data, and can provide shortlists for further analysis. Additionally, by reducing to three dimensions and assigning colour channels red, green, and blue to the now three dimensional data, segmentation maps can be produced which differentiate tissues in a similar fashion to clustering [122]. The main issue with dimensionality reduction methods, when used to directly interpret the image, is that they do not produce clear differentiation between regions and as such still rely upon user interpretation of the resulting colour-mapped image [136]. In some cases this is an advantage since there will not always be a distinct boundary between two regions but a gradient of change.

1.4.3 Classifiers

Classifiers use supervised learning to divide the dataset into classes and assign a single class label to each pixel and as such provide a clear categorisation of the data [137]. Firstly, a training dataset with known labelling is used to generate a classification model which is subsequently applied to the unknown data. Consider for example, a MALDI MSI dataset containing cancerous and non-cancerous tissue. A small subset of data are externally labelled and differentiated using traditional histopathological methods, either on serial sections or post MSI analysis on the same section. The corresponding MSI data are then segmented based on the staining and used as the training set to generate a classification model. This model is then applied to the remaining MSI data and thus the whole dataset is segmented into two catagories, cancerous and non-cancerous. This model can then be applied to any subsequent MSI datasets, provided there are no significant alterations

in experimental parameters. In the majority of cases, classifiers designed for the use with non imaging MALDI data have been used, primarily with a view towards biomarker discovery in various cancer types [138,139]. However as noted by Alexandrov [140], unlike MALDI spot analysis, MALDI MSI suffers from a number of issues, such as a high degree of spectral heterogeneity derived from experimental variance, and a large difference in comparison group size. In order to develop accurate classifiers tailored specifically towards MSI data these factors must be considered. Recently, Veselkov *et al.* developed workflows based on log transformations on the data, aimed at addressing some of these challenges, acheiving greater than 98% accuracy [141].

1.4.4 Clustering

Clustering is frequently used to segment different regions of an image for the purpose of diagnosis of diseases or to improve disease understanding, and to segment anatomical regions for comparison to histology in order to better understand the molecular composition of different anatomical regions [130,131,142,143]. Unlike classifiers, clustering techniques are an unsupervised method to divide the data, based on measures of similarity within the data. However, the idea of a cluster of data is arbitrary, relying on the notion of "similarity" which can be formulated in many ways. As a result there are a multitude of different clustering algorithms each with different assumptions about the data [144]. The following clustering algorithms have been applied to MSI data; k-means, bisecting k-means clustering, agglomerative hierarchical clustering, spatially aware clustering, spatially aware structrally adaptive clustering, and spectral clustering [120, 131, 142]. Outside of MSI, there are numerous other algorithms tailored towards differnt goals, of these, five of the more popular algorithms; density-based spatial clustering of applications with noise (DBSCAN) [145], fuzzy clustering by local approximation of memberships (FLAME) [128], affinity propogation [146], graph based clustering algorithms [147, 148] and two-phase k-means clustering [149] will be outlined in this introduction.

Distance metrics in clustering

The distance metric used by the clustering algorithms to compare one spectrum to another (Figure 1.11), and any normalisation strategies applied to the data prior to analysis have a significant effect on the results of any clustering algorithm. In MSI, there can be significant variations in the data that derive from a number of different experimental sources. For example, variations in sample preparation [12], and laser instability [150] both introduce a source of non-biological variance within the data. Minimising these effects by normalisation remains a challenge that is yet to be fully addressed [151]. Many different normalisation strategies make specific assumptions about the data and as such are only applicable to certain datasets. For example, normalisation to matrix peaks assumes that the matrix is homogeneously distributed over the tissue and as such should be constant, however most matrix peaks occur at low m/z (usually less than 300) and so will not account for changes in ion transmission at different mass ranges. Nevertheless, normalisation of the data, or pseudo normalisation achieved by the use of the cosine and correlation distances, reduce the effects of these variations, and thereby improve the clustering results. In the commonly applied total ion current (TIC) normalisation, each spectrum is normalised to have unit sum intensity (also referred to as l_1 norm). The cosine and correlation similarities are also intensity-independent and therefore also have potential to reduce the impact of some of these variations on clustering performance (Figure 1.11 b).

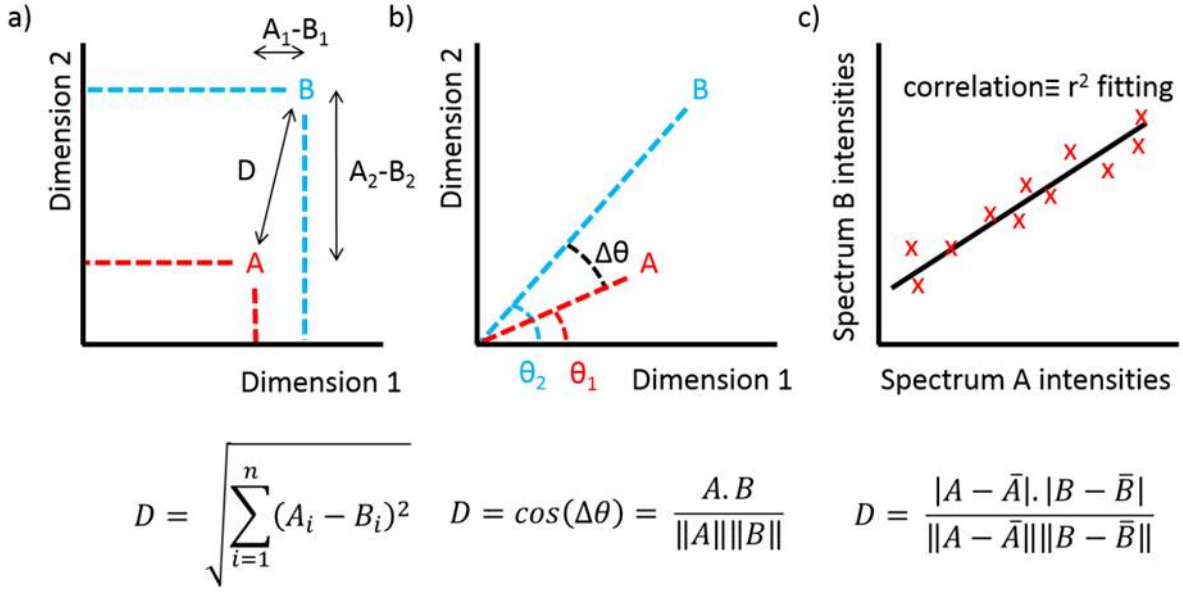

Figure 1.11: Visual representations of three of the distance metrics, a) Euclidean distance, b) cosine distance, and c) correlation.

Most applications of k-means clustering in MSI have used the Euclidean distance metric, and where normalisation has been used, TIC normalisation is most common [152, 153]. Most attempts to evaluate clustering results in MSI have used manual examination or comparison to complimentary modalities such as histological analysis [120]. Recently Oetjen et al. published a series of benchmark 3D datasets with histological information [154], and there are a number of publicly available datasets in repositories such as PRIDE and MetaboLights; however the limited chemical information provided by histology means that segmentations do not always match chemical information provided by MSI [152]. In order to accurately determine the most appropriate distance metric to choose, a sample of known chemical and spatial distribution would be required, which cannot be achieved from current MSI datasets.

K-means clustering

Due to its simplicity, relatively low computational requirements [155], and wide availability in many different languages [131], k-means clustering is one of the most popular

algorithms for clustering in MSI [129, 152, 153, 156, 157]. This can segment anatomies within different brain tissues [131, 133], distinguish tumour margins [158] and even intra-tumour heterogeneity [143]. Given a set of spectra, k-means clustering aims to partition the n spectra into k sets so as to minimize the intra-cluster sum of distances of each point in the cluster to its cluster centre. K-means clustering first starts by initialising the starting points of each cluster. This can be done in a number of ways, either by assigning a centroid as one of the data points within the sample, randomly assigning clusters to the data and calculating the mean of these, or by a clustering method itself, either k-means on a small subset of the data, or by hierarchical clustering (known as the Buckshot algorithm) [159]. Following this initialisation step, each point is assigned to its closest centroid as given by the distance metric specified. The new cluster centres are then recalculated as the mean of the newly formed cluster and the cluster assignment is repeated. These two steps are then repeated until convergence is achieved (Algorithm 1 and Figure 1.12). It is worth noting that k-means can be sensitive towards local minima and thus careful selection of the initialisation is required, as seen in figure 1.13 [160].

Input : Data matrix M of n pixels by d mass channels

Input : Initial centroid selection method i

Input : Distance metric m

Input : Number of clusters k

Input : Number of replicates r

Output: Vector length n of cluster assignment

for $i \leftarrow 1$ **to** r **do**

    Initialise k starting centroid locations c;

    while *not converged* **do**

        Assign each point to it's nearest centroid c;

        Calculate new cluster centroid as mean of it's data points;

    end

    Calculate performance $\sum_{p=1}^{k} \sum_{x \epsilon c_i} m(x, c_i)$ of this replicate;

end

Assign final clustering as the best result of all replicates;

Algorithm 1: K-means clustering algorithm

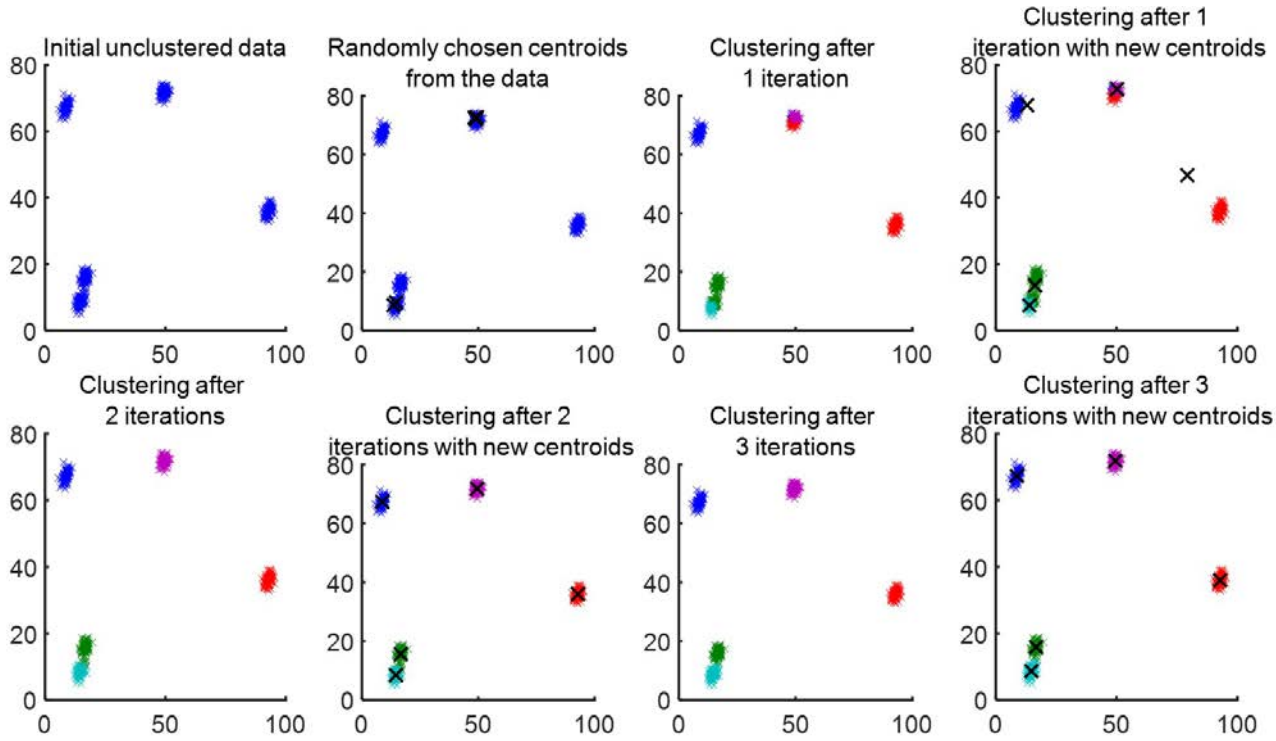

Figure 1.12: Iterative process of k-means clustering (k = 5) with initial centroids that leads to accurate clustering of the data. These data are comprised of five populations or normally distributed data that are separated from one another. In this case, these five populations are correctly clustered as separate from each other.

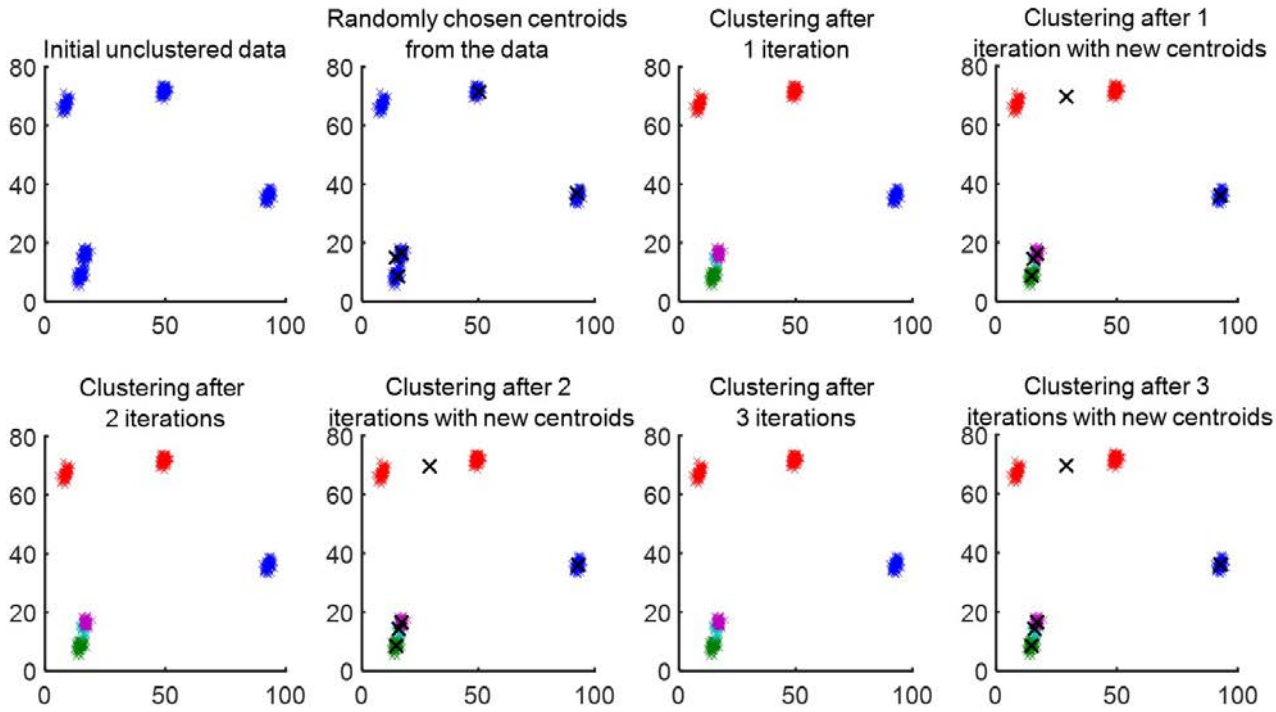

Figure 1.13: Iterative process of k-means clustering (k = 5) on the same data as figure 1.12, with initial centroids that leads to poor clustering of the data. In this case, the red cluster contains two of the populations of normally distributed data, and the green, cyan and purple clusters contain data from just two populations.

The earliest use of k-means clustering used in MSI was by McCombie et al. [131] who analysed two different tissues by clustering: mouse brain and rat head. These data were reduced to a specific m/z range of interest (4000-8500 Da), and then further reduced using PCA, followed by analysis by three clustering algorithms; hierarchical clustering, k-means clustering and Iterative Self-Organizing Data Analysis Technique Algorithm (ISODATA), a variant of k-means. In all cases the Euclidean distance metric was used. Muir et al. first investigated the use of the cosine distance measure for k-means clustering in MSI [156]. In this study, mouse cerebellum tissue was analysed in a mass range of 750-1200 Da. As with the previous example, a selection of multivariate tools were used to analyse the data, including PCA, linear discriminant analysis (LDA), multivariate analysis of variance

(MANOVA), as well as k-means clustering. This is one of the few examples of clustering in MSI data where the Euclidean distance was not chosen, however no justification for the choice of distance was presented in this study. Following these studies, k-means clustering was used as a comparison for edge-preserving denoising [121], and later a spatially aware clustering method by Alexandrov et al. [120]. This showed noisy results when using k-means clustering with the Euclidean distance metric with no normalisation applied, however in a more recent study, where TIC normalisation was applied it was noted that k-means clustering performed adequately [158]. In the two cases where k-means was determined to be unsuitable, the Euclidean distance measure was used, whereas in other studies the City block measure was chosen. This further limits any comparisons where different parameters and pre-processing were used within these experiments. Attempts were made to evaluate the clustering performance using the silhouette index, however they proved inconclusive [120]. K-means clustering has also been applied to cases where histological information is limited and so molecular imaging is vital [152]. In these cases, where there is knowledge of the sample, no comparison can be made to any underlying expected features. To overcome this, multiple multivariate tests were performed, and the degree of agreement between them was used to determine whether the result was correct or not. This is a promising means to determine erroneous results but can be difficult when these multiple tests do not concur with one another. K-means has also been applied with a variety of dimensionality reduction methods. Already mentioned were examples of PCA [156], peak picking [152], and binning [131]. Recently, memory efficient methods for performing PCA were developed, allowing larger datasets to be reduced, allowing k-means to be applied to large, even 3D MSI datasets [129]. In this example, the data was peak picked prior to PCA and the k-means was used with the Euclidean distance metric. Since this study was aimed at a demonstration of potential applications of the memory efficient PCA, no evaluation on the clustering result was performed. As, well as PCA, random projection has also been applied to MALDI-MSI data for dimensionality reduction [133, 157]. This was also clustered with k-means with the Euclidean distance.

Some comparison with anatomy was performed, however this was again refered to as a potential future application. K-means clustering has also been applied as a simple segmentation tool of tissue from background regions [153, 161]. In these cases, since differentiation between tissue and matrix is high, distance metric selection is likely to make minimal difference. In all of the studies where k-means clustering has been applied, there is little to no consistency between studies in parameters, either in image acquisition or processing. For example, some of the studies performed normalisation prior to analysis, whereas others did not. In addition to this, widely varying mass ranges, representing different molecules of interest, were used, as well as different matrix compounds to ionise these different molecules.

Bisecting k-means hierarchical clustering

Bisecting k-means hierarchical clustering involves recursively dividing the data into two by means of the k-means clustering algorithm [162]. After each division, the most appropriate cluster is selected and further divided into two until the desired number of clusters is reached. By only performing k-means clustering with two clusters at each step, the algorithm is more efficient than k-means clustering, with a complexity $O(n.d.\log_2 k.i)$ and so it is useful for clustering large datasets. The main difficulty when using the bisecting k-means algorithm is how to determine the appropriate cluster to divide. In the commercially available SCiLS Lab software this is selected manually by the user [163], however this can potentially introduce user bias into the data analysis. Other methods inlude selecting the largest cluster to divide or the cluster with the least overall similarity [164].

Input : Data matrix M of n pixels by d mass channels

Input : K-means parameters

Input : Number of clusters k

Input : Means of selecting which cluster to divide

Output: Vector length n of cluster assignment

Output: Hierarchy tree

```
for i ← 1 to k − 1 do
    if i == 1 then
        Divide the data in two using k-means (algorithm 1);
    else
        Divide the cluster determined in step 7 in two using k-means (algorithm 1);
    end
    Determine the next cluster to divide in two;
end
```

Algorithm 2: Bisecting k-means hierarchical clustering algorithm

Bisecting k-means clustering has primarily been used in clustering of large 3D MSI datasets [42, 165] but has also been used to analyse lipid compositions in different tumourous regions [166], or proteomic studies on various tissues [167]. The high efficiency over k-means clustering, along with the ability to produce a dendrogram to further interpret the data, gives the bisecting k-means algorithm significant advantages over many other clustering algorithms used in MSI such as standard k-means. This however comes with the added challenge of selecting which cluster to divide as mentioned above. Additionally, at any step, k-means may incorrectly segment the data into two, and this will impact the clustering results further down the tree.

Agglomerative hierarchical clustering

Agglomerative hierarchical clustering begins by constructing the pairwise similarity between the pixels within the data. This has a time complexity of $O(n^2.d)$ and a memory

requirement scaling of n^2 which is significantly higher than either k-means or bisecting k-means clustering. Once the pairwise similarity has been calculated, the most similar pixels are merged to form the prototype cluster, and a new distance to the remaining data is calculated. The means to determine the similarity of the merged data to the remaining data is known as the linkage method. There are a number of ways to define the linkage, and of these, four have been applied to MSI data; single, average, complete and Ward's [142, 168]. Given two clusters of data, the single linkage is the distance between the two closest points within each cluster (Figure 1.14), the average linkage is the mean of the distances between all points (Figure 1.15), and complete linkage the similarity between the two furthest points (Figure 1.16). Ward's linkage uses the difference in the variance between the individual clusters and the variance the data as one single cluster (Figure 1.17) [169]. This has been suggested as the most appropriate for clustering of mass spectrometry imaging data [168], however it is only applicable to clustering using the Euclidean distance metric, which, as discussed, may not always be appropriate for mass spectrometry imaging data. The complexity of linkage calculation is generally $O(n^2 \log n)$ but some cases can be faster; single and complete linkages can be calculated in $O(n^2)$ [170, 171], average in $O(k3^k n^2)$ [172] which can be more efficient when $k << n$.

Input : Data matrix M of n pixels by d mass channels

Input : Linkage method l

Input : Number of clusters k

Output: Vector length n of cluster assignment

Output: Hierarchy tree

```
Construct a pairwise distance matrix of similarityes D;
while number of clusters < k do
    Merge the two most similar points together;
    Calcutale a similarity of the merged data to remaining data using the linkage l;
end
```

Algorithm 3: Agglomerative hierarchical clustering algorithm

Figure 1.14: Single linkage, where the new distance is the smallest distance between all points in each cluster.

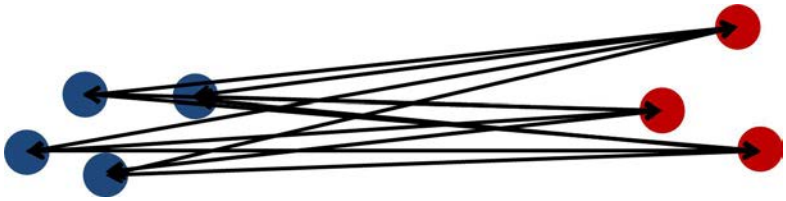

Figure 1.15: Average linkage, where the new distance is the average of all distances between data points in each cluster.

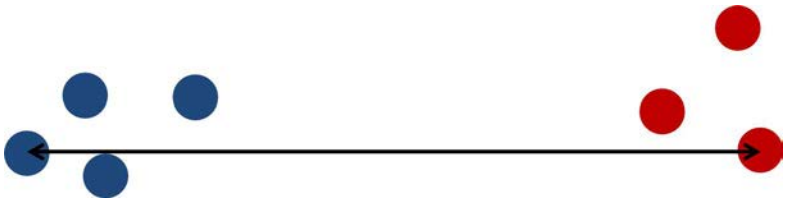

Figure 1.16: Complete linkage, where the new distance is the largest distance between all points in each cluster.

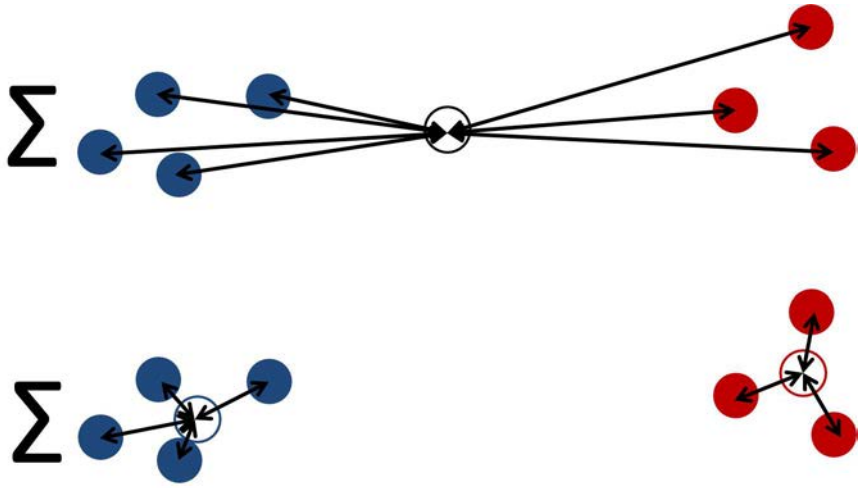

Figure 1.17: Ward's linkage, where the new distance is change in variance between the individual and combined clusters.

Agglomerative hierarchical clustering on MSI data has primarily been applied to cancer proteomic studies [39, 173, 174], and has been shown to be capable of differentiating cancerous from non-cancerous tissues [142], as well as differences within tumourous tissues [143]. As with bisecting k-means, the production of a dendrogram aids user interpretation of the data, and unlike k-means based algorithms, there are no random elements to this algorithm, thus improving its reliability if not necessarily its accuracy. The main limitation of agglomerative clustering is the high computational cost which limits it to datasets with small numbers of pixels.

Spatially aware and spatially aware structurally adaptive clustering

Spatially aware, and spatially aware structurally adaptive clustering are two algorithms introduced by Alexandrov *et al.* to overcome the problem of pixel variability in MSI [120]. These algorithms are similar to kernel k-means where the kernel function used is based on the spectra within the local spatial neighbourhood. In the non-structurally adaptive algorithm, for a given spectrum p of length d, a new concatenated spectrum is formed from all spectra within a given radius r, mulitplied by some weighting function (in the case of

Alexandrov *et al.* a Gaussian weighting was used) to create a new spectrum of length $d \times (2r+1)^2$. Following this, k-means clustering is applied to the new spatially weighted data. This increases the size of the data by a factor of $(2r+1)^2$ which would significantly increase the complexity of the k-means clustering. Therefore to reduce this, the FastMap algorithm [175] was used to perform dimensionality reduction prior to clustering (algorithm 4). In the structurally adaptive clustering, the weighting of the spectra applied in the kernel function is based on the similarity of the two spectra, where if the two spectra are more similar, a higher weighting will be applied. Alexandrov *et al.* demonstrated that these algorithms produce more accurate and less noisy clustering results by comparison with anatomical features and comparison to histological staining. The main limitation of this algorithm is that is requires the clustering to be performed with Euclidean based distance metrics which may not always be suitable for MSI data. As well as this, additional radius and weightings parameters must be specified by the user, which may have to be optimised for each dataset.

input : data matrix M of n pixels by d mass channels

input : Neighbourhood radius r

input : FastMap reduction dimension q

input : weighting function for neighourhood

input : k-means clustering parameters

output: Vector length n of cluster assignment

```
for i ← 1 to n do
    Initialise new spectrum s of length d × (2r + 1)^2;
    Create matrix N of spectra within r of pixel i;
    for j ← 1 to (2r + 1)^2 do
        if structurally adaptive then
            Create weighted spectrum w based on the similarity of N_j to pixel i;
        else
            Create weighted spectrum w of based on N_j weighted by its distance to
              pixel i;
        end
        Concatenate spectrum w onto s;
    end
    Add the fully concatenated spectrum s into new data matrix D;
end
Perform the FastMap transform on D to give a new matrix F with the desired
  number of dimensions q ;
Perform k-means clustering on F;
```

Algorithm 4: Spatially aware and structurally adaptive clustering algorithm

Spectral clustering

Spectral clustering is a graph theory based clustering algorithm that performs k-means clustering on eigenvalues of the graph Laplacian of the data (algorithm 5) [176]. For a graph with n vertices, The Laplacian L is an $n \times n$ matrix representation of a graph, and can be formulated in a number of ways. Typically, the Laplacian is described as follows;

$$L_{i,j} := \begin{cases} deg(v_i) \text{ if } i = j \\ -1 \text{ if } i \neq j \text{ and } v_i \text{—} v_j \\ 0 \text{ otherwise} \end{cases}$$

where $deg(v_i)$ is the degree of number of adjacent vertices (where vertices represent connected nodes on the graph) to the vertex v, and v_i—v_j vertex v_i is adjacent to vertex v_j if they have a similarity above a given threshold. The random walk normalised Laplacian used by Palmer for spectral clustering in MSI [177] is described as follows;

$$L_{i,j} := \begin{cases} 1 & \text{if } i = j \text{ and } deg(v_i) \neq 0 \\ \frac{-1}{deg(v_i)} & \text{if } i \neq j \text{ and } v_i \text{—} v_j \\ 0 & \text{otherwise} \end{cases}$$

input : data matrix M of n pixels by d mass channels

input : k nearest neighbours

input : weighting function for graph generation

input : Number of eigenvectors to cluster on e

input : k-means clustering parameters

output: Vector length n of cluster assignment

1 Construct a weighted graph W where $W_{i,j}$ represents the similarity between spectra i and j;

2 Perform eigendecomposition of W to get eigenvectors V and eigenvalues v;

3 Sort V in ascending order of the corresponding eigenvalues;

4 Perform k-means clustering on the V_{2-e+1};

Algorithm 5: Spectral clustering algorithm

The main issue with spectral clustering, as described by Palmer is the need to select an appropriate number of nearest neighbours for constructing the graph Laplacian [177]. Using too few will not connect adjacent vertices, and using too many will connect vertices which are not truly adjacent. In addition to this, the number of eigenvectors to perform the clustering on must also be selected by the user. In addition to these graph based clustering specific parameters, other parameters such as the number of clusters, and the measure of spectral similarity must also be selected by the user. While many of these parameters may be inferred from the data by a user with expertise in statistics, each additional requirement reduces the ease of use to a mass spectrometrist without a statistical background. Despite this, spectral clustering and other graph based clustering methods hold exciting opportunitites to overcome the issue of high dimensionality in mass spectrometry imaging.

DBSCAN

DBSCAN is a density based clustering algorithm that groups together dense regions of data points and assigns outliers in low density areas of data [178]. A threshold similarity

ϵ is defined, and pixels are classified as "core" if they have more than a user specified number of other pixels within the distance ϵ. All pixels that are within ϵ of this pixel are then defined as directly reachable from this pixel. Any other pixels that can be directly reached from the directly reached pixels through a path of core points are then defined as reachable, and all others defined as outliers. DBSCAN then groups together all directly reachable and reachable pixels and assigns the remainder as outliers (algorithm 6). Unlike many other algorithms, DBSCAN does not require a user specified number of clusters, merely a threshold similarity ϵ and minimum number of pixels to define core points. It can also handle the concept of outlier pixels thereby preventing the possible skewing of statistical analysis of the resulting clusters, and can group arbitrarily shaped clusters. The major limitations of DBSCAN are that it can be very sensitive towards ϵ and minimum points parameters, and without *a priori* knowledge of the data these can be hard to estimate. If ϵ is too low or minimum points too high, all data will be considered as outlier, and if ϵ is too high or minimum points too low, then all data will belong to one single cluster. In addition to this, MSI data have a very high dimensionality which results in a large degree of sparsity in the data (distances in high dimensional space converge to infinity) further increasing the difficulty in estimating ϵ.

input : data matrix M of n pixels by d mass channels

input : ϵ neighbour threshold

input : minimum data points for core definition $minPoints$

output: Vector length n of cluster assignment

```
1  cluster assignment c = 1;
2  for i ← 1 to n do
3      if pixel i ∼ assigned then
4          determine the pixels neighbours within ϵ of pixel i;
5          if size(nNeighbours) < minPoints then
6              assign pixel i to outlier class;
7          else
8              assign pixel i to cluster c;
9              for j ← 1 to nNeighbours do
10                 if neighbour(j) ∼ assigned then
11                     assign pixel neighbour(j) to cluster c;
12                     determine pixels neighbours′ in pixel neighbour(j);
13                     if size(neighbours′) > minPoints then
14                         add neighbours′ to neighbours;
15                     end
16                 end
17             end
18         end
19     end
20 end
```

Algorithm 6: DBSCAN clustering algorithm

FLAME

FLAME is another density based clustering algorithm based on each data point's neighbourhood. In this algorithm, each point's density is determined as the sum of distances to it's k nearest neighbours (Figure 1.18). Data points that have higher density than all of its neighbours are then defined as cluster support objects, and given fixed membership to itself (Figure 1.19). Any data points which have lower density than their neighbours, and lower than a user specified threshold are assigned as fixed outliers. All other data points are then assigned as cluster member objects and given equal membership to all cluster support objects as well as the outlier group. The membership of all cluster member objects are then updated iteratively by a linear combination of it's nearest neighbours until convergence is reached (Algorithm 7, Figure 1.20).

This, like DBSCAN offers the advantages that the number of clusters need not be specified, and can consider data points as outliers. In addition to this, only two parameters, the number of nearest neighbours and a threshold density for outliers needs to be set. The number nearest neighbours chosen must be selected carefully as they will dramatically affect the end result. Ideally this value should be equal to and no larger than the expected size of the clusters in the data. However, clusters may vary in size and the size of clusters may not be know prior to analysis. If a value for k is selected that is too low, the data will fragment into many small clusters, and if k is too high, too few cluster support objects will be determined.

input : data matrix M of n pixels by d mass channels

input : k nearest neighbours for density estimation

input : density threshold t

output: Vector length n of cluster assignment

Construct a weighted graph of each points distance d to it's k nearest neighbours;

for $i \leftarrow 1$ **to** n **do**

    determine density of point n_i as $\sum_{l=1}^{k} d_k$;

    classify point d by the following rules;

- Cluster support object: data has higher density than all it's neighbours
- Cluster outlier: data has lower density than all it's neighbours and lower than threshold t
- Cluster member object: all other data

end

Assign each cluster support object to have singular membership to itself;

Assign each outlier to have fixed membership to the outlier group;

Assign all cluster members to have equal membership to all cluster support objetcs and the outlier group;

while *not converged* **do**

    update all data points membership as a linear combination of it's k neighbours

end

if *Single membership wanted* **then**

    Assign each data point to it's highest membership cluster

end

Algorithm 7: FLAME clustering algorithm

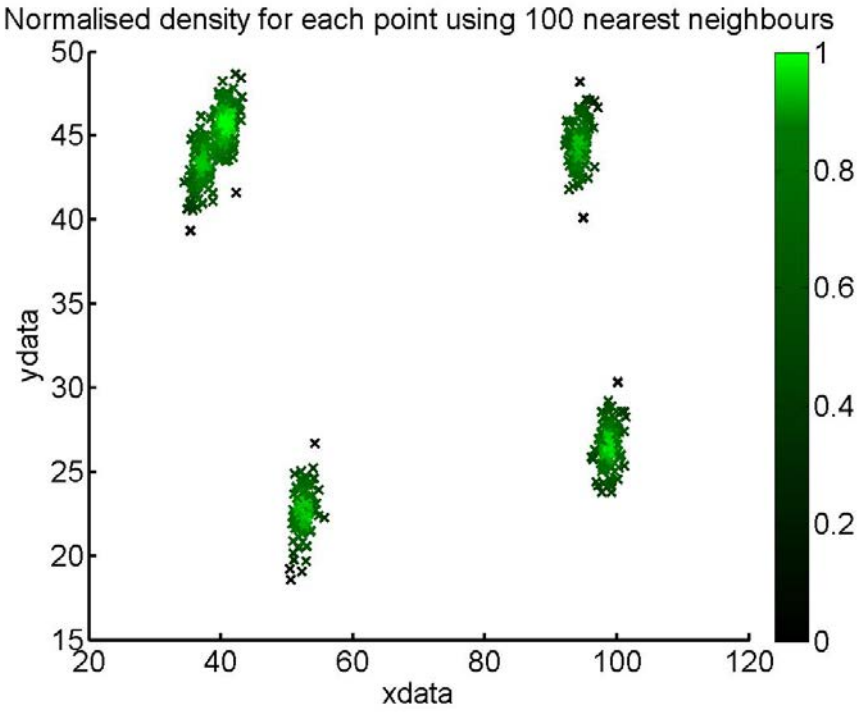

Figure 1.18: Five normally distributed clusters of data containing 100 samples each. The density of each data point is shown when 100 nearest neighbours are used for FLAME clustering.

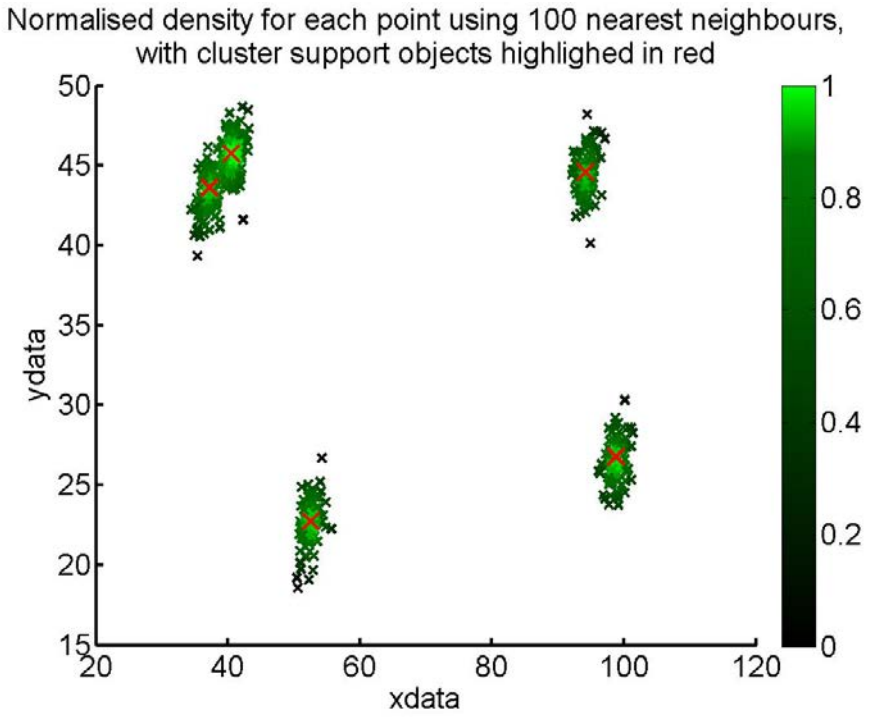

Figure 1.19: Density of data points in figure 1.18 when 100 nearest neighbours are used, with each of the five cluster support objects highlighted in red.

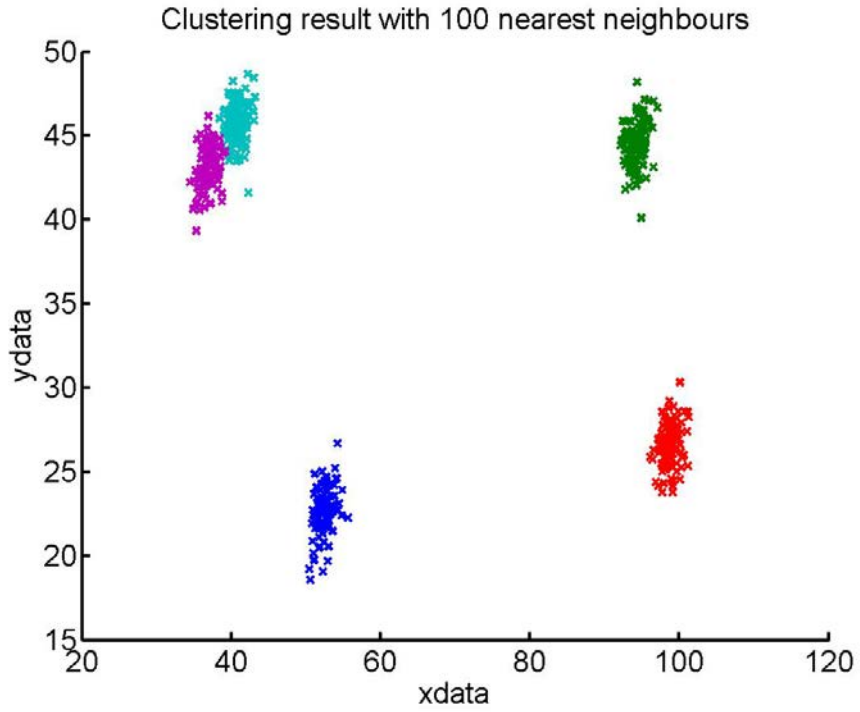

Figure 1.20: Final FLAME clustering result on the data in figure 1.18 when 100 nearest neighbours are used.

Affinity propagation

Affinity propagation is a clustering algorithm that uses the principle of "message passing between data points" to determine "exemplar" points which are good representations of other points [146]. Considering a pixel i and a potential exemplar j it iteratively updates two matrices; the "responsibility" matrix which determines how well j acts as an exemplar for i, and the "availability" matrix which measures how well j acts as an exemplar for i considering all other possible exemplars for i (algorithm 8). Like agglomerative hierarchical clustering, a pairwise similarity between all data points S is first calculated, using some distance metric d, and an initial measure of responsibilites can be assigned. The main advantage of this algorithm over others is that little to no user defined parameters are required, thus no *a priori* knowldege of the data is needed. This comes with a high compuational cost however as the time complexity of affinity propagation scales as $O(n^2.d + n^2.i)$ where n is the number of pixels, d the spectral channels and i the number

of iterations before convergence, and pairwise distance matrices scale by n^2.

input : pairwise distance matrix S size $n \times n$

input : Optional initialisation of responsibilities

output: Vector length n of exemplar locations

Initialise responsibility matrix R and availability matrix A;
while $\sim$ *converged* **do**
  for $i \leftarrow 1$ **to** n **do**
    for $j \leftarrow 1$ **to** n **do**
      update $R(i,j) = S(i,j) - \max_{i \neq j}\{a(i,j) + s(i,j)\}$;
    end
  end
  for $i \leftarrow 1$ **to** n **do**
    for $j \leftarrow 1$ **to** n **do**
      if $i \neq j$ **then**
        update $A(i,j) = \min(0, R(j,j) + \sum_{i \notin \{i,j\}} \max(0, R(i,j)))$;
      else
        update $A(i,i) = \sum_{i \neq j} \max(0, R(i,j))$;
      end
    end
  end
end
for $i \leftarrow 1$ **to** n **do**
  if $R(i,i) > 0$ **then**
    pixel i is an exemplar
  end
end

Algorithm 8: Affinitiy propagation algorithm

Additional graph theory based methods

Along with spectal clustering, there are many variants on the graph based clustering algorithms, based principally on the idea of dividing the data while preserving the maximum number of connected vertices [147, 148, 179]. The popular normalised cuts algorithm uses the second smallest eigenvector of the graph to bipartition the data into two [147]. This done by either segmenting the data either side of the median, the mean, zero, or a user specified number into the two clusters. When greater than two clusters are desired, the normalised cuts algorithm can be used to further segment the subsets of data in the same manner as the bisecting k-means clustering algorithm. As with bisecting k-means, this introduces the issue of how to appropriately select the data to be further segmented, with the additional issue of which method to use for partitioning the eigenvectors. It is desireable for end users to have the minimum number of parameters to be specified to reduce the number of variables involved in an algorithm, thereby increasing the ease of use. To remove the necessity for specifying the number of nearest neighbours for graph construction, a full weighted graph can be constructed, where the Laplacian contains the degree of similarity for all vertices [180]. This does however come at an increase in the computational cost, as the Laplacian is no longer sparse. Since these algorithms are based around connectivity, they are ideally suited towards non Gaussian shaped clusters, and high dimensional data.

Two-phase k-means clustering

Two-phase k-means clustering is an algorithm developed by Pham *et al.* [149] for clustering large databases that cannot be loaded into random-access memory (RAM). Small subsets of the data are loaded into RAM, and clustered via the k-means algorithm. The cluster centroids of this data, and all other subsets form a compression set which is also clustered using by k-means. Since clustering is only performed on a small subset of the data, the RAM requirement and computational complexity is significantly reduced. In addition, the clustering on all subsets after the first one can be initialised using the cluster centroids from

the first subset clustering. This means that the clustering will have a higher probability of converging in fewer iterations than when random seeding is used, thereby further reducing the complexity. This is of particular interest in MSI since datasets are rapidly increasing in size with new instrument developments and three dimensional MSI.

input : data matrix of n pixels by d mass channels
input : desired number of subets s
input : K-means clustering parameters
output: Vector length n of cluster assignment

```
Randomly assign the data into s equal sized subsets S_{1-s};
for i ← 1 to s do
    load into RAM data assigned to subset S_i;
    cluster S_i into k clusters using the k-means algorithm 1;
    if i == 1 then
        create a compression set, a compressed representation of subset S_i from the
         cluster centroids;
    else
        add clusters centroids of subset S_i to the compression set;
    end
end
cluster the compression set using the k-means algorithm 1;
assign the cluster identities from the compression set to the pixels from subsets
```

Algorithm 9: Two-phase k-means clustering algorithm

Clustering evaluation

There are many different methods for quantitatively evaluating the success of clustering, which can be divided into two types; internal and external. Internal evaluation uses the intrinsic properties of the clustering result, usually by comparing the data within each cluster to the data outside of the cluster [181]. Examples of internal evaluation measures

include the Calinski-Harabasz index, which compares the sum of inter to intra cluster distances (equations 1.1 to 1.3) [182], the silhouette index which measures the similarity of a data point to all others within the cluster, compared to the other clusters (equation 1.4) [183], and the Dunn index which compares the minimum inter-cluster distance to the maximum intra-cluster distance (equation 1.5) [184].

For a dataset, with n pixels, segmented into k clusters, with each cluster c containing n_k pixels, the Calinski Harabasz index (CH_{index}) is defined as;

$$CH_{index} = \frac{S_{intra}(n-k)}{S_{inter}(k-1)} \tag{1.1}$$

where the inter cluster sum of squares;

$$S_{inter} = \sum_{i=1}^{k} ||m_i - m||^2 \tag{1.2}$$

and;

$$S_{intra} = \sum_{i=1}^{k} \sum_{x \in c_i}^{n_k} ||x - m_i||^2 \tag{1.3}$$

Where m_i is the cluster centre for cluster i, and x is a spectrum from cluster c.

The silhouette index (s_i) for a given pixel is defined as;

$$s_i = \frac{b_i - ai}{max\{a_i, b_i\}} \tag{1.4}$$

where a_i is the average dissimilarity of pixel i to the data within its own cluster, and b_i is the lowest average dissimilarity of pixel i to any other cluster.

The Dunn index (D_{index}) is measured by;

$$D_{index} = \frac{min_{1 \leq i \leq j \leq k} d(i,j)}{max_{1 \leq x \leq k} d'(x)} \tag{1.5}$$

where $d(i, j)$ is the distance between clusters i and j, and $d'(x)$ is the intra cluster distance of cluster x.

Previous attempts to evaluate clustering in MSI have used the silhouette [185,186] and Calinski-Harabasz [187] indices but these have proven inconclusive at best. This could be in part due to the fact that most internal evaluation measures also rely on a measure of spectral similarity, and primarily use a Euclidean based metric for this. This means that comparison between distance metrics is not possible as the metric used for the clustering should reflect that used to evaluate the result. In addition to this, the high dimensionality of MSI data leads to a high degree of sparsity, and the curse of dimensionality, and thus severly hinders the effectiveness of these evaluation metrics.

External evaluation on the other hand compares the clustering results to known ground truths to determine the true and false positive and negative results (table 1.5). Using this information, values such as sensitivity $\frac{true\ positive}{\sum positive}$ and specificity $\frac{true\ negative}{\sum negative}$ can be calculated alongside validation measures such as the Rand index $\frac{true\ positive + true\ negative}{total\ population}$ (often referred to as accuracy) and Jaccard index $\frac{true\ positive}{true\ positive + false\ positive + false\ negative}$ [188]. Since the comparison is to known information there is no concern of bias towards a given algorithm or distance metric and so can be used as a method for accurately and reliably comparing and evaluating clustering algorithms or workflows. The main limitation of external evaluation is the need for a ground truth to compare against. Since MSI is generally used as an exploratory tool, usually on biological samples, most datasets will not have a ground truth and thus these external evaluations are usually not possible [189].

		Measured Result	
		Positive	Negative
Ground truth	Positive	True positive	False negative
	Negative	False positive	True positive

Table 1.5: Summary of true and false positive and negative measures for external evaluation.

1.5 Introduction to this thesis

In this introduction, the principle of mass spectrometry imaging, all the way from fundamentals of desorption and ionisation, through to data interpretation and multivariate analysis are described. The remaining chapters of this thesis describe experimental and computational methods to study fundamental processes in MSI, with a focus on providing a better understanding of MSI data.

In chapter 2, multivariate analysis is applied to the investigation of the relationship between different laser parameters in raster mode MALDI imaging experiments. In addition to this, good practice guides are put forward for both instrumental setup and experimental design for larger scale fundamentals studies in tissue imaging.

Chapter 3 tackles the increased complexity of ion mobility based separation, and provides novel and exciting methods to investigate the effects of different FAIMS field parameters using multivariate analysis. Furthermore, deconvolution methods and their challenges for LESA-ESI based protein imaging with the addition of FAIMS separation are investigated.

Chapter 4 goes on to investigate the need for the means to evaluate many of these multivariate analysis techniques, and details the generation of printed ink standards as a basis for defined spatial patterns to test different external evaluation metrics for clustering algorithms. Different metrics for both internal and external clustering evaluation are investigated, along with the use of multivariate normality testing to better understand the underlying distribution of MSI data to appropriately select distance metrics to use when clustering these data.

The multivariate normality from chapter 4 is then used as the basis for statistical modelling to generate synthetic MSI data by resampling from multivariate normal distributions in chapter 5. This is then combined with simulations of instrument variances to generate controllable variabilty to allow quantitative evaluation of aspects of MSI data procesing.

Synthetic data generated using the statistical modelling in chapter 6 are then used to

evaluate the clustering algorithms described in this introduction. A novel efficient and accurate graph based clustering algorithm combined with the subset sampling approach of two-phase k-means is then described for segmentation of large MSI datasets.

Finally, chapter 7 describes the conclusions of this thesis, and outlines area for future work in these areas.

CHAPTER 2

MULTIVARIATE ANALYSIS OF MALDI FUNDAMENTALS

2.1 Introduction

Despite the widespread and popular use of MALDI MSI as an imaging technique, there is still a lack of full understanding of the fundamental processes involved in MALDI ion formation [49]. Further complicating this, there are a huge number of variables involved in MALDI MS image formation, from laser parameters such as fluence and repetition rate, to instrument variables such as m/z range, the mass analyser used, and ion accumulation time. Consequently, the study of these effects is particularly difficult, especially where many of these variables are not independent of one another and so must be studied at the same time.

The work presented in this chapter aims to develop experimental approaches, alongside multivariate analysis methods to study variables involved in laser operation when performing raster mode MALDI-MSI. When considering variables in laser operation, there are a number of different parameters to consider. The laser energy incident upon the sample per unit area, or fluence, has been previously determined to have a power law relationship with detected ion intensity $I \propto H^m$ where I is the detected intensity, H the laser fluence, and m a fitting parameter [55]. Below a threshold fluence, little to no ions are detected, followed by a rapid increase in intensity up until a plateau at higher energies [56].

While seemingly simple to control such as through variable attenuating filters, the laser fluence is affected by a number of variables, such as the size of the laser spot incident on the sample, the energy delivered per pulse, the number of laser pulses incident on the sample over the given sampling time, which will in itself be determined by both the laser pulse repetition rate and the sampling rate of acquisition. Current commercially available MALDI ion sources have little to no control or feedback on these experimental variables. For example, energy delivered is typically controlled on an arbitrary 0-100% attenuation scale, with no reference to how this relates to Joules of energy delivered per pulse, a variable that has been shown to be particularly influencial in MALDI-MSI of biological tissues [56]. Furthermore, there is no measurement of this energy at any point, meaning that fluctuations in the laser are neither accounted for, nor even observed. This means that even if the laser energy is first optimised on an extra piece of tissue (requiring an additional tissue section), there is no guarantee that the energy delivered to the sample to be imaged is the same as this. The effect of the incident fluence on the ion intensity of tissue samples has also been to vary sigmoidally with increasing fluence, and therefore any small variations in fluence can have a drastic effect on the ion intensity observed [56]. In addition to this, there are often limited options available to alter the laser pulse repetition rate, and the effect of this has been the subject of a few studies but never with a controlled and monitored fluence output [58, 59]. Laser spot size in MALDI is often constrained by the method of delivery such as by fibres or lenses. Recently, Steven *et al.* published methods for determination of laser spot size based on a fluorimetric method, allowing this variable to be measured in most systems, if not controlled [190].

This chapter details experimental setup to control and monitor these effects, as well as using multivariate methods to investigate the relationship between laser parameters of fluence and repetition rate, along with the stage speed in raster mode MALDI imaging.

2.2 Experimental

Materials

HPLC grade Methanol was purchased from Fisher Scientific (Leicestershire, UK), and water was purified using an ELGA Purelab Option system (Marlow, UK). Trifluoroacetic acid (TFA) (99.9 % purity) and CHCA were purchased from Sigma Aldrich (Dorset, UK). Stainless steel MALDI target plates for MSI from Sciex (Ontario, Canada) were used for all samples. Energy per pulse was measured using a pyro-electric sensor (PD10-C, Ophir Photonics), and where stated, a 2.5 cm diaphragm shutter, with 10 ms full aperture opening time (Thorlabs Ltd, Ely, UK) was used.

Tissue preparation

Mice were sacrificed humanely at [redacted] in accordance with the Home Office Animals (Scientific Procedures) Act 1986. Mouse brain was flash frozen in liquid nitrogen immediately after excision. Tissue sections ($10\mu m$ thick) were collected and thaw mounted onto stainless steel MALDI imaging plates (ABSciex, Warrington, UK). Sectioning was performed on a CM 1850 Cryo-microtome (Leica, Milton Keynes, UK). All nine tissue sections for the repetition rate raster speed experiment were mounted onto a single stainless steel MALDI imaging plate (ABSciex, Warrington, UK).

Matrix application

In all experiments, blank or tissue containing sample plates were sprayed with 5 mg/mL CHCA in MeOH/H_2O/TFA (80/20/0.1), using automated spray deposition (TM Sprayer^{TM}, HTX Technologies, Carrboro, NC), at a nebuliser temperature of 90 ^{o}C, a solvent flow rate of 0.115 mL/min, a gas pressure of 10 psi, and a nebuliser speed of 1333 mm min^{-1}. Eight sequential passes across the whole plate were used, each with a spacing of 3 mm between lines, even passes were performed horizontally, and odd passes vertically, and an

offset of 1.5 mm was used on passes 3, 4, 7, and 8. This gave a density of matrix on the plate of 0.115 mg cm^{-2}.

Mass spectrometry

For the investigation into repetition rate and raster speed, a separate, serial section of mouse brain was used for each set of laser and stage conditions therefore nine serial sections of coronal mouse brain were analysed in total. The image of each hemisphere collected separately, thus acting as an internal control, giving a total of eighteen images collected. The order of data acquisition for all eighteen hemispheres of brain were randomised, and between each change in repetition rate, the laser was allowed to equilibrate for 30 minutes. Images were then acquired using an AB Sciex QSTAR XL QqTOF instrument with an oMALDI 2 ion source using Analyst QS 1.1 with oMALDI server 5.1 (ABSciex, Warrington, UK), at a range of stage speeds (slower, medium and fast), using an Nd:YVO_4 laser (Elforlight: SPOT-10-100-355, Daventry, UK), operated at repetition rates of 2, 6.6666, and 20 kHz, and energies of 1, 2, and 3 J per pulse. All images were acquired with pixel sizes of 300 μm in x and 100 μm in y, using a declustering potential 1 of 0 V, a focusing potential of 60 V, and a declustering potential 2 of 20 V, and an m/z range of 50-1000.

Data processing

Data processing was performed on an Intel Xeon quad core CPU E5-2637 v2 (3.50 GHz) with 64 GB of RAM. All data were converted from AB Sciex proprietary file format (.wiff) to .mzML using AB MS Data Converter (version 1.3; Sciex) and then converted to imzML using imzMLConverter [118]. These were then analysed using a combination of SpectraAnalysis software [191], and Matlab (version R2014a and statistics toolbox, The Math Works, Inc., Natick, MA, USA). The spectra were preprocessed as follows; QSTAR specific zero filling, three passes of sequential Savitzky-Golay smoothing, window size 7, polynomial order 2, and negatives in the spectra removed. In some cases, where stated,

TIC normalisation was also applied. The mean spectrum was then peak picked using a gradient method, and the top 2000 peaks retained for the shutter work, and top 5000 for the repetition rate, raster speed studies.

2.3 Results and Discussion

In order to study fundamental aspects of laser characteristics in MALDI, a laser setup is required that allows for controlling and monitoring these characteristics throughout the course of an experiment. To achieve this, a laser setup was used where a small portion ($\sim 1\%$) of the photons from the laser are diverged perpendicular to the path by the use of a low reflectivity mirror (Figure 2.1). An energy meter is then placed in this path, thereby allowing a continual online monitoring of the laser energy delivered throughout the course of a MALDI-MSI experiment.

2.3.1 Energy measurement calibration

To accurately measure the energy delivered to the sample inside the mass spectrometer, a calibration between the energy delivered to the sensor from the beam splitter, and the energy delivered in source must be performed. This is done in a two-step process, firstly the energy that passes through both the low reflectivity mirror, and through the fibre optics are measured. This allows for calibration of the percentage of energy delivered through the mirror, along with losses observed through the fibre optics (Figure 2.2). Following this, the losses through the optics in the instrument itself were measured by removing the ion source lens setup and measuring the subsequent loss through this system. Once this has been performed, the energy meter can remain in place, perpendicular to the low reflectivity mirror, thereby allowing for a constant monitoring of the in source energy delivered throughout the MS experiments.

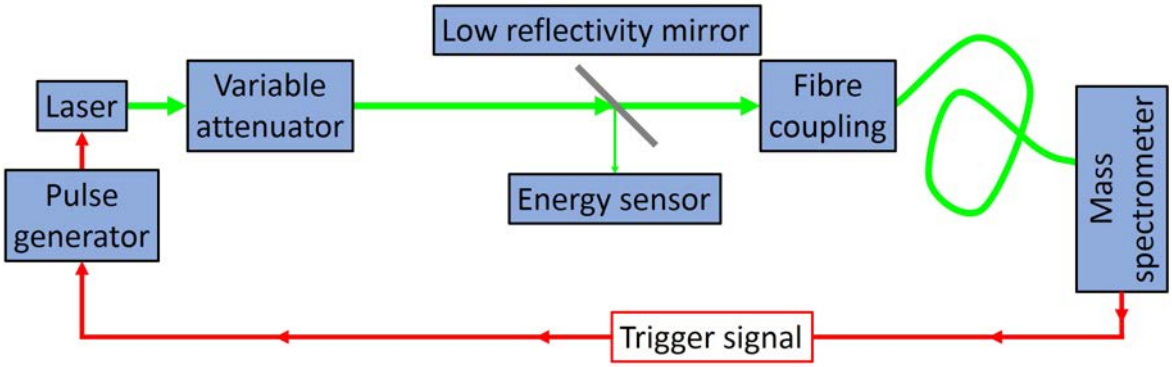

Figure 2.1: Schematic of the laser setup, where a small amount of laser photons ($\sim 1\%$) are directed onto an energy sensor via a low reflectivity mirror.

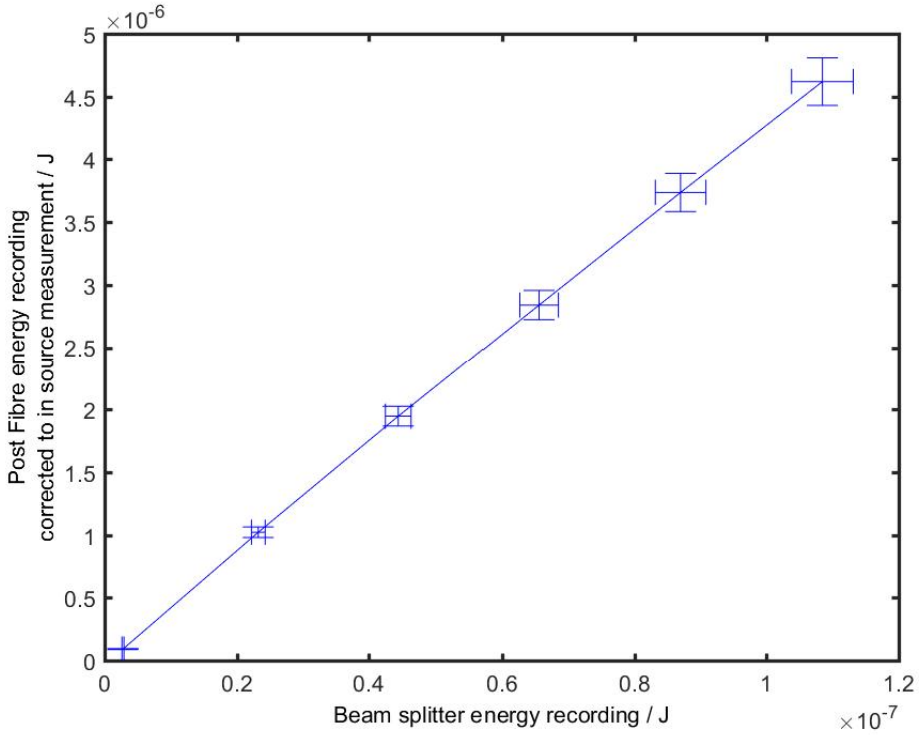

Figure 2.2: Errorbar plot of the energy delivered to the online measurement via the beam splitter compared to through the fibre optics. The horizontal and vertical errorbars represent the standard deviation of the energy per pulse delivered over ∼5000 laser shots.

The calibration of energy delivered through the mirror relative to through the fibre optics, corrected for the loss of energy through the optics of the QSTAR instrument

yielded a calibration of energy in source(J) = energy meter reading(J) $\times 37.14 + 10.44$ nJ (Figure 2.3). This allows the energy delivered in source to be measured by correcting the measured energy through the beam splitter using this formula.

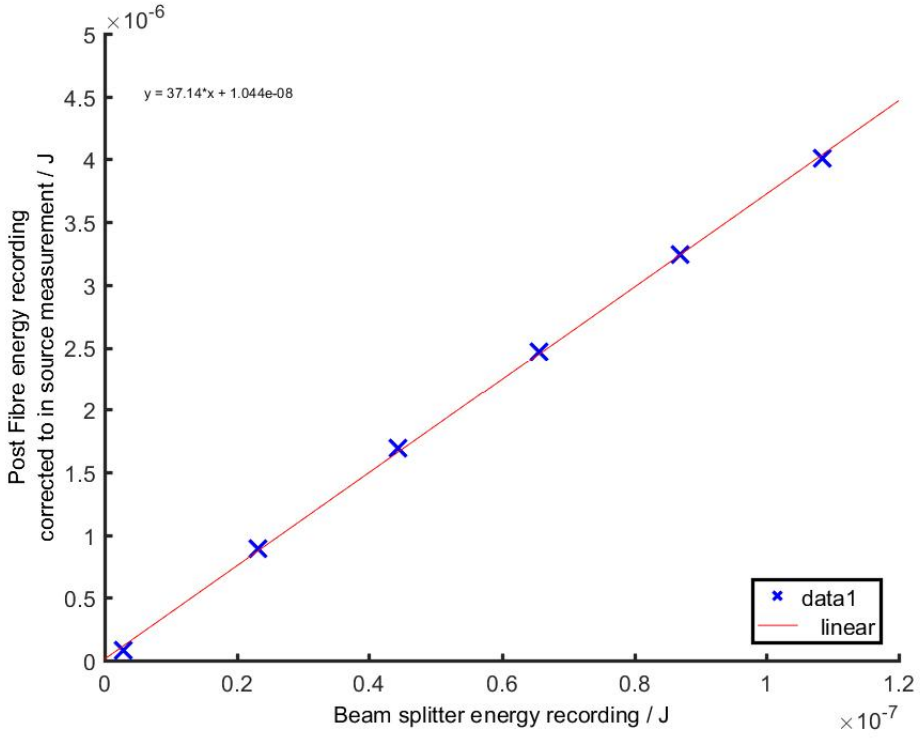

Figure 2.3: Plot of the energy delivered to the online measurement via the beam splitter compared to through the fibre optics (Figure 2.2, corrected for the loss of energy through the QSTAR optics). This was then fit with a linear regression to give the correction formula.

2.3.2 Controlling energy stability

When operating Nd:YAG diode-pumped solid-state (DPSS) lasers at relatively high repetition rates (> 3kHz), a "spiking" phenomenon was observed in the energy delivered (Figure 2.4). The initial amount of energy delivered was much higher, slowly dropping to a steady state after $\sim 10^4$ laser pulses (around 30s of acquisition time)(Figure 2.4). In tissue imaging experiments operating in raster imaging mode, this timeframe corresponds approximately to the length of a single raster line acquisition. Shutters have been em-

ployed in some MALDI experiments, such as Westmacott *et al.*'s [62] use of a mechanical shutter to reduce the variability of the laser between pulses when studying the effects of fluence in MALDI-MS. In other laser based MSI techniques such as laser ablation - inductively coupled plasma (LA-ICP), it is considered good practice to maintain the laser at steady state operation, and trigger sampling of a surface through opening and closing of a shutter rather than triggering of the laser itself [192]. The use of a similar shutter based acquisition (Figure 2.5) was investigated for MALDI-MSI experiments, and the effects on the resulting spectra were investigated through univariate and multivariate analysis.

When sampling a thin film of CHCA, variation in the TIC can be seen to mirror the variations measured in the energy that is delivered to the sample (Figure 2.7). In comparison, when the shutter is used to maintain steady state lasing conditions, the TIC is much more stable (Figure 2.8), and matches the online energy output measurement (Figure 2.6).

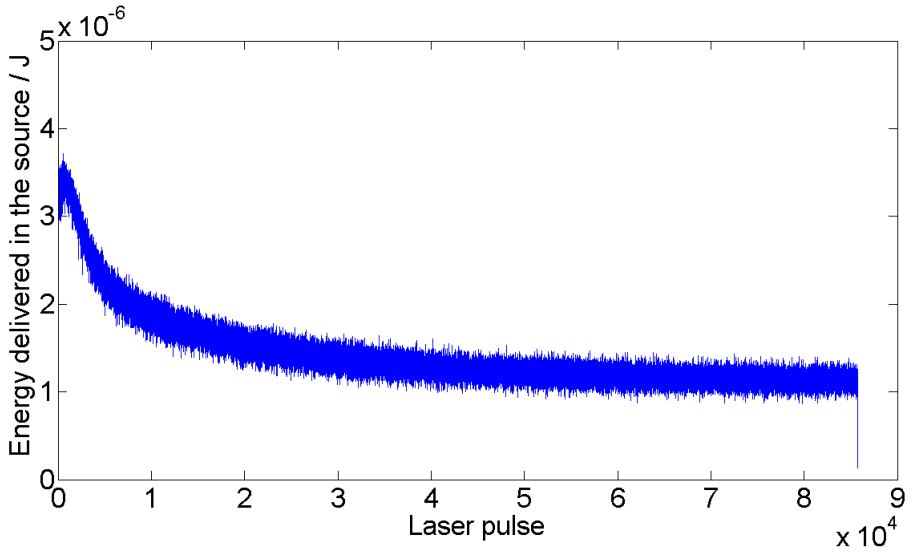

Figure 2.4: Energy delivered across a single raster line without the use of a shutter to maintain a steady state of operation. There is a much higher amount of energy delivered in the first 10,000 pulses than in the subsequent ones.

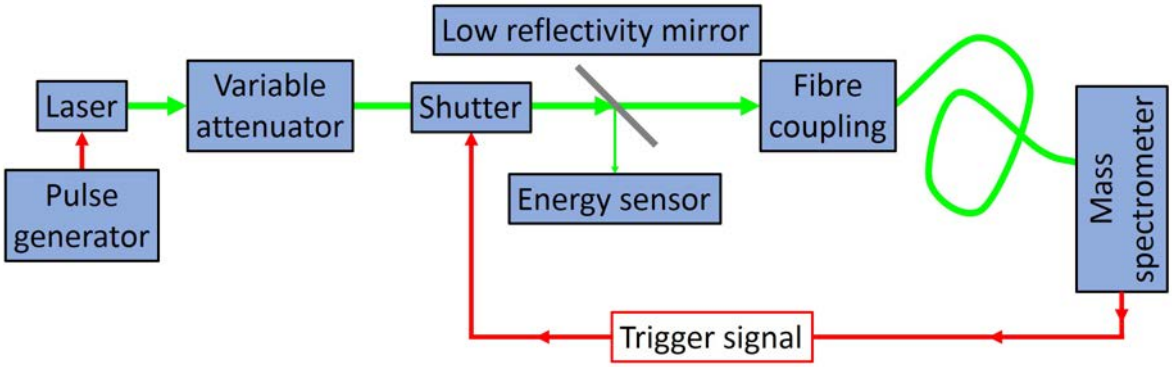

Figure 2.5: Modified laser setup from that shown in Figure 2.1, where the laser is permanently firing, and the instrument triggers the opening and closing of the shutter.

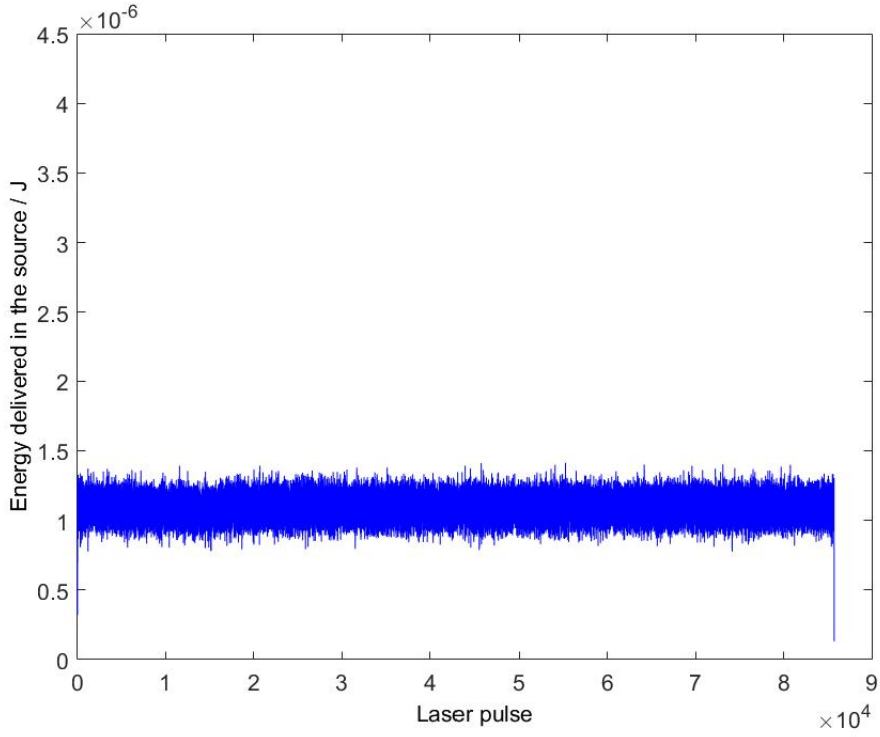

Figure 2.6: Energy delivered across a single raster line including the use of a shutter to maintain a steady state of operation, and triggering opens and closes a shutter. When the laser is allowed to maintain at a steady state operation, the energy delivered to the sample across a single raster line is much more consistent than the energy observed in figure 2.4 when triggering fires the laser itself. This is because the optics and lasing medium have been allowed to reach an equilibrium state.

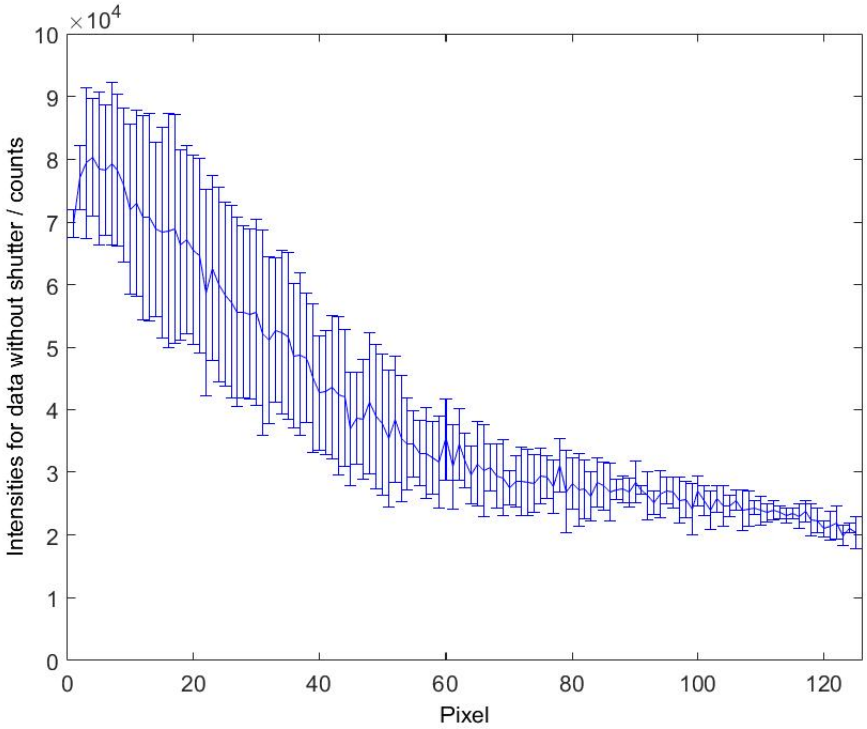

Figure 2.7: Errorbar plot of the TIC from the four raster lines of CHCA analysed by MALDI without the use of the shutter. The TIC reflects the changes in laser energy observed in Figure 2.4. Errorbars represent the standard deviation from the mean of the four raster lines.

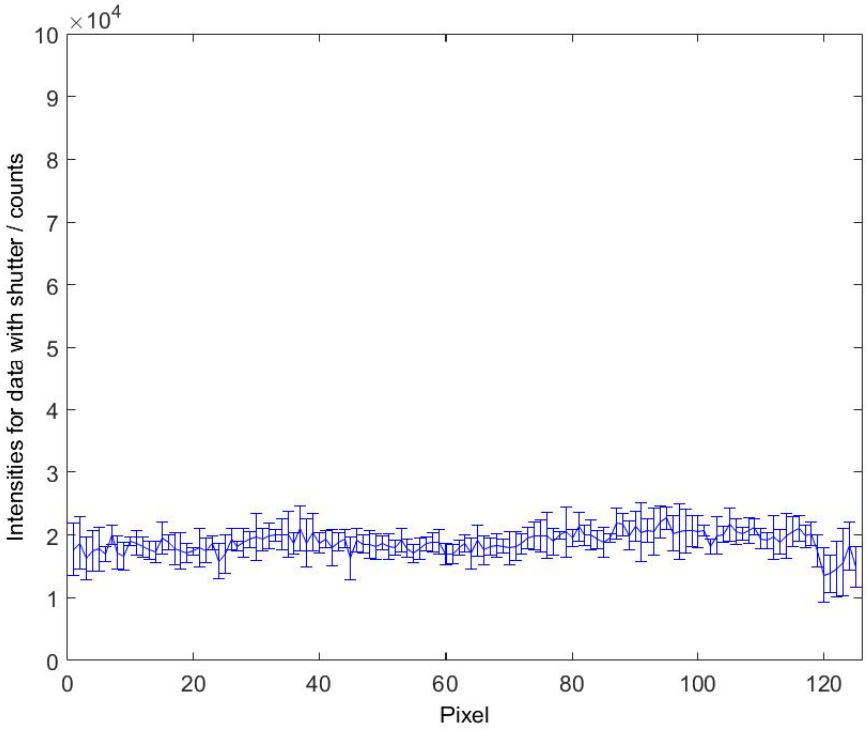

Figure 2.8: Errorbar plot of the TIC from the four raster lines of CHCA analysed by MALDI with the use of the shutter. The TIC is much more stable than in figure 2.7 and reflects the stable laser energy observed in Figure 2.6. Errorbars represent the standard deviation from the four raster lines.

To further characterise what changes are occurring as a result of energy changes with and without the shutter, PCA was applied to these data before and after TIC normalisation was applied. PCA is used to determine where the largest sources of variance lie within the data, in order to relate these back to the fundamentals of what may be occurring. The first principal component will highlight the largest source of variance within the data, with subsequent components showing the remaining variance in order, under the constraint of orthogonal axes. Prior to TIC normalisation, principal component 1 clearly shows a diffference in the data with and without the use of the shutter, but these differences are purely an intensity change in these regions (Figure 2.9). Since all principal components must be orthogonal to all previous ones, the large intensity change in component one may skew all subsequent ones, therefore it is appropriate to account for

this before analysing the higher components. If these effect of the changing energy were purely an intensity change, then it could be removed by TIC normalisation. To further test this, PCA was applied following TIC normalisation, and the result of this "spiking" can still be clearly observed (Figure 2.10). An increase in lower m/z ions can be seen from the data acquired during this spiking, likely as a result of increased fragmentation caused by the higher laser energy.

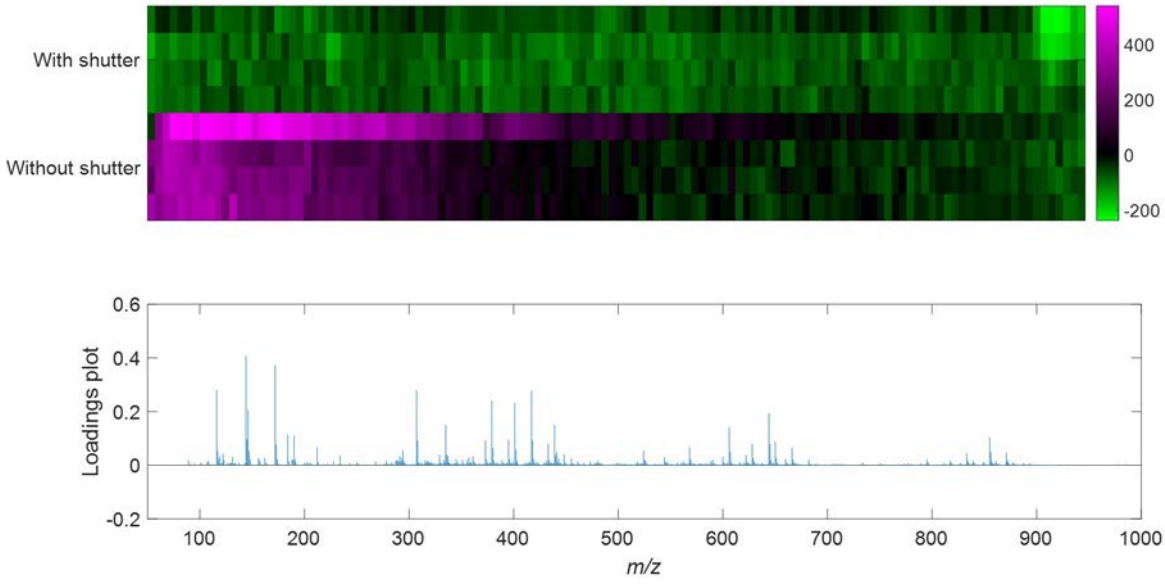

Figure 2.9: Principal component 1 from CHCA thin film the data acquired with and without the use of the shutter. The effect of the laser "spiking" can clearly be seen in the positive areas of the scores images at beginning of the raster lines, and the spectral loadings show this as an intensity increase in these regions.

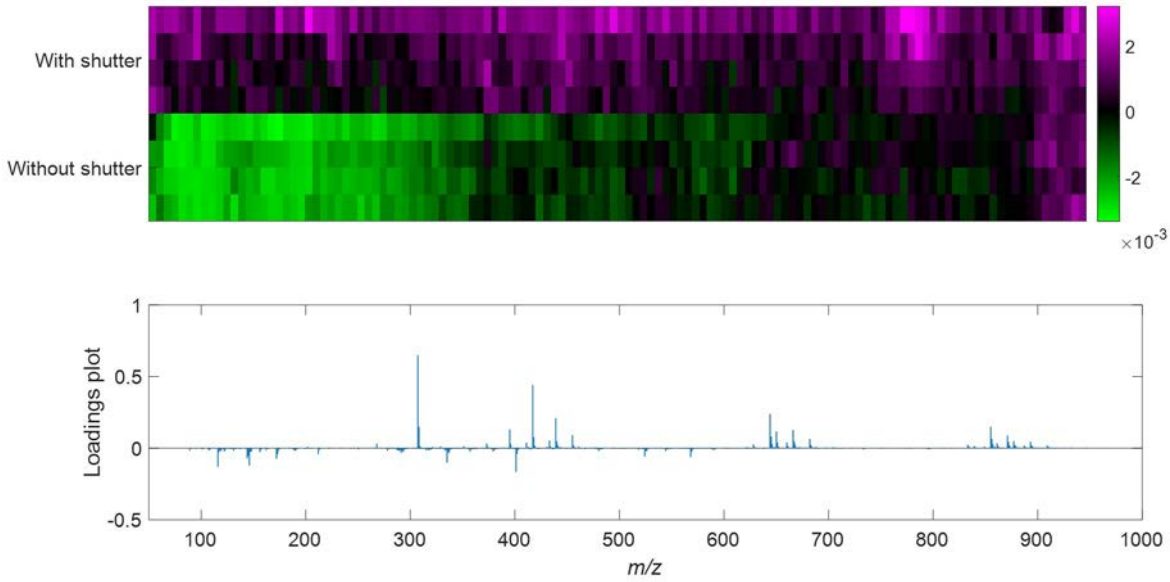

Figure 2.10: Principal component 1 on the CHCA thin film with and without the use of the shutter when TIC normalisation is applied. The scores images still separate the regions of laser energy "spiking" indicating that these are not just intensity changes in these regions.

While characterisation of these effects in thin films is important, it is necessary to investigate their effects in tissue samples as well, as these are the most common sample to be analysed by MALDI-MSI. When acquiring data on a QSTAR instrument, square image regions must be selected, and so when imaging a full tissue section, the first pixels in a raster line will be just matrix. Therefore, half a coronal mouse brain section was analysed, with the raster line beginning in the central line of the brain. Serial sections were analysed with and without the use of the shutter to ensure the maximum biological similarity within the sections, and minimise any possible biological variation within the sample.

As with the thin film data, the "spiking" phenomenon was observed across a raster line, and changes in ion intensity were observed corresponding to the laser energy (Figure 2.11). The ratios between different ions such as m/z 56 (PC head group fragment), 184 (PC head group), and 798 (intact lipid), reveals that the effect of the laser energy on acquired in

tissue is also not merely an intensity change (Figures 2.12 and 2.13). Principal component 2 shows clear difference between the shuttered and non-shuttered data, as an intensity difference between these data (Figure 2.14), with PC 1 separating tissue from matrix. However as with the thin film data, following TIC normalisation, principal component 4 still shows variance between the shuttered and non-shuttered data, primarily resulting from changes between the low mass fragments ion and intact lipid species (Figure 2.15).

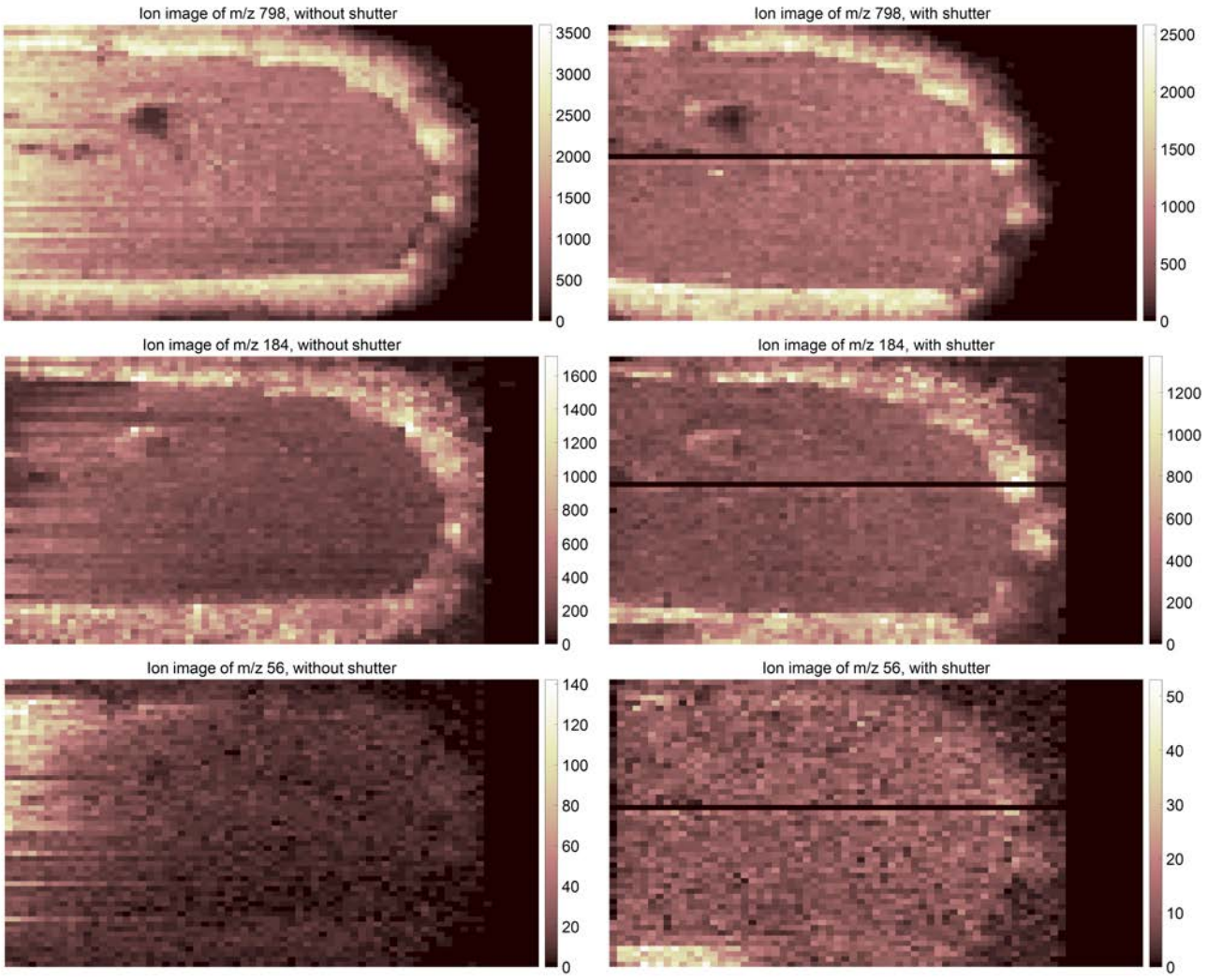

Figure 2.11: Ion images for the peaks at m/z 798, 184 and 56, with and without the use of a shutter. The ions m/z 798 and 56 have much higher intensities during the "spiking" region of the non-shuttered data, whereas the peak at m/z 184 is much lower. In comparison, the ion intensities of the data acquired with the use of a shutter is much more homogeneous.

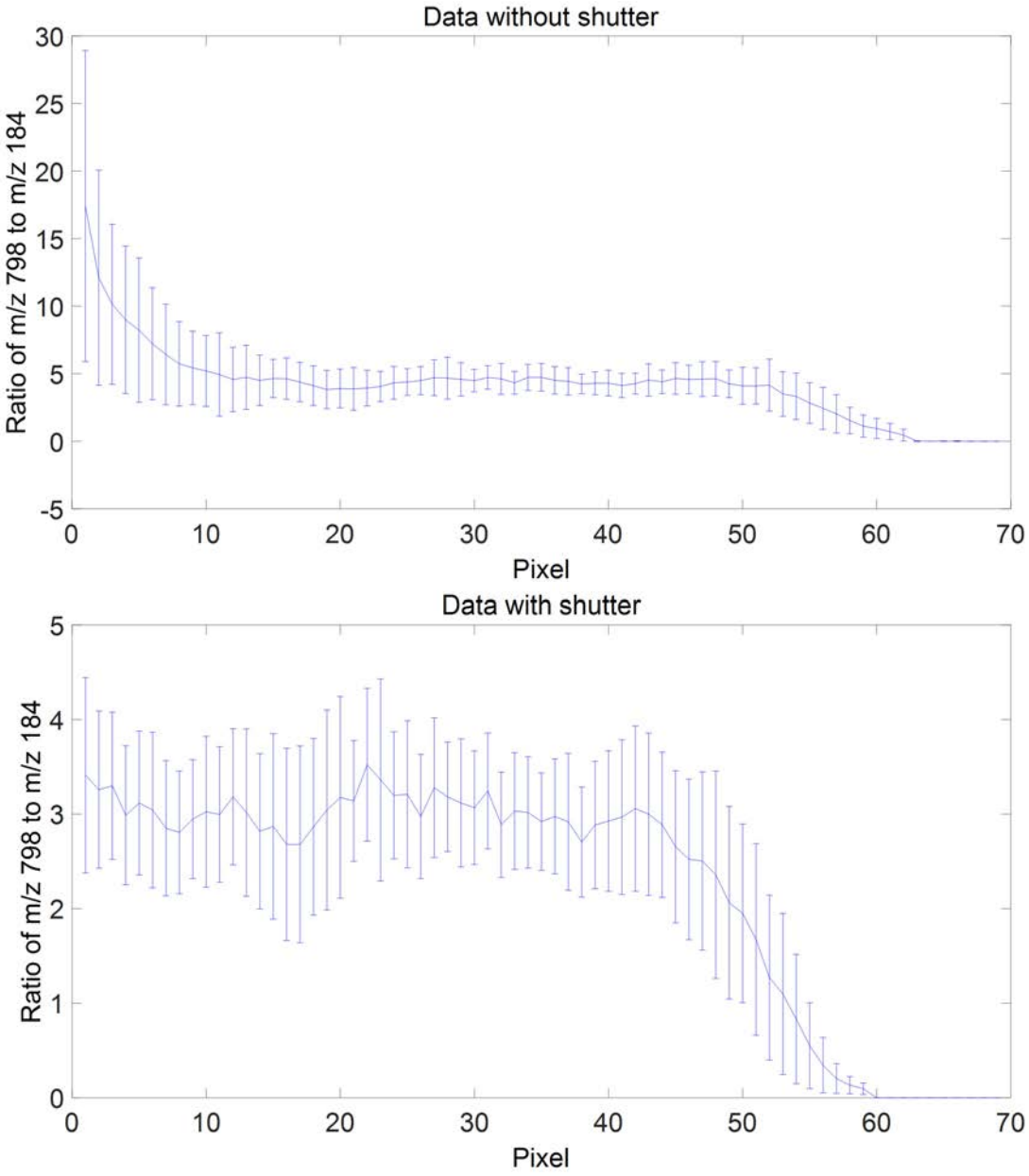

Figure 2.12: Errorbar plots for the ratio of ions m/z 798 to 184 across six raster lines of the images in figure 2.11. There is a large variance in this ratio in the shuttered data as compared to the data acquired with the use of the shutter. Errorbars are the standard deviation from the mean. The ratio changes a lot because the changes in energy delivered when the shutter is not used result in different levels of fragmentation of the parent m/z 798. Since the laser is not in equilibrium without the shutter, there is also a greater amount of variance in energy delivered resulting in the larger error bars.

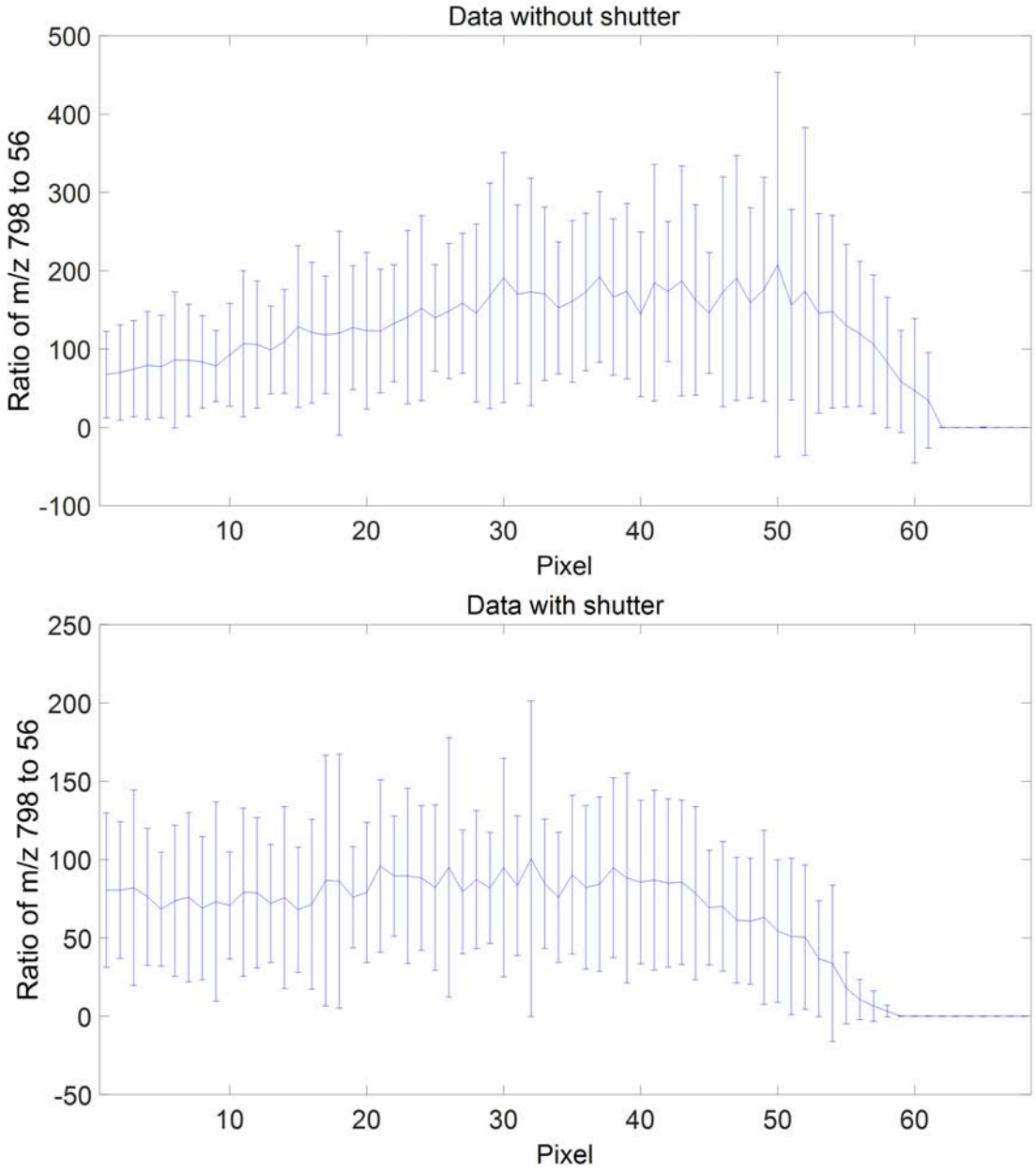

Figure 2.13: Errorbar plots for the ratio of ions m/z 798 to 56 (without normalisation) across six raster lines from the centre of the images in figure 2.11. There is a larger variance in this ratio in the unshuttered data as compared to the data acquired with the use of the shutter. Errorbars are the standard deviation from the mean.

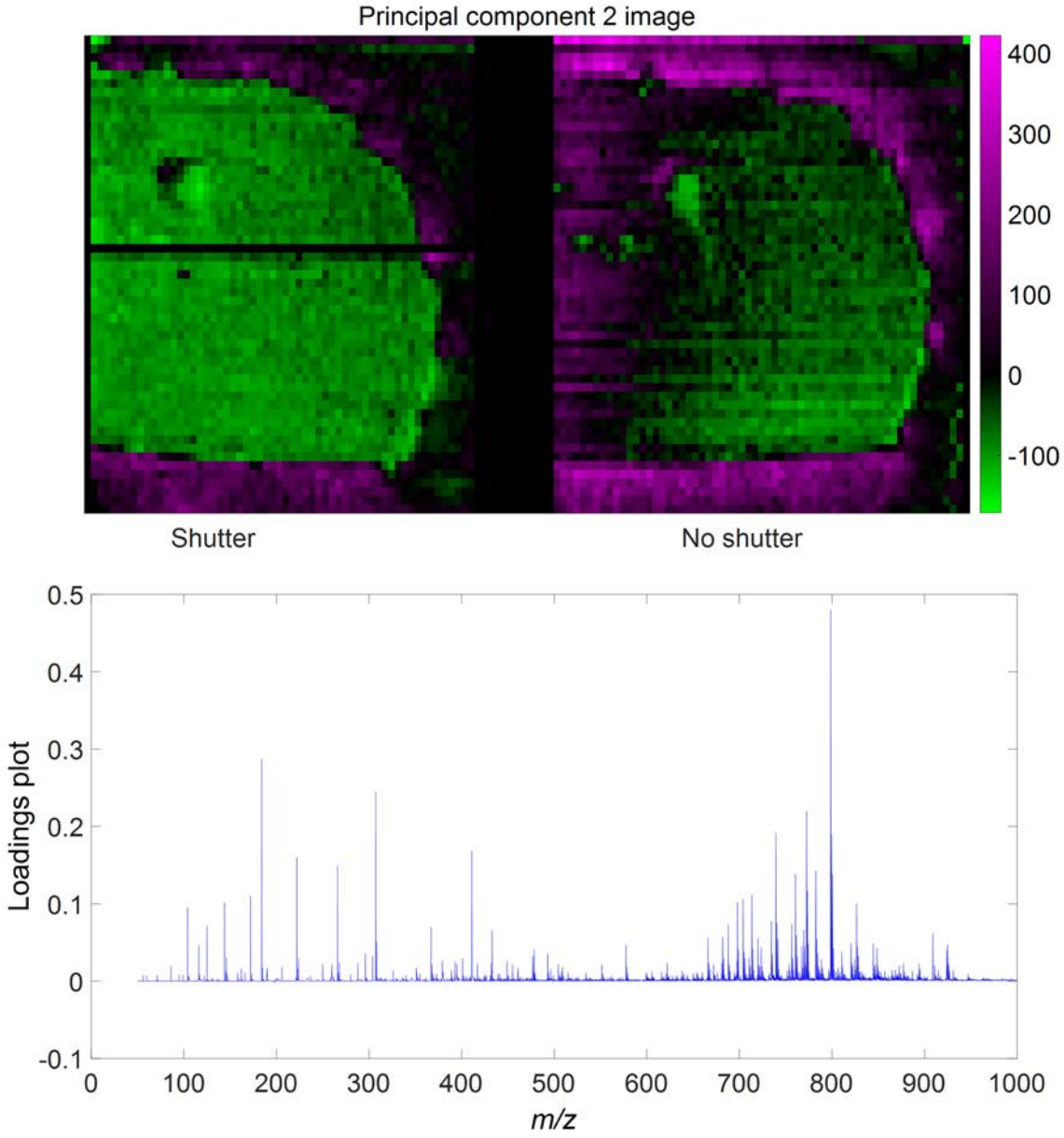

Figure 2.14: Principal component 2 on the tissue data with and without the use of a shutter. The variance between these data, as seen by the scores image, is in the initial "spiking" region of laser firing, and the loadings plot shows this as an intensity increase in these regions. This is likely due to a greater amount of energy being delivered in these regions resulting in more ion formation.

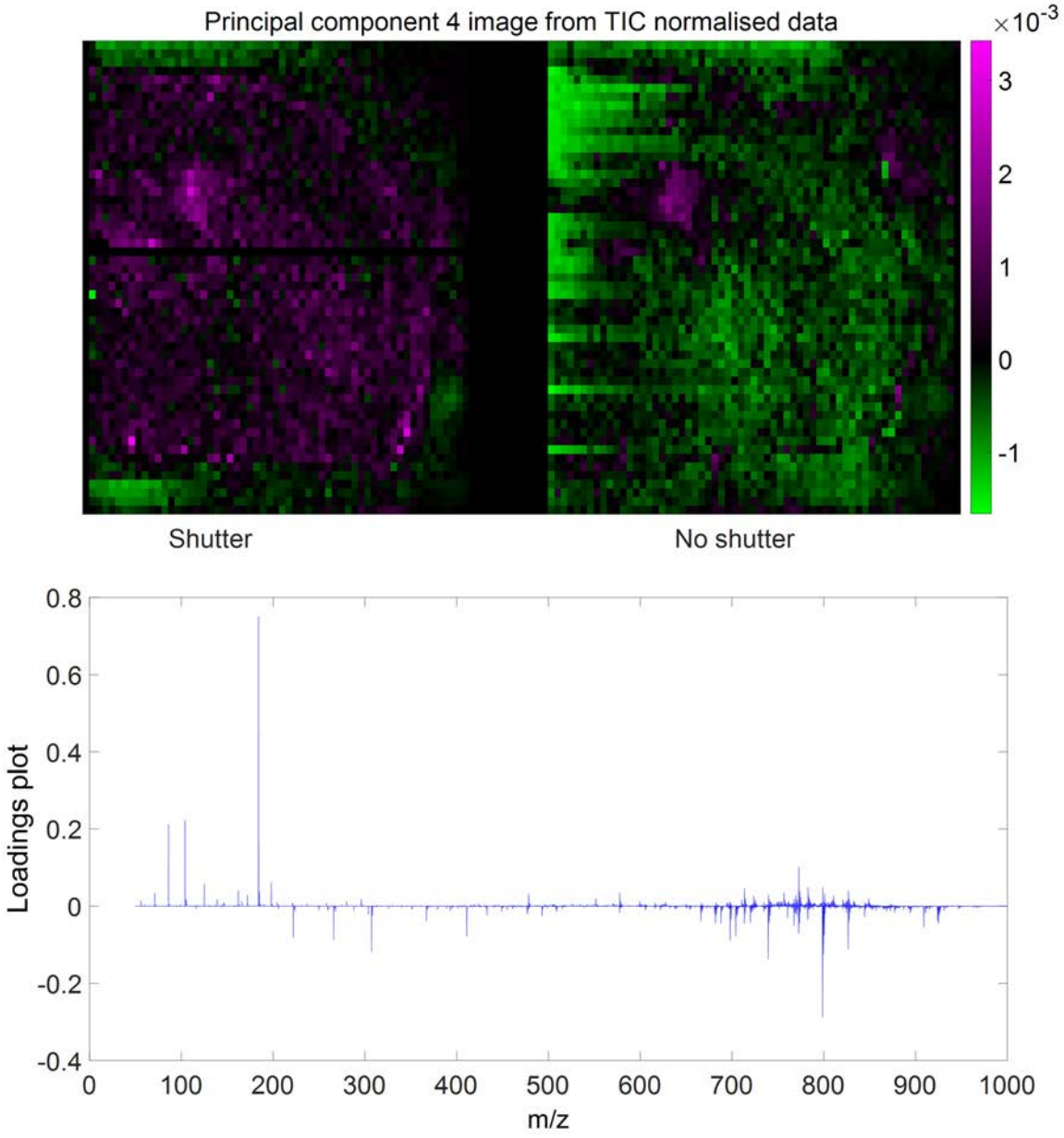

Figure 2.15: Principal component 4 on the tissue data with and without the use of a shutter after TIC normalisation has been applied. There remains a source of variance from the initial "spiking" region of laser firing but this is no longer an intensity change, and can be mainly seen by a difference in the ions at m/z 798 and 184.

2.3.3 Investigating the effect of repetition rate, stage raster speed and fluence

One of the main developments in throughput in MSI has been the use of high repetition rate neodynium based lasers [58, 59]. These typically operate at repetition rates between 1 and 20 kHz. Operating at higher repetition rates will deliver more laser shots to the sample for the same amount of time, and by also moving the stage faster, the same number of shots per pixel could be achieved, but with faster imaging speed. Acquiring data at higher repetition rates and faster stage speeds does not necessarily mean that the resulting data will be the same however, and so in order to acquire data at these faster speeds, it is vital to understand more of the relationship between the stage speed and repetition rate.

Energy delivered at varying repetition rate

The maximum available energy that a laser can achieve will be dependent on the repetition rate employed. Using an Nd:YAG laser, a sharp drop off in energy is observed above 1 kHz (Figure 2.16). In comparison, when employing the Nd:YVO_4 laser, at repetition rates below 3 kHz there is significantly less energy available, but there is a much shallower drop off in energy, and so above 3 kHz there is more available energy than with the Nd:YAG laser. In addition to this, the maximum repetition rate available when using the Nd:YVO_4 is 20 kHz, compared to 10 kHz using the Nd:YAG. Based on these data, it is more useful to carry out the study into the effect of repetition rate using the Nd:YVO_4 laser as it offers a higher useable energy across a wider repetition rate range.

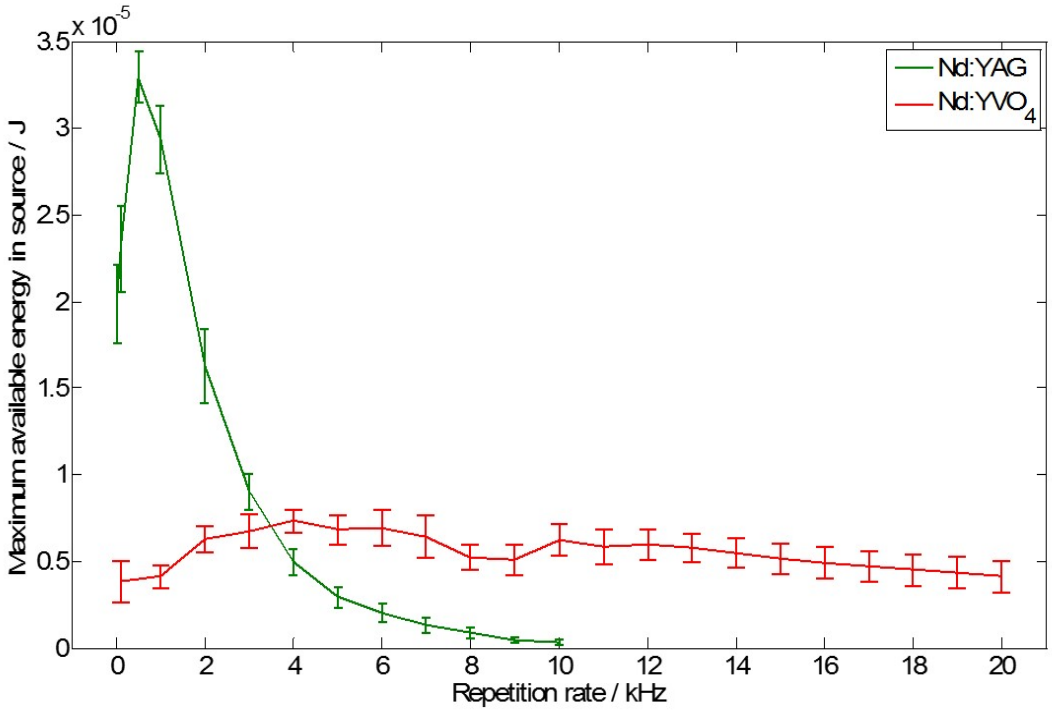

Figure 2.16: Maximum energy per pulse that can be delivered to the sample using the Nd:YAG and Nd:YVO_4 laser at several repetition rates.

Controlling laser fluence

Along with repetition rate and stage speed, the laser fluence will also have a significant effect on the resultant ions formed. As shown however, for the same attenuation of the laser, at different repetition rates, the energy delivered to the sample can vary significantly (Figure 2.16). In previous studies into the effect of repetition rate on ion intensities, laser fluence was not fixed, and optimal repetition rates were found to correspond with the highest fluence outputs from the laser [58]. In order to control this, for each following experiment, the laser was attenuated and monitored at each repetition rate such that the energy delivered to the sample was the same. In addition to this, the shutter system described above was employed to maintain steady state operation of the laser.

Investigation into repetition rate, stage raster speed and fluence

To study the effects of repetition rate, stage speed and laser fluence, in tissue imaging, the following conditions were used. Tissue sections were imaged at all combinations of stage speeds of (0.3 and 3.5 mm s^{-1}), repetition rates of 2 and 20 kHz, and laser fluences of 1 and 3 μJ, as well as one pair of images at 1 mm s^{-1}, 6 kHz and 2 μJ. These represent a range of possible combinations of these three variables, such as high and low repetition rates, fast and slow stage speeds, and threshold ionisation energy and plateau ionisation energy [56]. In addition to this, there are conditions whereby a similar amount of energy is delivered per pixel, by operating at either fast stage speed and high repetition rate, or slow speed and low repetition rate (Table 2.1). Nine serial coronal mouse brain sections were taken, one for each set of imaging conditions. To act as an internal quality control throughout the experiment, both hemispheres were imaged separately, and the order of acquisition was randomised throughout. This allows comparison of the two hemispheres to check for potential experimental variance that may have occurred throughout the course of the experiment.

Repetition rate / kHz	Stage speed /$mm\ s^{-1}$	Laser energy / μJ	Energy per pixel
2	0.3	1	0.002
2	0.3	3	0.006
2	3.5	1	0.000171
2	3.5	3	0.000514
6.66666	1	2	0.00400
20	0.3	1	0.02
20	0.3	3	0.06
20	3.5	1	0.00171
20	3.5	3	0.00514

Table 2.1: Combinations of repetition rates, stage speeds, and laser energies employed to perform imaging studies, and their corresponding energy delivered per pixel for these conditions.

From an initial univariate analysis, it can be seen that there is a complex interrelationship between the detected ion intensity for a given m/z and the stage raster speed, laser fluence and laser repetition rate. As expected, ion intensity increases with fluence, but the extent of this increase is largely dependant on the repetition rate and stage speed, and is dependant on the m/z of interest (Figures 2.17 to 2.20). The m/z dependence can be due to differences in the degree of fragmentation that occurs for different molecules at varying fluence. For example, the peaks at m/z 184, and 739 are typically associated with lipid fragments and exhibit much lower intensities at 1 μJ than 3 μJ at almost all raster speeds and repetition rates (Figures 2.19 and 2.20). In comparison, the peaks at m/z 798 and 826 are from intact lipids, and have very similar intensities at both 1 and 3 μJ when operating at the fast raster speed (Figures 2.18 and 2.17).

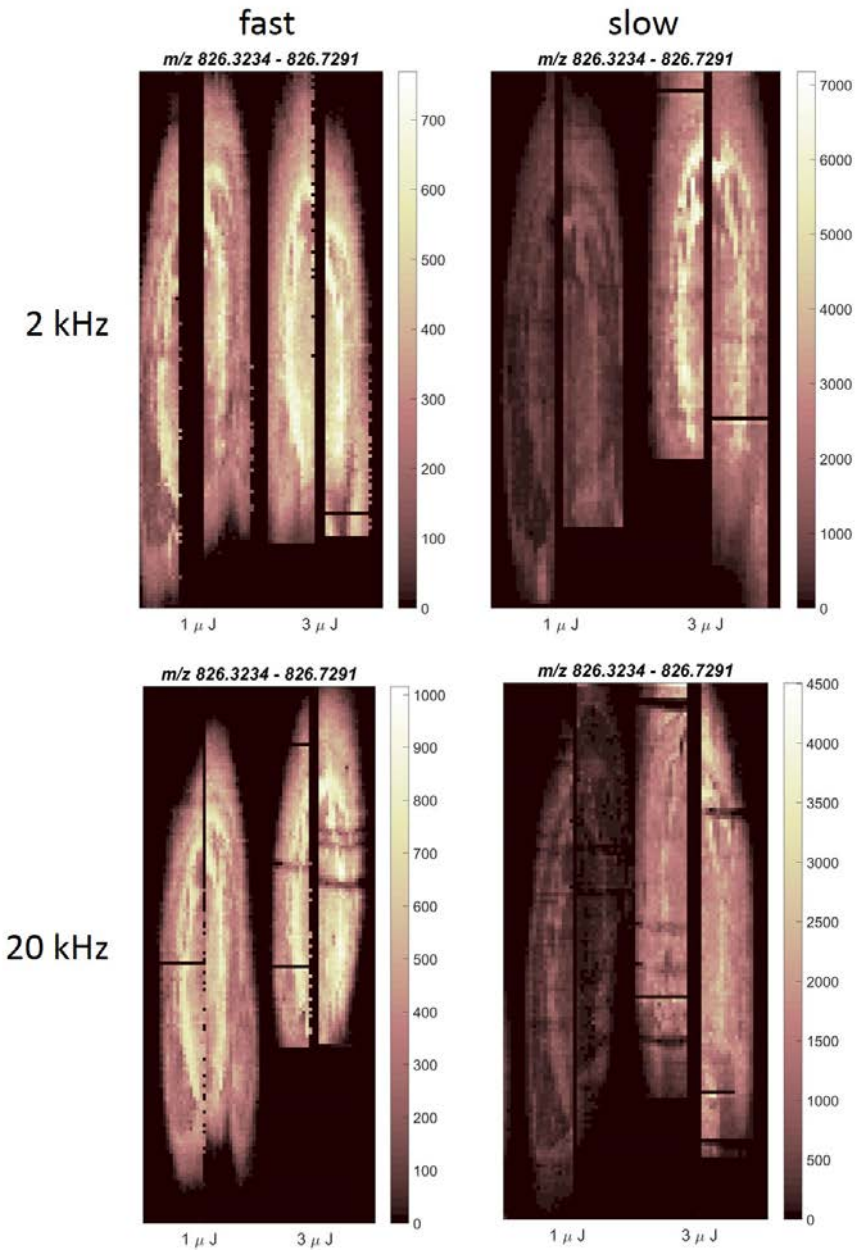

Figure 2.17: Composite images of m/z 826.6 acquired using the slow and fast raster speeds, 2 and 20 kHz repetition rates, and at 1 and 3 μJ laser energy per pulse. As with the images from m/z 798.5 the energy per pulse has a huge effect on ion intensity when using the slow raster speed, particularly at 2 kHz repetition rate, but at the fast raster speed, this energy change has little to no effect.

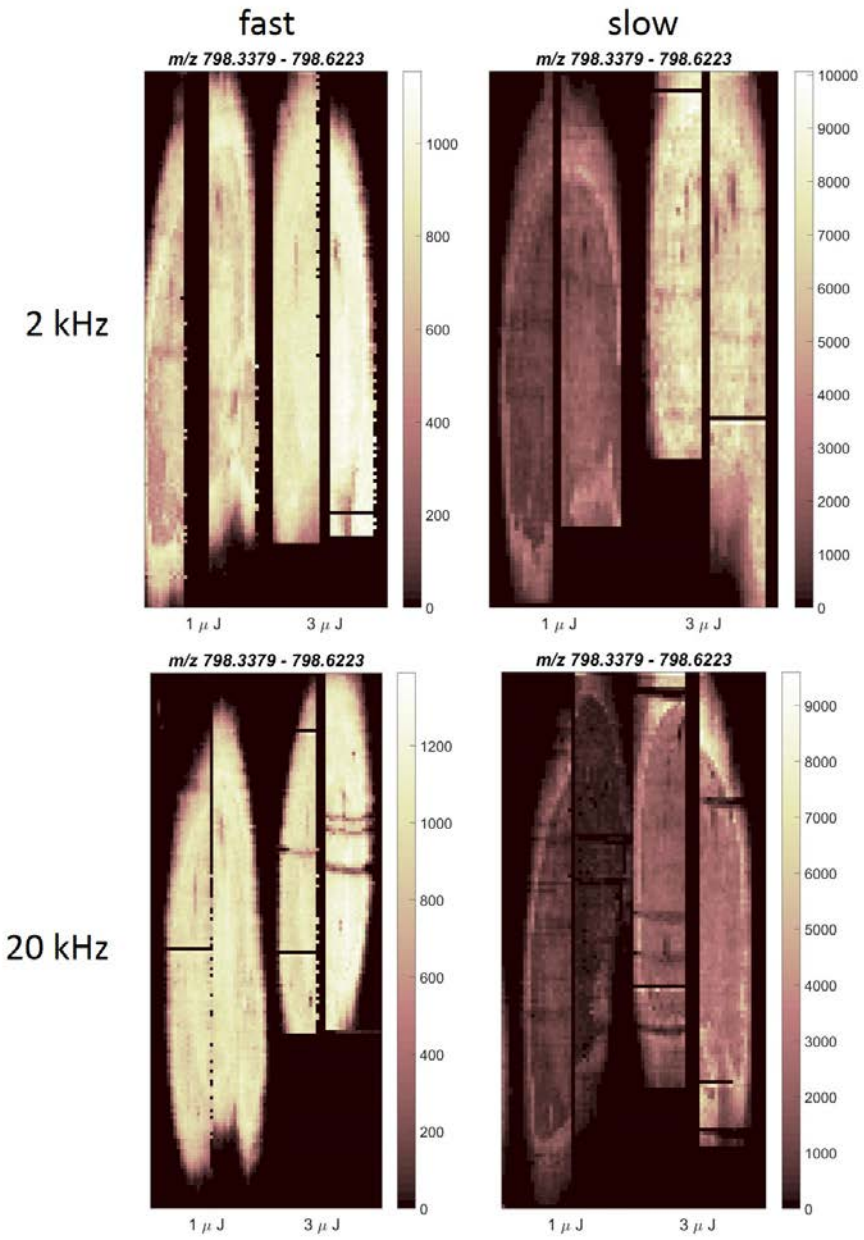

Figure 2.18: Composite images of m/z 798.5 acquired using the slow and fast raster speeds, 2 and 20 kHz repetition rates, and at 1 and 3 μJ laser energy per pulse. The energy per pulse has a huge effect on ion intensity when using the slow raster speed, particularly at 2 kHz repetition rate, but at the fast raster speed, this energy change has little to no effect.

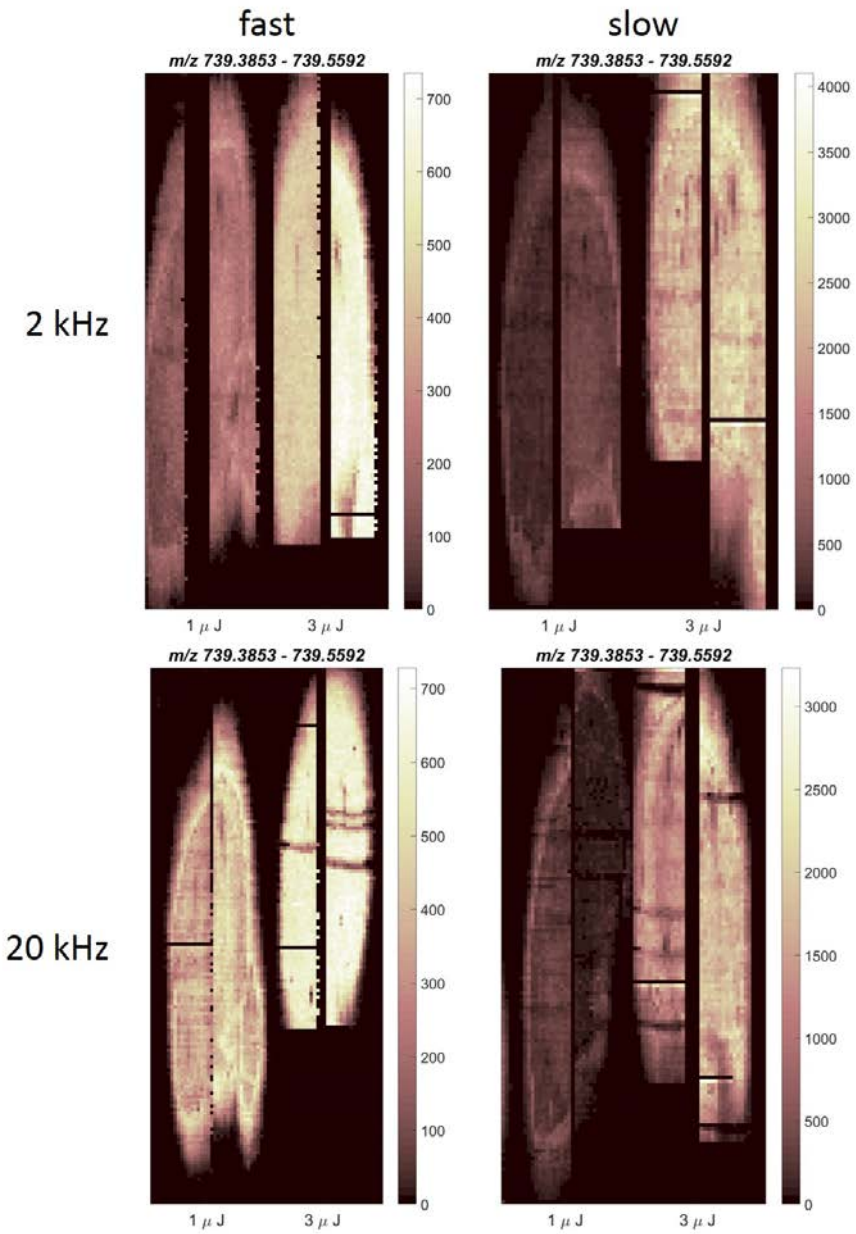

Figure 2.19: Composite images of m/z 798.5 acquired using the slow and fast raster speeds, 2 and 20 kHz repetition rates, and at 1 and 3 μJ laser energy per pulse. Unlike the images at m/z 798.5 (figure 2.18), using the fastest raster speed at 2 kHz repetition rate produces very different intensities at 1 and 3 μJ energies per pulse. However at 20 kHz repetition rate the effect is fairly minimal.

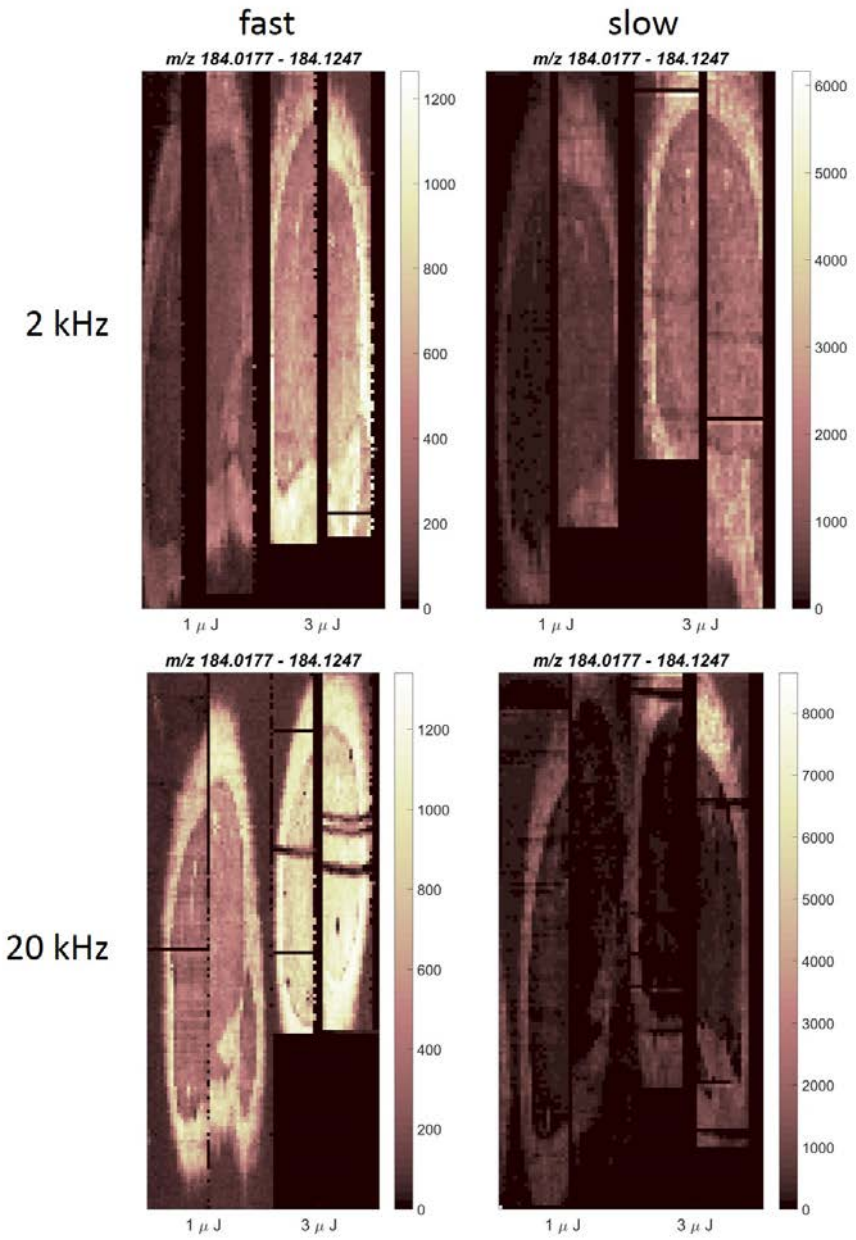

Figure 2.20: Composite images of m/z 184.1 acquired using the slow and fast raster speeds, 2 and 20 kHz repetition rates, and at 1 and 3 μJ laser energy per pulse. The energy per pulse has a huge effect on ion intensity at all raster speeds and repetition rates.

In order to get an unbiased and more complete overview of these data, PCA was applied to each of these pairs of images. Comparing these data using PCA also shows that the main source of variance in almost all cases is an increase in intensity between 1 and 3 μJ energy per pulse (Figures 2.21 to 2.24). Unlike the other datasets, in the

case of the data acquired at the fastest raster speed, and 20 kHz repetition rate, the first principal component does not separate the two energies per pulse and instead separates the outer "halo" often observed in tissue imaging studies (Figure 2.24). The variance between the low energies delivered is then separated by the second principal component and as with the others is largely an intensity difference in the data (Figure 2.25).

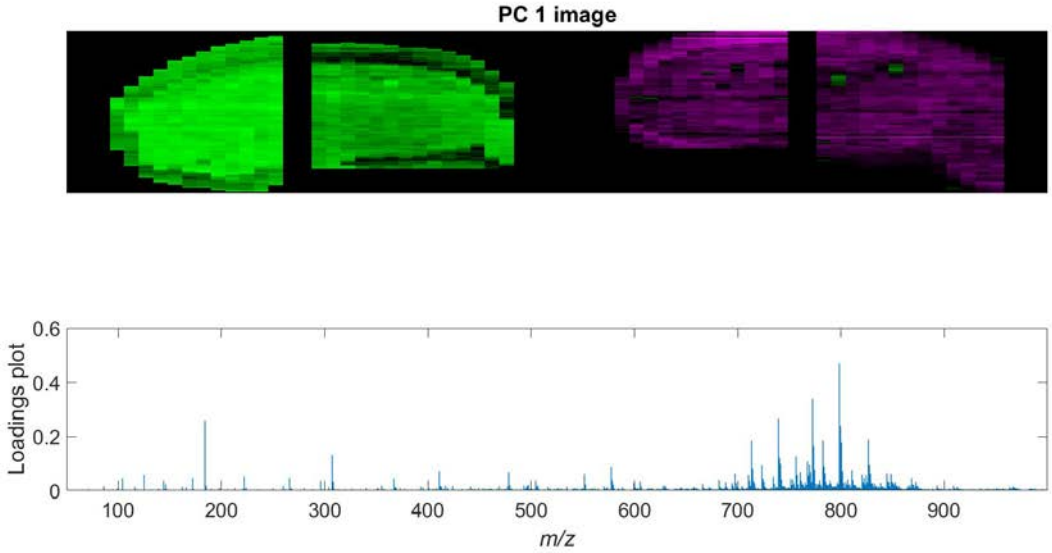

Figure 2.21: First principal component from the combination of 1 and 3 μJ datasets, the slow raster speed and 2 kHz repetition rate. Like the fast raster speed data (figure 2.23), the first component separates these data by the energy per pulse used, and the loadings are entirely dominated by an increase in intensity with increased energy.

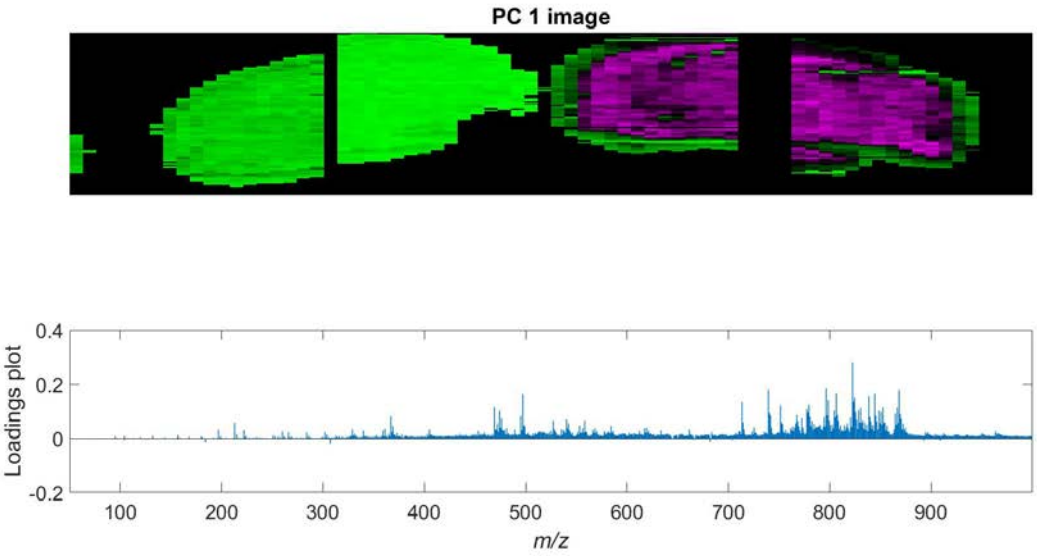

Figure 2.22: First principal component from the combination of 1 and 3 μJ datasets, the slow raster speed and 20 kHz repetition rate. As with the data acquired at 2 kHz, the first component separates these data by the energy per pulse used, and the loadings are almost entirely dominated by an increase in intensity with increased energy.

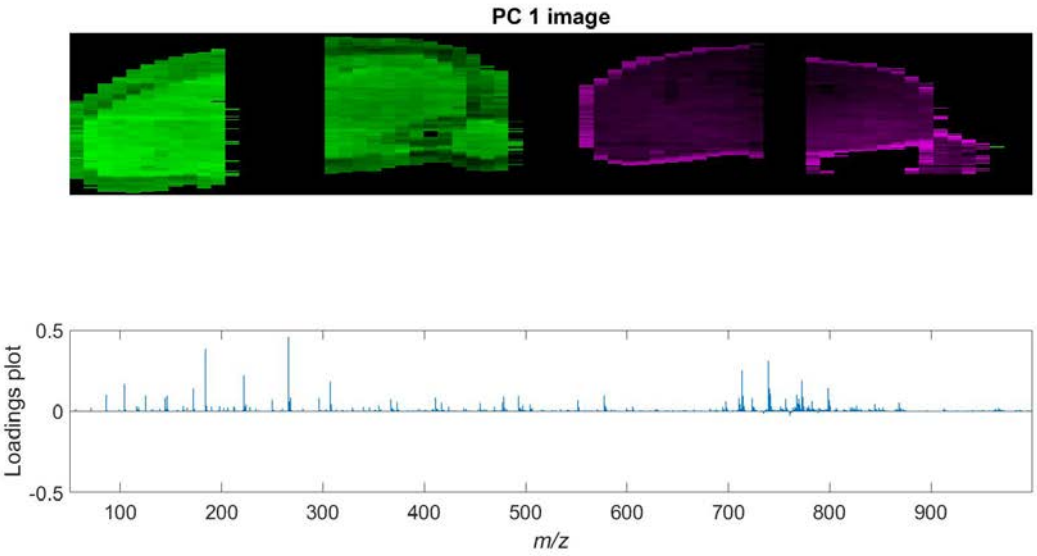

Figure 2.23: First principal component from the combination of 1 and 3 μJ datasets, the fast raster speed and 2 kHz repetition rate. The first component separates these data by the energy per pulse used, and the loadings are almost entirely dominated by an increase in intensity with increased energy.

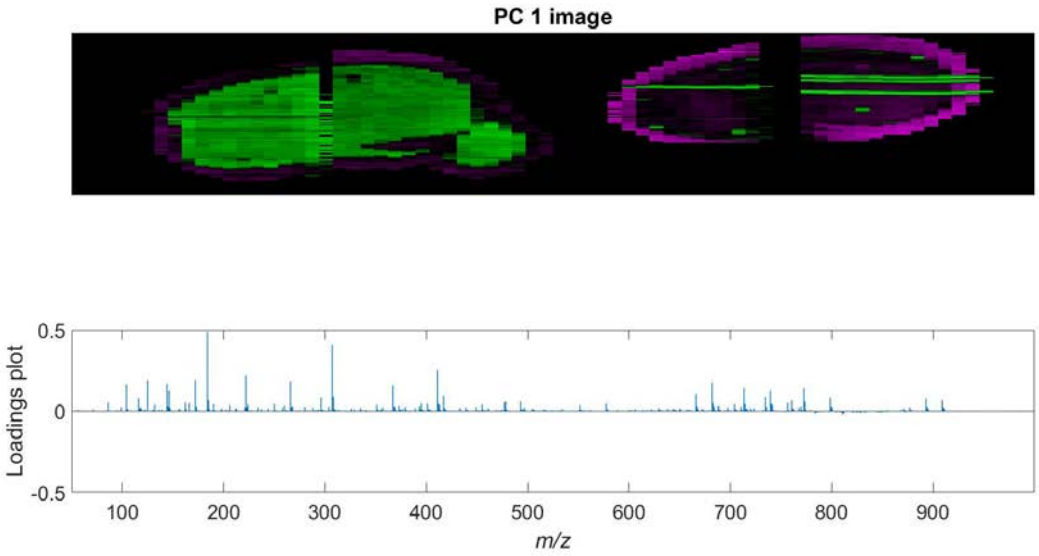

Figure 2.24: First principal component from the combination of 1 and 3 μJ datasets, the fast raster speed and 20 kHz repetition rate. Unlike the 2 kHz data, the first component does not separate these data by the energy per pulse used, and instead separates the tissue "halo" from the main tissue.

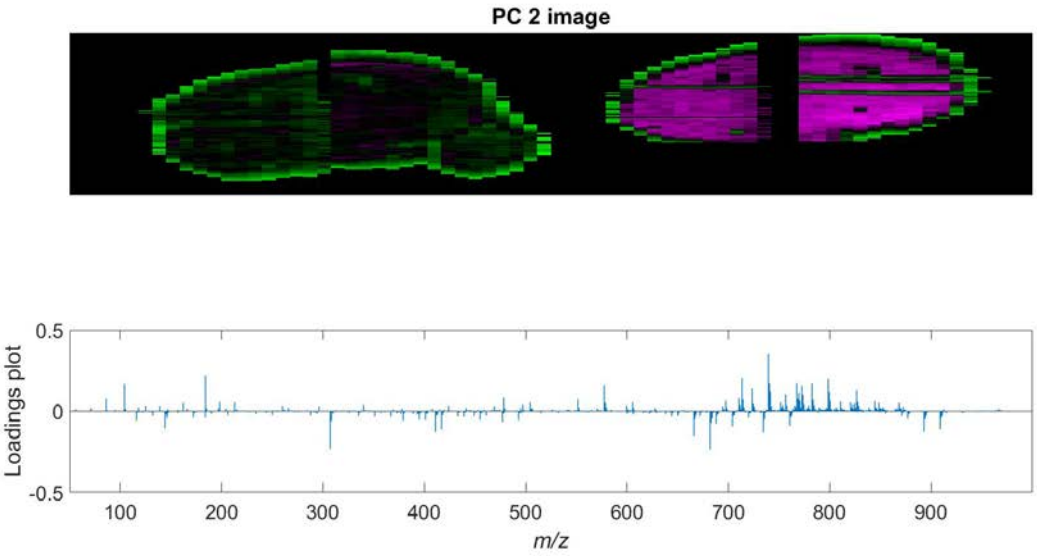

Figure 2.25: Second principal component from the combination of 1 and 3 μJ datasets, the fast raster speed and 20 kHz repetition rate. Unlike the 2 kHz data, the first component does not separate these data by the energy per pulse used, and this separation is seen in this second principal component.

By comparing just the data at 1 or at 3 μJ the primary source of variance can be removed and then the effect of repetition rate and raster speed analysed separately by PCA. To analyse this, the datasets at all raster speeds and repetition rates at 3 μJ were analysed together. This reveals that the main source of variance within the data is the raster speed, as seen by the first principal component of these data (Figure 2.26). In these images, the two brain datasets acquired at the slowest raster speed (left) are grouped together and separate to the faster raster speed data (right), and the spectral loadings are largely dominated by the difference in intensity between these datasets. Following this, the next largest source of variance within the data is the difference between the 2 and 20 kHz repetition rates for the slow raster speed data (Figure 2.27). The spectral loadings of which show positive loadings (associated with the 2 kHz data) for many of the abundant lipid ion peaks analysed such as m/z 798.5, 782.5, and 826.5, as well as fragments such as 739.5 and 184. In comparison, the negative loadings contain many peaks within the lipid

region of the mass spectrum (m/z 700-900) which could be the less abundant species. Of note, the loadings for the ion at m/z 798.5 (presumed PC 34:1 based on literature [63,193]) are positive, yet m/z 796.5 (presumed PC 34:2) are negative. Therefore this change is not a lipid class dependence, and since the ions at m/z 798.5, 782.5, and 760.5 all have positive loadings, this is also not an adduct affect. Other possible explanations could be either be the number of double bonds present in the lipid, or an abundance dependence.

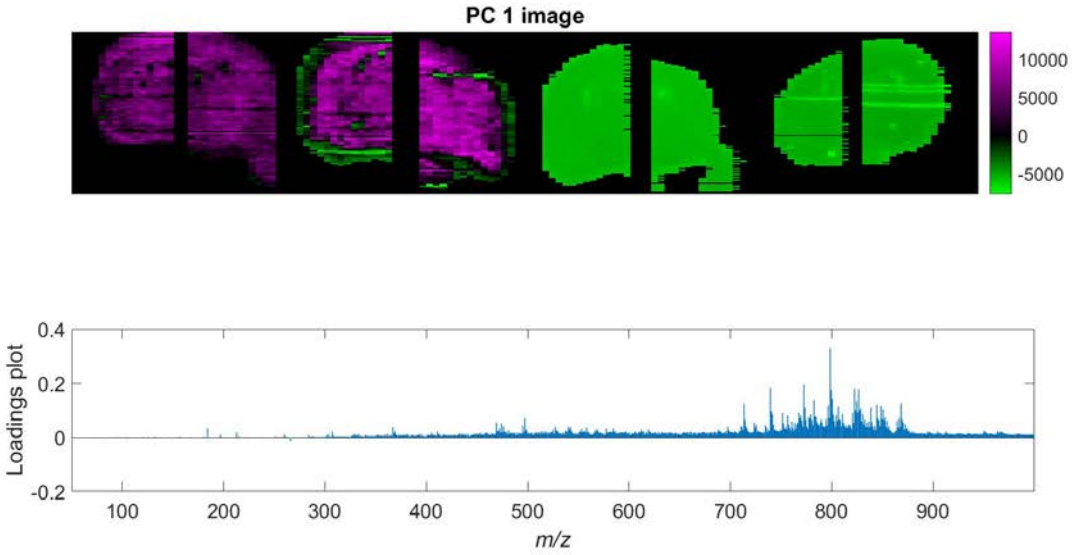

Figure 2.26: First principal component from the combination of all 3 μJ datasets, from left to right 2 kHz slow, 20 kHz slow, 2 kHz fast, and 20 kHz fast. The first component separates these data by raster speed, and the loadings are almost entirely dominated by an increase in intensity at the slowest raster speed.

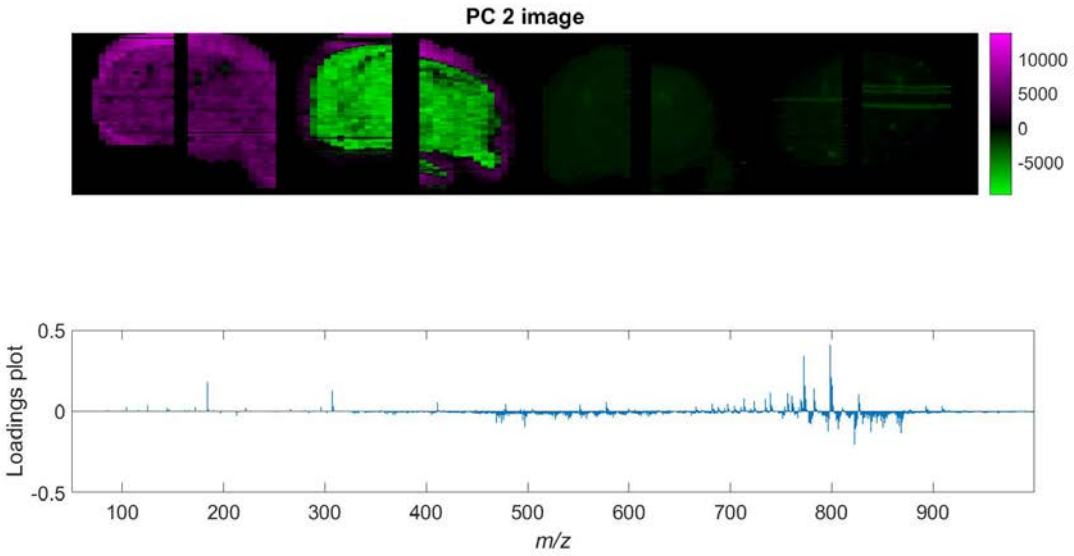

Figure 2.27: Second principal component from the combination of all 3 μJ datasets, from left to right 2 kHz slow, 20 kHz slow, 2 kHz fast, and 20 kHz fast. This component primarily separates the slow raster speed data by repetition rate. The loadings show an increased intensity for the abundant lipids (m/z 798.5, 782.5, and 826.6) at 2 kHz, and an increased intensity for many others at 20 kHz.

2.4 Conclusions

While the full influence of repetition rate, pulse energy, and stage speed in raster mode MSI of biological samples remains undetermined, some of the effects of these variables are now better understood. The predominant factor that affects the signals observed is by far the energy per pulse, however in most commercial systems, the means to measure this is not present. Furthermore, it was demonstrated that when operating lasers at high repetition rates, the energy delivered in the first ten thousand laser pulses is much greater than after this. Therefore the use of a shutter to maintain steady state operation is suggested for all laser systems in MALDI. Following on from this, there is a m/z dependant change in an ions intensity with varying repetition rates, and this is not a variability that

is based on either lipid class or adduct formation. Increasing signal for a given ion of interest cannot simply be achieved by using a higher repetition rate, or a slower raster speed. Additionally for newer generations of instruments, when using fast raster speeds and high repetition rates (20 kHz), changes in laser energy output contribute less towards variance in the data than at lower repetition rates or slower raster speeds. Furthermore, the experimental aspects of this chapter, including the use of a shutter to reduce laser variability, serial coronal mouse brain sections with each half imaged separately, and multivariate data analysis, improve the quality and quantity of information gained from these fundamentals studies for MALDI tissue imaging, and can also apply to many other fundamentals studies as well as emerging fields in MSI.

CHAPTER 3

DATA PROCESSING FOR LESA FAIMS MSI

3.1 Introduction

Another promising new area for MSI is in LESA-ESI imaging. LESA coupled to ESI-MS provides a means to analyse a number of different classes of molecules at discrete spatial locations in a tissue sample [83, 84, 194, 195]. This can be used to generate MS images by sampling in a regular grid however these typically have a limited number of pixels (< 400). The two main reasons for this are the throughput (10s of seconds per pixel), and the sampling area of LESA ($\sim$1 mm). Despite this limitation, LESA is a popular technique due to a number of factors; it provides a highly sensitive technique to map the spatial distribution of one of the broadest range of possible molecules from a sample, including drugs [90], peptides [196], lipids [91], and intact proteins [83]. Critically, unlike MALDI-MSI since the ionisation is electrospray based, multiply charged ions are produced, meaning large proteins can be analysed on high resolution Orbitrap mass analysers despite their limited m/z range. In addition to this, little to no sample preparation is required, and LESA can be performed on a number of different sample types, and has been demonstrated for tissue sections, bacterial colonies, and dried blood spots [83,87,89]. The other primary advantage of LESA is the large sample volumes that are used. While these limit the spatial resolution to $\sim$1 mm, they allow for long MS analysis times which mean that fragmentation experiments can be performed on selected ions allowing for top

down characterisation of proteins, and confirmatory analysis of drugs and lipids [83]. One potential challenge within LESA-MSI is the complexity of the resulting spectra, with even a single molecule giving rise to many peaks from both multiple adduct possibilities and multiple charge states. For one single molecule, if there are three possible adducts (such as $[M+H]^+$, $[M+Na]^+$, and $[M+K]^+$), five different charge states, and five isotope peaks observed, there will be a total of 75 peaks in the mass spectrum for this molecule. In reality, there can be as many as 20 charge states, and this could arise from a number of possible adduct combinations e.g. for a 9+ charge state this could be $[M+9H]^{9+}$, $[M+9Na]^{9+}$ or any mixture of [M+H] and [M+Na].This results in the possibility of many overlapping peaks in the spectra. The two potential approaches to handle this are to employ computational methods to deconvolute these spectra, or additional separation methods such as LC or ion mobility. For imaging purposes, LC timeframes are far too long to be feasible, whereas ion mobility separation is much more rapid and therefore suitable for LESA imaging experiments.

FAIMS and other ion mobility separation methods are an emerging field in MSI which have a great deal of potential. This allows additional separation of ions with the same m/z, in much faster timescales than chromatography methods, thereby suitable for imaging experiments. Additionally, unlike standard ion mobility separation where there is a correlation between cross section and m/z, FAIMS is based on differential mobilities of ions in high and low fields, and as such there is no direct relationship between FAIMS mobility separation and m/z. The two main challenges of these data are the additional variables involved, and the data analysis and visualisation of these data. FAIMS experiments require optimisation of CF and DF variables, and ion trajectories are highly dependant on both of these variables, requiring optimisation of both of these simultaneously. In addition to this, currently, FAIMS and other ion mobility data analysis are limited to univariate analysis which negates much of the benefit of the additional dimensionality. The main challenge in visualisation of data containing mobility based separations is the complication arising from the additional dimensionality. Typically, multivariate analysis

of MSI data discards spatial information to create a 2D matrix as seen in section 1.4 but the additional mobility dimension means that even with this, the data contains three orthogonal variables.

The work presented in this chapter explores data processing methods to allow interactive and multivariate analysis to optimise these FAIMS conditions. Following this, LESA FAIMS MSI experiments are carried out, with different data processing approaches to determine which methods provide the most useful and informative data.

3.2 Experimental

3.2.1 Sample preparation and LESA sampling

Single lambs brain hemispheres were acquired from a local abbatiore, and stored at -80^{o}C until sectioning. Coronal tissue sections were taken at 60 μm thickness, and thaw mounted onto polyvinylidene fluoride (PVDF) membrane. LESA was performed using the Advion nanomate robot, using a solvent consisting of water, acetonitrile and formic acid (49.5%/49.5%/1%). For LESA FAIMS 2D sweep analysis, a 5 μL volume of solvent was drawn into the LESA pipette, and a volume of 1.5 μL was dispensed onto the sample and held there for 1 minute before reaspiration, this extraction process was repeated three times before infusion into the mass spectrometer. For the comparison between the 1D sweep and stepped static data, a smaller solvent volume of 1.5 μL was used, and 1 μL was dispensed onto the sample for 1 minute prior to reaspiration and two extractions were performed.

3.2.2 Mass spectrometry

For the comparison of FAIMS using nitrogen and air, a mixture containing a ubiquitin (50 μM) standard added to the Thermo calmix positive ion calibration solution (Thermo Scientifc, UK) was introduced into an Orbitrap Elite mass spectrometer (Thermo Scientific,

UK) set to a *m/z* range of 150-2000 by nano electrospray using the Triversa Nanomate robot (Advion Biosciences, Ithaca, USA), using a gas pressure of 0.3 psi and a voltage of 1.7 kV. FAIMS separation was then performed using the Owlstone ultraFAIMS device (Owlstone, Cambridge, UK) with a chip temperature around 100 ^{o}C. Currently due to the setup of this device, there is no control or monitoring of the temperature, and any heating comes as a result of conduction from the instrument heated cone. Future improvements could involve direct heating of this device to regulate and monitor the temperature of the chip. CF was scanned between -1 and 4 Td over a 5 minute period, and DF values of 130 to 270 Td in 20 Td increments. Either nitrogen or air was used as the FAIMS carrier gas as described. This sample represents a complex mixture of a number of different molecular classes including peptide (MRFA), protein (ubiquitin), drug (caffeine), and polymer (Ultramark 1621) allowing information to be gained as to the effects of FAIMS variables on different molecular classes.

For the LESA 2D FAIMS sweep data acquisition, the extracted solution from the lambs brain was introduced into an Orbitrap Elite mass spectrometer (Thermo Scientific, UK) set to a *m/z* range of 600-2000 by nano electrospray using the Triversa Nanomate robot m (Advion Biosciences, Ithaca, USA), using a gas pressure of 0.3 psi and a voltage of 1.7 kV. In these experiments, the ultraFAIMS device was set to 120 ^{o}C, and a CF sweep of 0 to 6 Td was performed at DFs of 208 to 308 in 20 Td increments, and air was used as the FAIMS carrier gas.

For the comparison between the FAIMS stepped static, 1D sweep or non-FAIMS acquisition, extracted solutions were introduced into an Orbitrap Elite mass spectrometer (Thermo Scientific, UK) set to a *m/z* range of 600-2000, and resolution of 120,000 FWHM at *m/z* 200 by nano electrospray using the Triversa Nanomate robot m (Advion Biosciences, Ithaca, USA), using a gas pressure of 0.3 psi and a voltage of 1.7 kV. For the 1D sweep, the DF was set to 308 Td, and CF was ramped between 0 and 6 Td over three minutes. For the stepped static acquisition, the DF was set to 308 Td and the CF was held at 0, 1, 2, 3, 4, 5, and 6 Td for 30 seconds each.

3.2.3 Data processing

All data processing was performed on an Intel Xeon quad core CPU E5-2637 v2 (3.50 GHz) with 64 GB of RAM.

FAIMS 2D sweep data conversion

All data were converted from proprietary Thermo .raw fromat to mzML using the ProteoWizard converter [197]. For the data processing of FAIMS 2D sweep data, spectra were loaded into Matlab (version R2014a and statistics toolbox, The Math-Works, Inc., Natick, MA, USA) using the imzML converter [118]. Each spectrum was then zero filled using the Orbitrap zero filling function in SpectralAnalysis [191]. Following this, the CF axes at each m/z were then interpolated using the methods described by Sarsby *et al.* [93] to produce concurrent CF axes. The zero intensity values were then removed from each spectrum to produce the common sparse representation (processed format), and each spectrum was then exported as a new mzML file. These mzML files were then combined into a single imzML file using the imzML converter [118], using the imzML x and y pixel co-ordinates to represent CF and DF respectively. Currently there are no supported parameters within the imzML vocabulary for FAIMS parameters, therefore there is currently no retention of the CF and DF when viewing the data. The imzML format however does support user defined parameters which could act as a means for custom FAIMS processing software to incorporate this in the future.

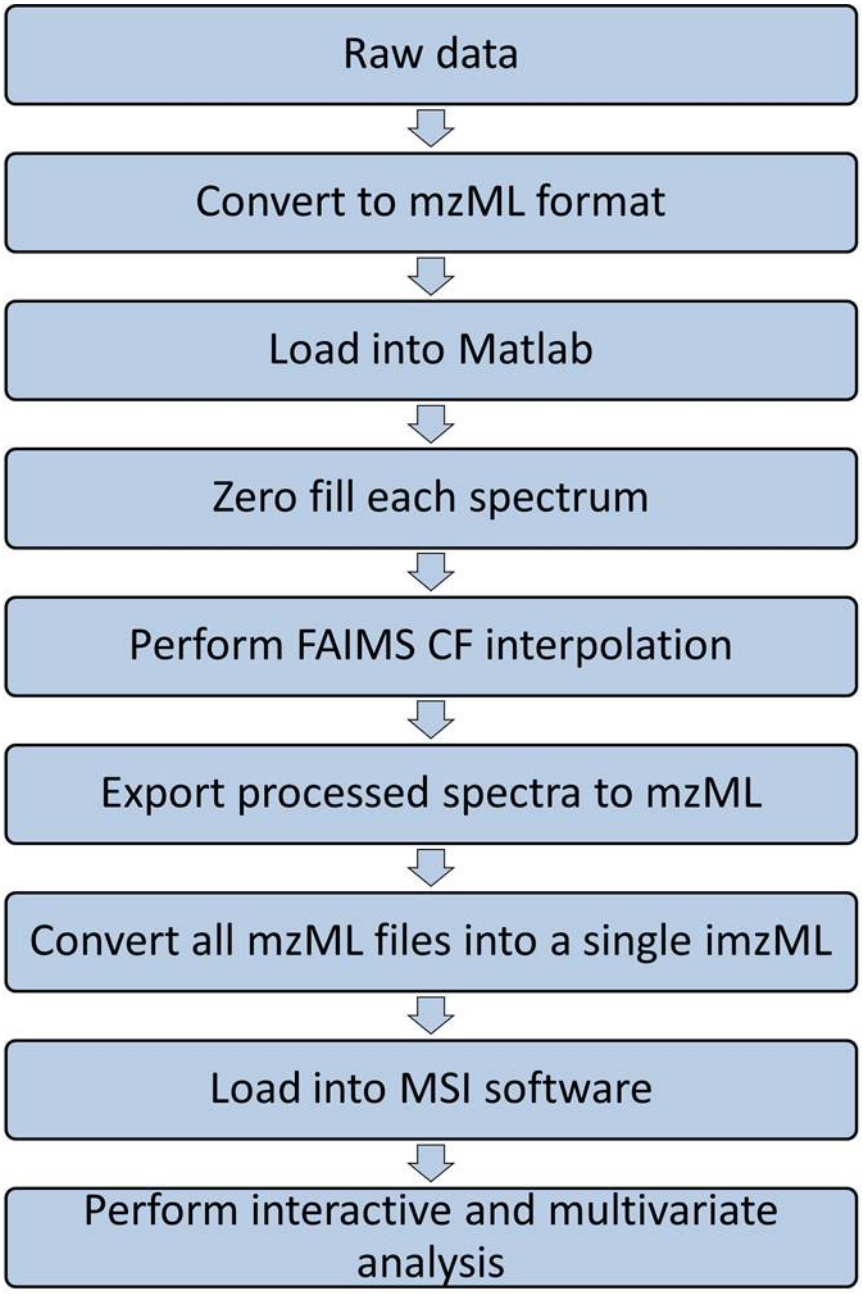

Figure 3.1: Workflow for conversion of 2D FAIMS sweep from raw data to imzML to allow loading into imaging software packages and provide users with interactive and multivariate analysis.

FAIMS 2D sweep data processing

Following imzML conversion, the 2D sweep data were imported into SpectralAnalysis software [191], and zero filled using the Orbitrap zero filling function. The mean spectrum was then peak picked using the gradient method with an intensity threshold offset to above noise level based on manual inspection of the mean spectrum. These data were then exported into Matlab (version R2014a and statistics toolbox, The Math-Works, Inc., Natick, MA, USA), and NMF and PCA was performed using the Matlab functions "nnmf" and "princomp" respectively.

LESA FAIMS MSI data processing

For the comparison of LESA MSI data acquired with and without FAIMS separation, where stated, prior to this conversion the spectra were deconvoluted using the Xcalibur Xtract function (Xcalibur 2.10 software, Thermo Fisher Scientific) both on each indivudual spectra within the chromatogram, or on a single summed spectrum. These data, either deconvoluted or not, were converted from proprietary Thermo .raw format to mzML using the ProteoWizard converter [197]. Where the Xtract deconvolution had not been performed, the data were then loaded into Matlab (version R2014a and statistics toolbox, The Math-Works, Inc., Natick, MA, USA) using the imzML converter [118], and deconvolution performed using custom scripts based on the "Thorough High Resolution Analysis of Spectra by Horn" (THRASH) algorithm [198]. Following this, these deconvoluted spectra were exported as new mzML files. All mzML files were then converted into imzML using the imzML converter [118] and analysed using SpectralAnalysis [191].

3.3 Results and Discussion

3.3.1 FAIMS 2D sweep data processing

When performing a LESA FAIMS MSI experiment, it is necessary to first determine the desired CF and DF parameters to use. This is typically done by performing a 2D sweep of the CF and DF parameters, whereby at a number of fixed DF values, the CF is continuously scanned across a range [91, 93]. This then produces data at a wide range of CF and DF combinations, thereby allowing the user to determine the optimal CF and DF for any ions of interest. When acquiring a 2D FAIMS experiment on an Orbitrap mass analyser with automatic gain control (AGC) on, the time points along the chromatogram will not be evenly spaced. Since the sweep of the CF is continuous, this means that the change in CF between each spectrum will not be the same. Recently, Sarsby *et al.* published data processing methods to interpolate the data within the chromatogram to give a concurrent CF axis [93]. This allows the user to generate total ion transmission (TIT), and single ion transmission (SIT) maps, showing intensity at different CF and DF for a given m/z, or CF and m/z for a given DF which can be viewed either as a 2D image or in 3D [91, 93]. This has been used to determine optimal CF and DF conditions for LESA FAIMS experiments. The main limitation of this is that only a single m/z or DF can be considered at once, and the processing has to be repeated each time a new m/z or DF needs to be analysed.

These data can be considered a similar format to MSI data where instead of pixel co-ordinates in x and y, CF and DF values are considered. By performing the workflow described in section 3.2.3, this will then generate a dataset containing spectra with both concurrent m/z and DF axes. This allows the data to be analysed using MSI software packages such as SpectralAnalysis [191], and allows a more interactive viewing as well as a wide range of additional analyses to be performed. The features that are particularly useful for FAIMS 2D sweep processing are image and spectral overlays, the combination of multiple experiments into a single dataset, and multivariate analysis methods such as

PCA and NMF.

Overlays

In mass spectrometry imaging software packages such as spectral analysis, images from selected ions of interest, and spectra from selected pixels are commonly overlayed [191]. Overlaying selected ions in the newly converted FAIMS data allows users to determine at what DF and CF given ions are separated. For example, in a 2D FAIMS sweep of a mixture of ubiquitin and calmix, it can be seen that the charge states 11^+, 7^+, and 5^+ are well separated at DF above 250 Td, but not at lower DFs (Figure 3.2) . This could be achieved by generating multiple SIT maps, but would be more time consuming and would require more user interpretation of the data. Similarly, spectra under different conditions can be compared, allowing users to determine which ions are being transmitted under different conditions (Figure 3.3). Again this could be achieved by manually looking through the chromatogram, but then requires users to determine what the CF and DF conditions were for each time point. This more interactive view of the 2D FAIMS data allows users to optimise their parameters much more quickly and effectively.

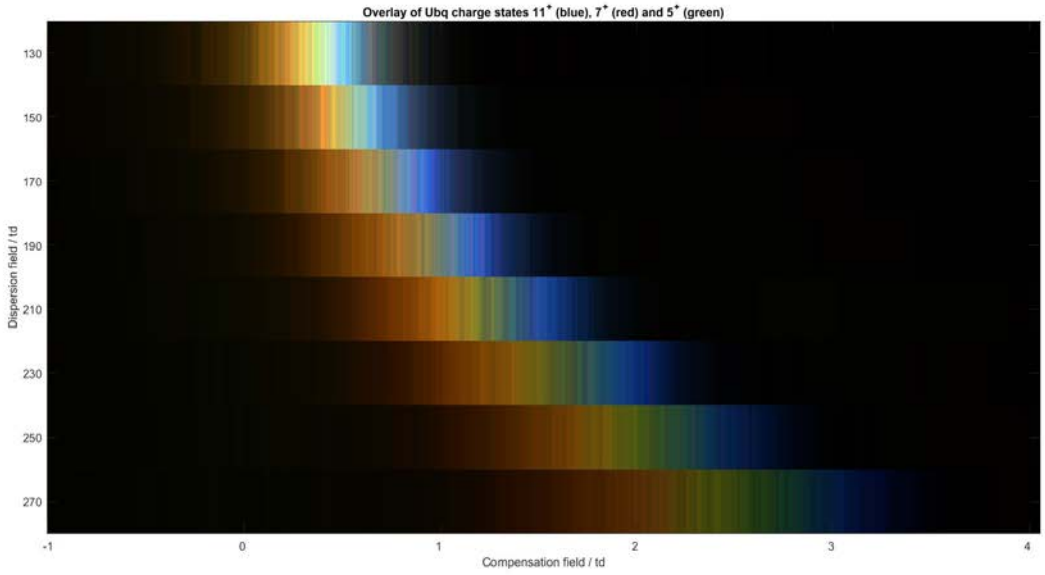

Figure 3.2: Overlay of Ubq charge states 11^+ (blue), 7^+ (red) and 5^+ (green) at a range of CF and DF conditions. This show good separation of these ions at DF above 250.

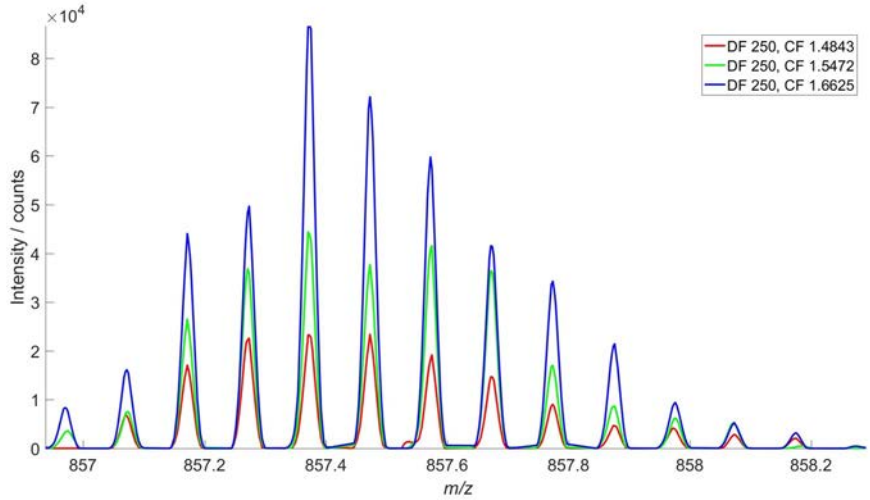

Figure 3.3: Zoom in on the Ubq 10^+ charge state from an overlay of spectra from different CF and DF conditions showing how the changing from low to high CF results in increased transmission of more of the higher charge state ions (lower m/z).

Analysis of multiple experiments

In the imzML converter software, there is functionality to allow the combination of multiple imzML files together to allow the comparison of different experiments [118]. Using this with the FAIMS 2D sweep data allows comparison of different experimental conditions. This is demonstrated on two datasets from a direct infusion of a mixture of ubiquitin and Calmix (Thermo Fischer Scientific) where either air or nitrogen is used as the FAIMS carrier gas. The ion transmission at various charge states clearly shows improved CF resolution with the nitrogen carrier gas, and by combining this with the overlay feature, it can be clearly seen that the CF spread of ions is much lower, and charge states (11^+, 7^+ and 5^+) are separated at DF above 210 Td when nitrogen is used but only above DF 250 Td when air is used(Figure3.4).

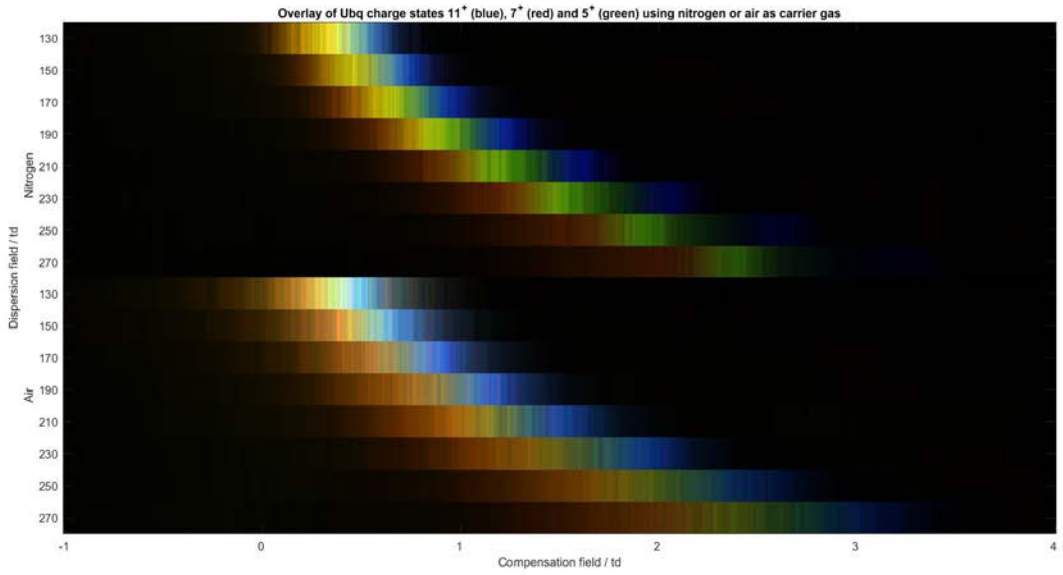

Figure 3.4: Overlay of Ubq charge states 11^+ (blue), 7^+ (red) and 5^+ (green) at a range of CF and DF conditions using both nitrogen and air as carrier gasses. This clearly shows increased separation of these ions when nitrogen is used as the carrier gas, but a slight decrease in the observed ion intensities.

Multivariate analysis

Within some mass spectrometry imaging software, there is functionality to perform multivariate analysis methods such as PCA and NMF [191, 199]. By applying multivariate analysis to FAIMS 2D sweep data, all ions within the data are considered rather than just single ions, giving a much more rapid overview of the data, allowing for the analysis of variance, and reducing potential for user bias within the data analysis. Methods such as NMF can be used to reduce the data into a predefined number of components, thereby allowing the simultaneous determination of optimal static conditions for multiple ions rather than just individual ones. PCA on the other hand determines the largest sources of variance within the data, and can be used either in the studies of the fundamental processes involved in FAIMS, or in conjunction with the functionality to combined multiple datasets to determine the effect that these variables have. For example, PCA applied to the single dataset of infused ubiquitin shows the second largest source of variance (after regions of ion transmission vs regions with very little ion transmission (Figure 3.5)) between the (9^+ and below and 10^+ and above) charge states of ubiquitin (Figure 3.6). This has been shown through a combination of electron capture dissociation (ECD) and photofragment spectroscopy experiments to be a change from loss of tertiary structure above charge state 9^+ [200]. This indicates that conformation is critical to the changes in mobility observed in FAIMS experiments. While this has already been shown in other studies using FAIMS [201, 202], by applying multivariate analysis to FAIMS 2D sweep data, all ions within the data are considered rather than just single ions, giving a much more rapid overview of the data, allowing for the analysis of variance, and reducing potential for user bias within the data analysis. Application of multivariate methods to 2D FAIMS data has the potential to deepen our understanding of the mechanisms involved which are still not fully understood [203].

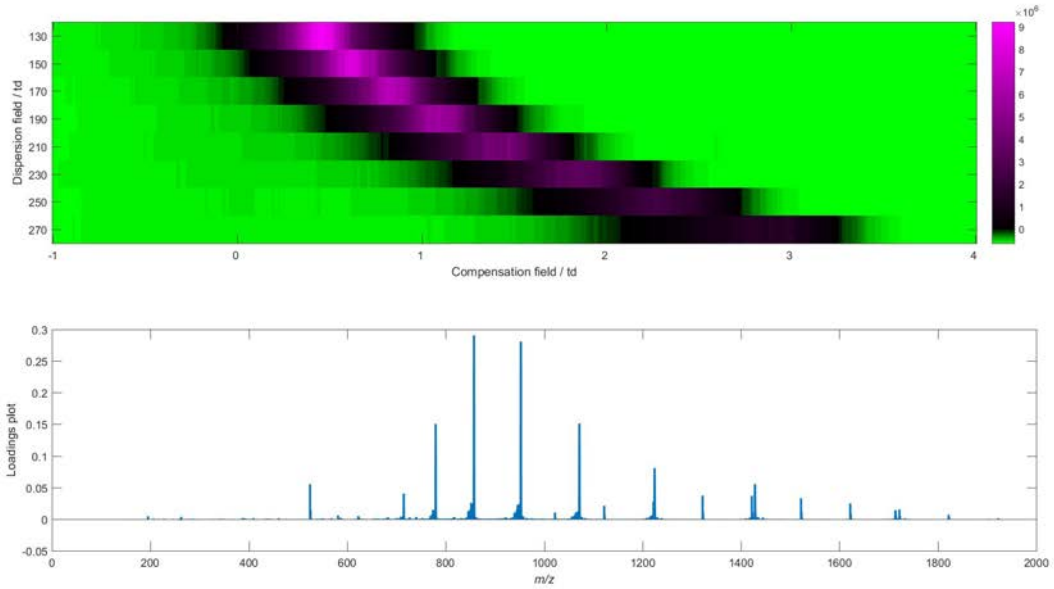

Figure 3.5: First principal component from the 2D sweep data of calmix and ubiquitin showing regions of very high ion transmission (positive loadings) and little to no transmission (negative).

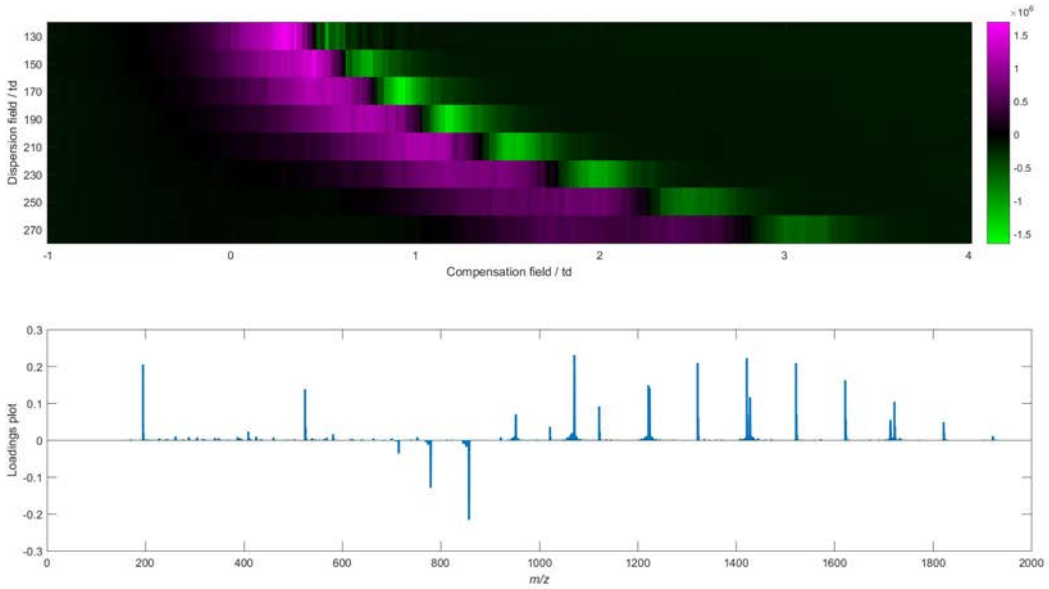

Figure 3.6: Second principal component from the 2D sweep data of calmix and ubiquitin. The positive loadings correspond to ubiquitin charge states 10^+ and above, and the negative to charge states 9^+ and below. This indicates that this is one of the highest sources of variance between the CF and DF conditions, and above charge states of 9^+ ubiquitin tertiary structure is lost.

In addition to this, PCA can be applied to the combined datasets with air and nitrogen as carrier gass described previously. When the datasets with nitrogen gas and air are combined and PCA is applied, no principal component entirely separates the two experiments from one another. This indicates that aside from the improvements in CF resolution described previously, no additional changes to the spectra are caused by the use of nitrogen as the carrier gas. This can be potentially used to investigate the effect of different carrier gas mixtures on the results obtained when using FAIMS separation.

Another aspect of FAIMS 2D sweep acquisition is to determine optimal conditions to perform static separation experiments for specific ions. In this case, the use of NMF allows a reduced representation of the data to be generated which shows the optimal conditions for transmission of not a single ion, but instead ions which have similar properties under FAIMS separation. By performing NMF analysis on the 2D FAIMS sweep data from

a LESA MS experiment on lambs brain tissue, optimal FAIMS conditions for multiple analyte classes can be optimised simultaneously. For example, performing NMF with $k = 5$ discovers factors relating to larger proteins (up to 15 kDa) in charge states of 10^+ to 15^+ (factor 2), small proteins (8-10kDa) in charge states 9^+ to 13^+ (factor 4), and small proteins (also 8-10 kDa) in charge states 5^+ to 8^+ (factor 3), along with two factors for small peptides in different charge states (all 2^+ in factor 1 and 3^+ to 5^+ in factor 5) (Figure 3.7).

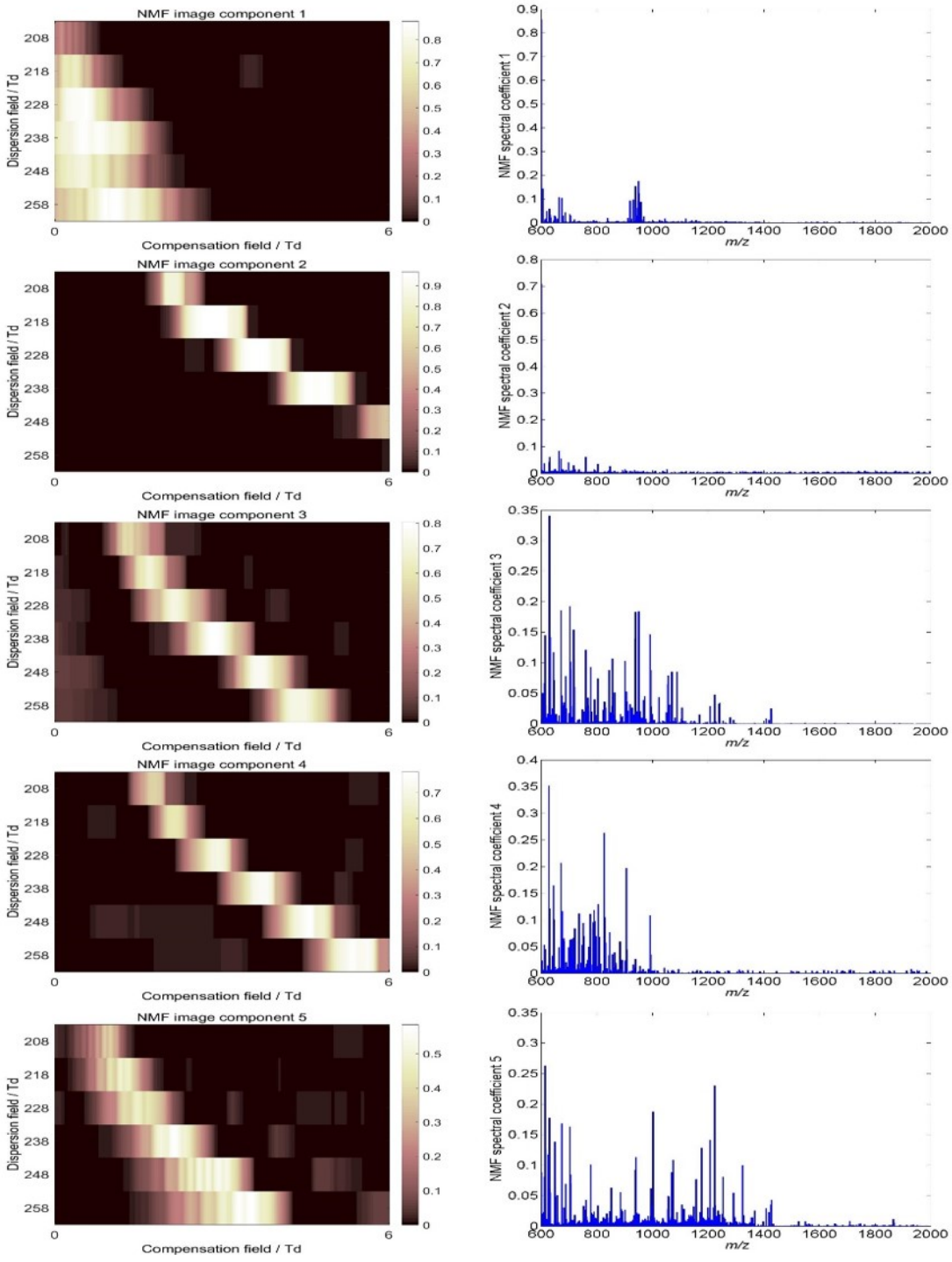

Figure 3.7: Image and spectral factors from NMF of a 2D FAIMS sweep of a LESA extract from lambs brain tissue. Different groups of molecules are preferentially transmitted under different conditions, such as small doubly charged peptides at low compensation fields (factor 1).

In summary, the conversion of the FAIMS 2D sweep data into an imzML format allows a suite of different data mining techniques to be applied to these data. This takes data analysis in FAIMS from simple univariate analysis into much more powerful data analytics. These methods can allow an unbiased and untargeted approach to optimise FAIMS conditions for further experiments, or can be used to further investigate the fundamental processes in FAIMS. A future more streamline development for this conversion could involve the direct export of these data into imzML, thus removing the extra processing step, and reducing the amount of data that needs to be stored on disk.

3.3.2 LESA FAIMS imaging

LESA sampling, in combination with FAIMS separation can also be performed in a spatially discrete manner to create FAIMS separated, MS images. By LESA sampling in a spatially discrete grid pattern, and triggering FAIMS separation, either statically, or in a 2D sweep, differential mobility separation can be achieved alongside MS measurement. There are three possible modes of operation within this; static imaging, where a single DF and CF are used throughout, multiple static imaging, where a number of static CF and DF conditions are used per pixel for a proportion of the injection time, and 1D sweep imaging, where the CF is ramped between two fields over the course of a single pixel (Figure 3.8).

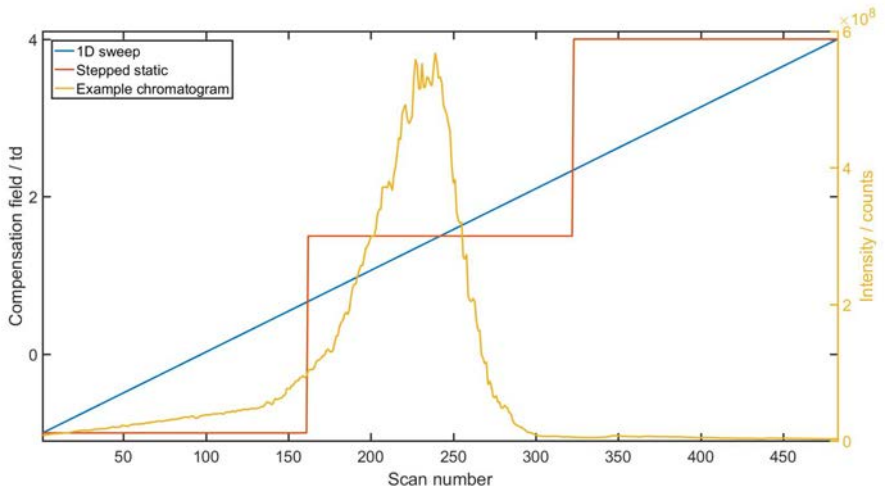

Figure 3.8: Diagram of the compensation fields applied for a 1D sweep and a stepped static FAIMS experiment. Alongside this is an example TIC obtained from a 1D sweep experiment.

Some of this choice will be made based on the considerations of the experiment, such as whether a targeted or untargeted approach is needed, but there are a number of considerations when choosing the experiment to perform. A single static field experiment will give the greatest ion transmission for a given ion of choice, but this requires prior knowledge of both the ion of interest, and the optimal CF and DF transmission fields. By performing multiple static experiments within a given pixel, more than one ion can be imaged by this approach. However, the ion transmission will be greatly reduced as only one ion will be optimally transmitted at once. Additionally, any ions whose optimal transmissions fall in between these optimal conditions will either have their sensitivity greatly reduced, or not be detected at all. By sweeping the CF across the desired range over the course of an experiment, each ion will, at some point experience optimal transmission fields, but will also experience sub optimal conditions for the main duration of the acquisition at each pixel. Therefore this approach is best suited towards a fully untargeted imaging approach. The comparison of static FAIMS against without FAIMS has been previously analysed, with the FAIMS providing increased S/N by filtering out a large amount of chemical noise [93]. Therefore a comparison of a FAIMS 1D sweep was compared to LESA without FAIMS,

and different data processing and analysis methods were compared.

Spectral deconvolution

Since the ionisation is electrospray based, LESA generates multiply charged protein species, therefore high resolution mass analysers such as Orbitrap and FTICR can be used to resolve isotopic structure. This improvement in mass resolution improves protein identification but complicates the data analysis as the intensity ratios of the charge states present may not always be consistent. The effect of this can be reduced by the process of deconvolution [204]. Deconvolution of protein MS data involves reducing the charge states and isotope distributions of the mass spectra into a single monoisotopic mass peak, usually the mass observed where each element is in its most abundant isotopic form.

The first step is to select a small m/z window in which to perform the deconvolution as demonstrated by the 1 m/z window between m/z 952 and 953 shown in (Figure 3.9).

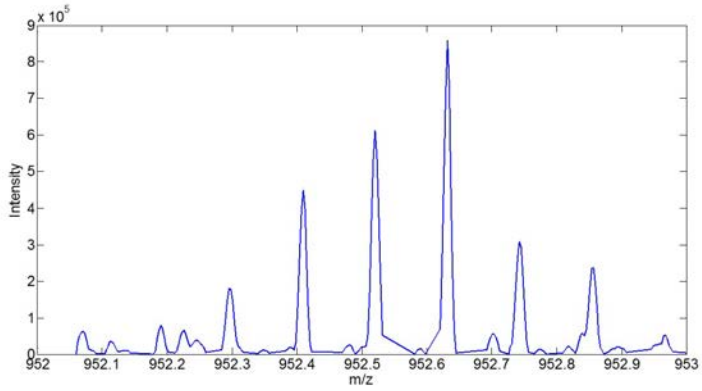

Figure 3.9: Small 1 Da window between m/z 952 and 953 containing the ubiquitin $[M + 9H]^{9+}$ ion for spectral deconvolution.

Within this window, the charge state of the ions present is then determined. This can be done using either the Patterson routine [205], or a Fourier transform based approach [206]. The Patterson function takes a selection of preselected possible charge states, and for each charge state determines the $\Delta m/z$ associated with a hydrogen to deuterium isotope shift for that charge (1.006277 / z). The for each $\Delta m/z$ shift, at selected m/z

intervals across the selected portion of the mass spectrum, all the intensities at the $\Delta m/z$ shifts are multiplied, and then summed together;

$$\sum_{m/z_{min}:m/z_{max}} (m/z - \frac{\Delta m/z}{2}) \times (m/z + \frac{\Delta m/z}{2})$$

This works on the basis that isotope peaks from a given charge state z will have a separation of $\frac{1.006277}{z}$ m/z. For a specified charge z, intensities at a given $m/z + \frac{1.006277}{2\times z}$ are multiplied by the intensities at $m/z - \frac{1.006277}{2\times z}$. These values are then summed across the whole m/z window selected, and the calculation is performed for each possible charge state entered. This results in peaks where the actual charge present is a multiple of the specified charge, since the isotope peaks will be separated by an m/z of $\frac{1.006277}{z}$ (Figure 3.10). The Fourier transform approach uses the fast Fourier transform (FFT) on the data when zero padded to the nearest power of two. The zero padding of the data allows for faster FFT algorithms to be used [207]. By Fourier transforming the data, peaks will be observed in the Fourier amplitudes at the charge states of the ions present(Figure 3.11). Fourier transform based methods have been shown to produce best results with higher charge states, and Patterson methods at lower charge states [206]. The popular algorithm THRASH [198] uses a combination of both these methods to give optimal charge state determination for all data as shown in (Figure 3.12).

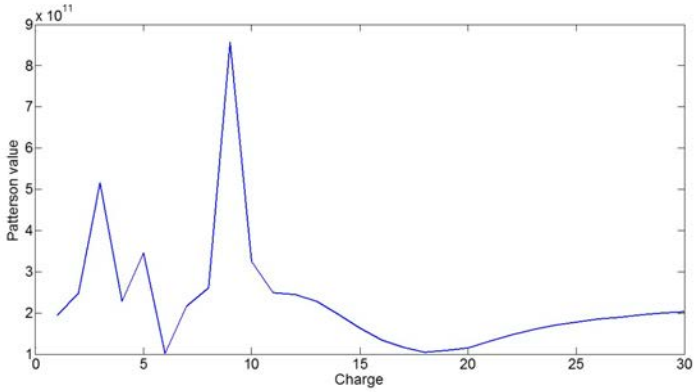

Figure 3.10: Patterson routine performed on the 1 Da window between m/z 952 and 953 containing the ubiquitin $[M + 9H]^{9+}$ with possible charge states of +1 to +30 used. There is a clear peak at a charge state of 9 where the theoretical and actual isotope peaks combine.

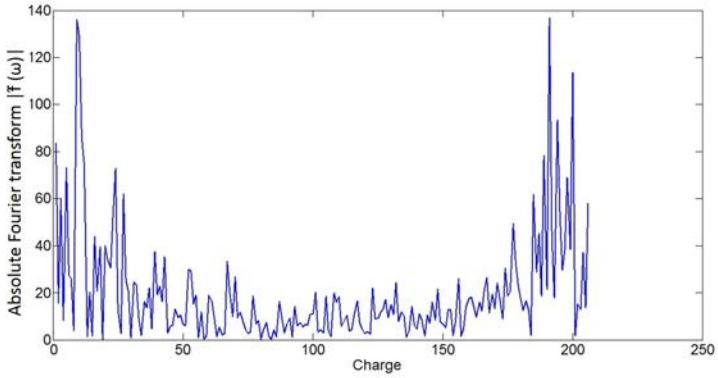

Figure 3.11: Absolute Fourier transform of the 1 Da window between m/z 952 and 953 containing the ubiquitin $[M + 9H]^{9+}$. As with the Patterson function, there is a clear peak at a charge state of 9, but also peaks at very high charge states.

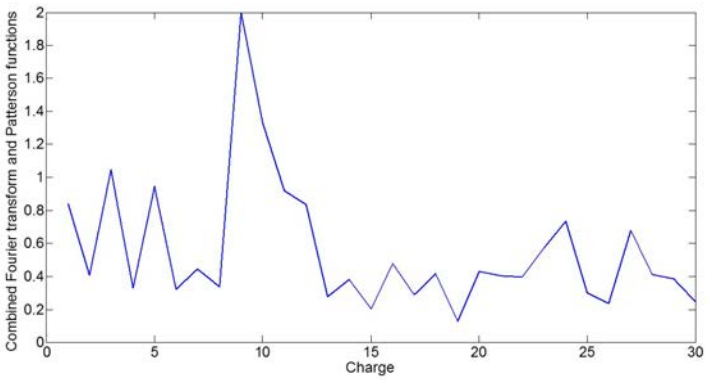

Figure 3.12: Combination of both Patterson and FT performed on the spectrum from figure 3.9 after scaling each one to between 0 and 1. By combining these two methods, the correct charge state is easily identified as 9+.

The charge state determination can be further improved through the use of a peak "folding" approach [208]. This has been shown to improve identification in low S/N or when overlapping isotope clusters are present. This approach selects a given peak, and "folds" the spectrum by multiplying together the lower m/z side of the peak by the higher m/z side of the peak. When pairs of peaks are equidistant from the targeted peak such as for isotope peaks, the intensity value of the "folded" peak will be significantly greater. To demonstrate this, a spectrum with much lower intensity peaks for ubiquitin $[M+9H]^{9+}$ was selected (Figure 3.13) and folded around the basepeak (m/z 952.6). This gave much higher intensities where the isotope peaks aligned, potentially allowing low S/N peaks to be deconvoluted (Figure 3.14). This then improves charge state determination of overlapping or low S/N isotope clusters as all other peaks will be multiplied by the baseline intensity.

Charge state determination allows the m/z axis to be converted to a mass axis (Figure 3.15). Following this, the isotope distribution of peaks needs to be converted to a monoisotopic mass peak. This can be done by performing a least squares fitting between the detected peaks, and that of the statistical distribution from *averagine*. Averagine is the mean mass of an amino acid based on the statistical occurrence of amino acids in a

large database of proteins. $C_{4.94}H_{7.76}N_{1.36}O_{1.48}S_{0.04}$ [209]. Some threshold for this fitting is then applied to determine if the protein is then present or not. If these criteria are met then the relevant peaks are then removed and the process repeated within the same m/z window. If not then the window is moved a given increment and the process repeated. This whole process is then repeated until the entire spectra has been processed [198]. This has been implemented in a number of software packages including DeconMSn [210] and YADA [211].

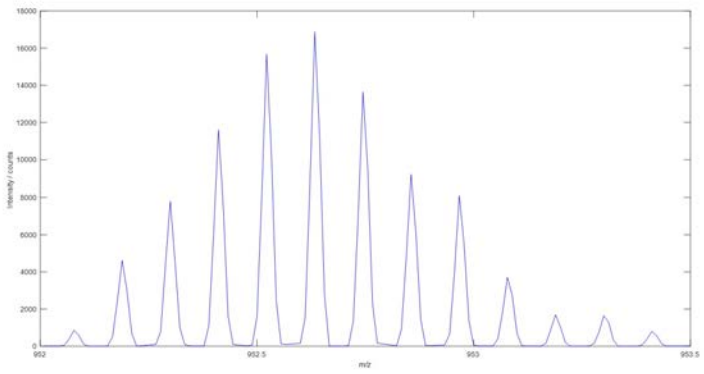

Figure 3.13: 1.5 Da window from m/z 952 to 953.5 for a lower intensity spectrum of ubiquitin $[M + 9H]^{9+}$.

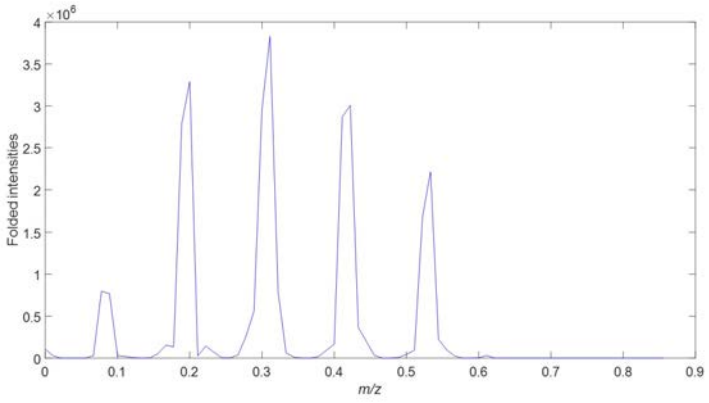

Figure 3.14: Target peak folding around the peak at m/z 952.6 from the spectrum in figure 3.13. The intensities for the isotope peaks are much greater, thus potentially identifying low S/N species.

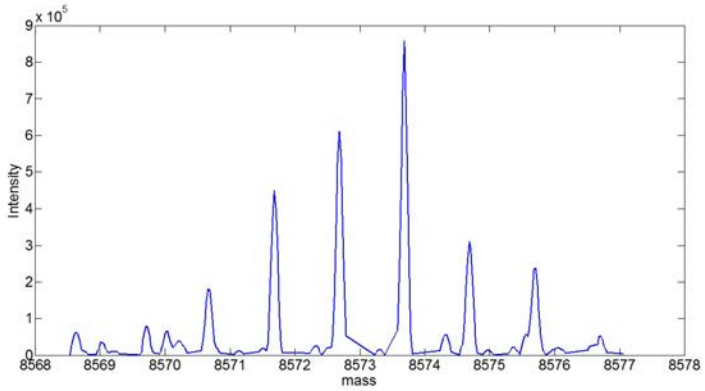

Figure 3.15: Spectrum of the m/z 952 region in figure 3.9 converted into a mass measurement rather than m/z by multiplying by the determined charge state.

For protein imaging experiments by LESA it is not realistic to manually select every spectrum for deconvolution, especially when deconvoluting every spectrum within the chromatogram for each pixel, such as when performing a 1D FAIMS sweep. The popular THRASH algorithm was implemented in Matlab (version R2014a and statistics toolbox, The Math-Works, Inc., Natick, MA, USA). This was also combined with the target peak folding method for charge state determination described by Chen *et al.* [208]. As described by Mayampurath *et al.* [210], the THRASH based algorithms significantly outperform the Xtract functions provided in the Xcalibur software in terms of number of peaks detected(Figure 3.16). It is possible however that these methods could just be more likely to result in false positive identifications. To assess this, it would be necessary to analyse these algorithms on well characterised samples.

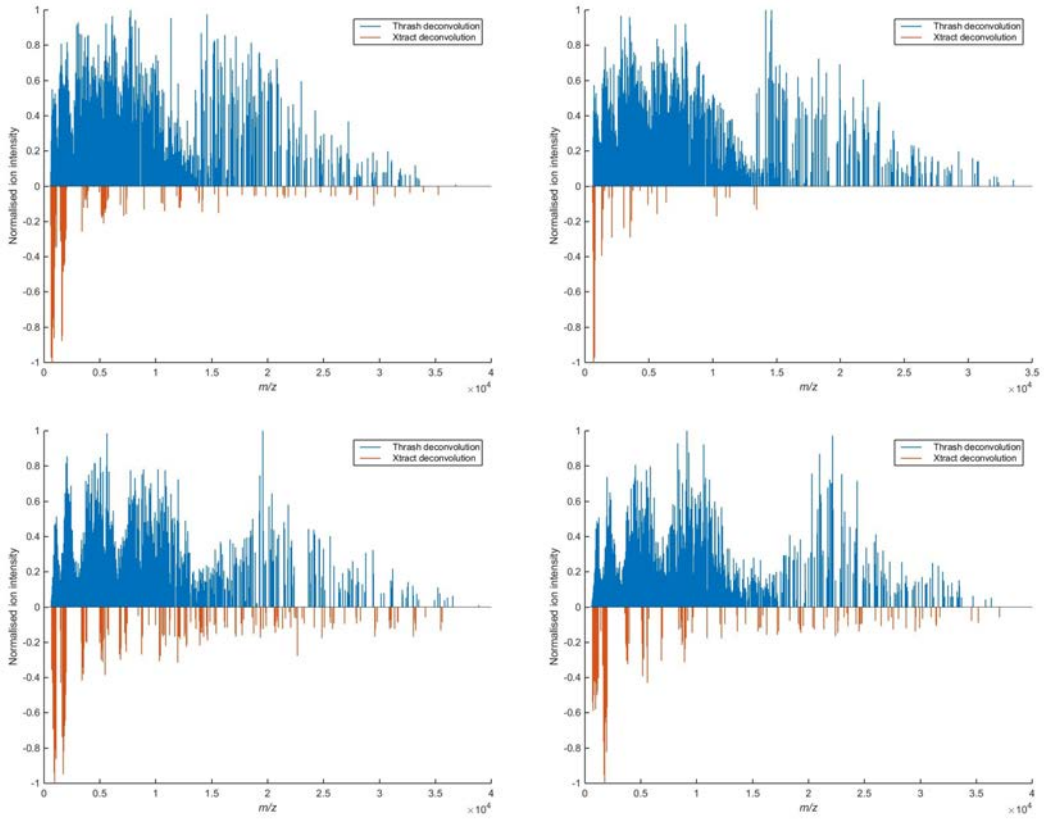

Figure 3.16: Example of THRASH vs. Xcalibur Xtract deconvolution on a number of spectra. The THRASH deconvolution identifies many more peaks than Xtract, however further investigation would be required to confirm that these are not false positive results.

The main limitation of these algorithms is the speed of their performance. Deconvolution of a single spectrum can take up to two hours with the THRASH algorithm, and while the process can be automated and parallelised, it is unrealistic to use this algorithm on even small imaging datasets, especially if each individual spectrum of a chromatogram is to be deconvoluted. If deconvolution is to be routinely performed on MSI data, a significant improvement in speed is required.

Comparison of LESA with and without the use of FAIMS

To compare the different deconvolution methods, and possible FAIMS conditions, a 4 by 4 pixel section of three serial sections of lambs brains were analysed by LESA, either

using a FAIMS 1D sweep per pixel, a series of stepped static conditions, or without the use of FAIMS separation. When comparing the results of spectral deconvolution on a LESA 1D FAIMS sweep dataset against a LESA stepped static FAIMS, and a standard LESA experiment, there are different orders that can be considered for data processing. For a standard LESA experiment, the chromatogram of each LESA extraction is typically summed together and the deconvolution performed on this [93]. Assuming that the ions are detected during every scan in the chromatogram, this will give the highest possible S/N for the data, thereby maximising the likelihood of the algorithm identifying the proteins present in the data. This is not necessarily suitable for the FAIMS 1D and stepped static experiments however, since some of the proteins will only be present in a few spectra of the chromatogram. Due to the slow speed of the THRASH algorithm, these data were deconvoluted using the Xcalibur Xtract function, and the effects of deconvolution performed on the whole summed chromatogram, were compared to deconvolution on each individual spectrum, wth the resulting spectra then summed.

When deconvoluting summed spectra of the chromatogram, some species can be seen that are not present when deconvoluting the individual spectra. Of particular note, a region in each of the three tissue sections where the peak at a mass of 15040 was identified which was tentatively assigned as a single haemoglobin subunit (Figure 3.17). The ion detected only when deconvoluting the individual data is the result of ions which are ubiquitously present throughout the chromatogram but at low intensities. Therefore when the chromatogram is summed, these signals combine to be above the noise. There are however ions which also can only be seen when the individual spectra from the chromatogram are deconvoluted, typically smaller ions around 1-3 kDa (Figure 3.18). This is further highlighted by performing PCA on these data, where there is a strong separation between the two processing workflows, and the spectral loadings show the mass 15040 to dominate the negative component where the data are summed before deconvolution, and peaks between 1-3 kDa for the positive loadings where the data is deconvoluted on the individual spectra (Figure 3.19). These peaks only present in the individual deconvolution are most likely

to be the result of ions that are transmitted at specific optimal FAIMS conditions and so are only present in a few scans. These become overtaken by the noise in a summed chromatogram but can be above the noise level in individual spectra. As a result of these two phenomena it is hard to define an optimal processing workflow. Ideally, each chromatogram could both be summed then deconvoluted, and deconvoluted then summed, and the results of these combined with some appropriate scaling. This would however be even more time consuming, further highlighting the need for efficient deconvolution methods, and would also require some appropriate method to scale intensities within the different methods which would complicate gaining any quantitative information from these spectra.

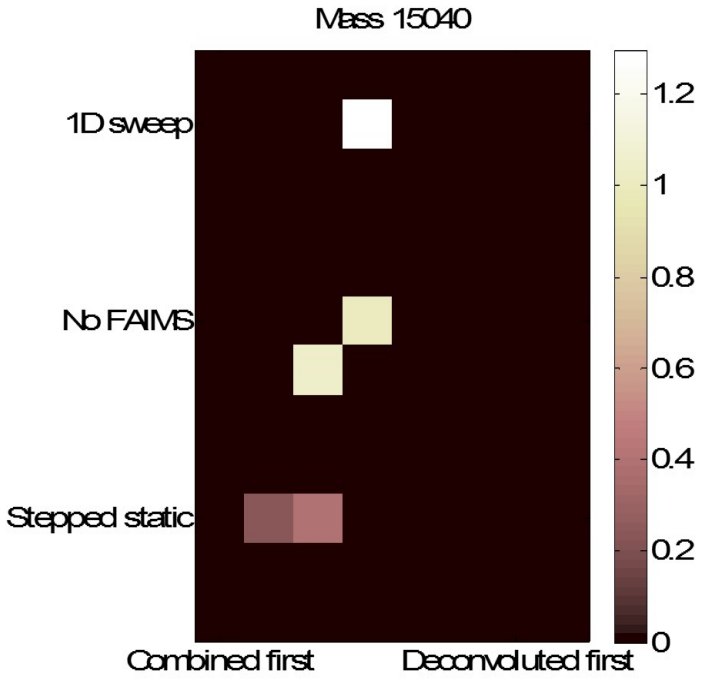

Figure 3.17: Comparison of the distribution of mass 15040 using FAIMS 1D sweep, stepped static, or no FAIMS. These data were either deconvoluted on the summed spectrum for each pixel, or on each individual spectrum per pixel then summed. This mass is only seen when the deconvolution is performed on the summed spectra.

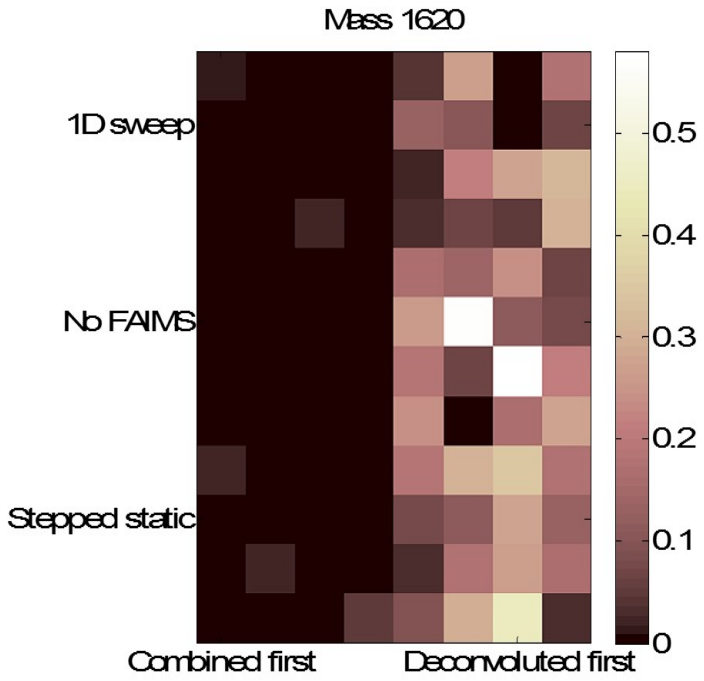

Figure 3.18: Comparison of the distribution of mass 1620 using FAIMS 1D sweep, stepped static, or no FAIMS. These data were either deconvoluted on the summed spectrum for each pixel, or on each individual spectrum per pixel then summed. This mass is only seen when the deconvolution is performed on the individual spectra in each pixel, suggesting that it is only transmitted under certain FAIMS conditions.

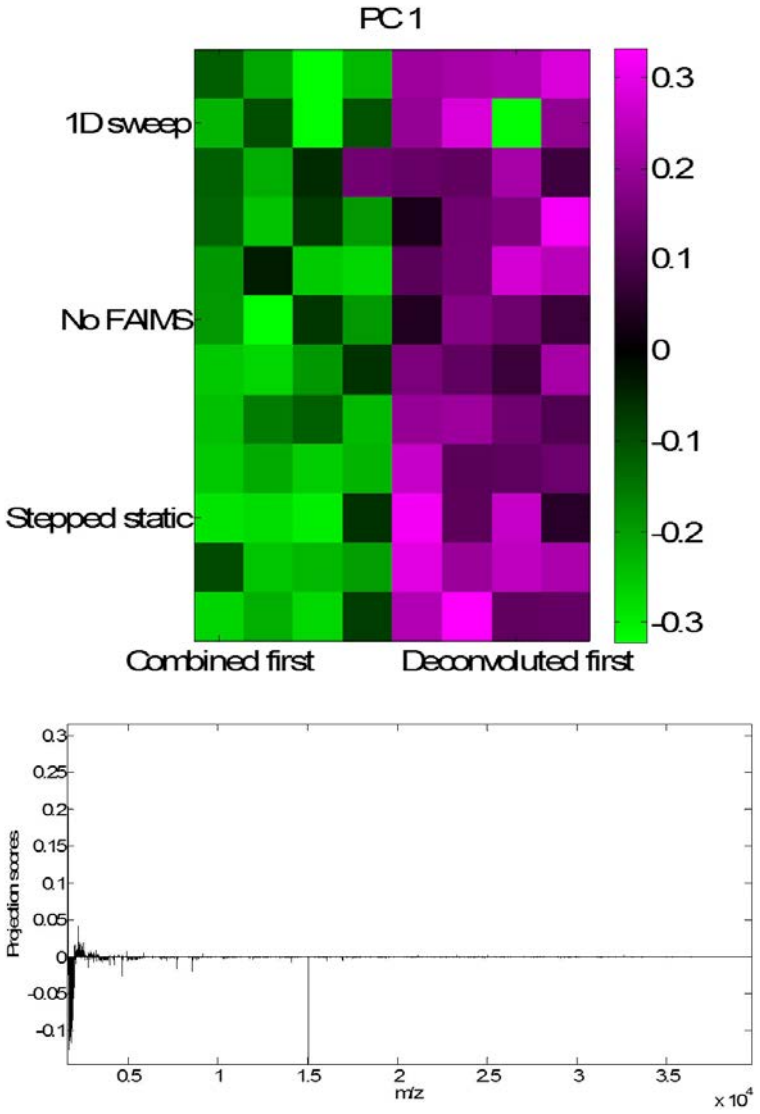

Figure 3.19: Principal component analysis on the FAIMS 1D sweep, stepped static, and no FAIMS data. This shows the primary source of variance as the difference in deconvolution method, and the spectral loadings are dominated by the mass of 15040 correlating with the negative regions in the image.

The other possible workflow that could be applied to the stepped static data is to deconvolute each of the individual sets of static conditions, and then to sum the resulting data together. This could potentially allow some of the very low S/N ions to be deconvoluted, but would not be summing together unnecessarily large portions of noise. To do this however would be extremely time consuming, particularly for large images, and so as

before, some method of automation would be required.

3.4 Conclusions

In conclusion, analysis of MSI data with additional FAIMS separation remains a challenge, but the data processing methods presented here allow a more comprehensive and detailed investigation of these data. The novel processing and conversion of FAIMS 2D sweep data into imzML format allows a huge number of different analyses to be performed, providing a deeper understanding of these data to be achieved. This could be used in the future for both rapid FAIMS conditions optimisation, and to further understand fundamentals of FAIMS mechanisms. While these processing methods allow for complex multivariate analysis to be performed, one challenge that still remains in MSI is the means to quantitatively evaluate these different machine learning approaches.

CHAPTER 4

CLUSTERING EVALUATION IN MSI

4.1 Introduction

As discussed in section 1.4, external clustering evaluation is preferable to internal evaluation as it can be used to compare algorithms based on different principles, and requires no assumptions about the data. The limitation for segmentation in MSI is that in order to perform external evaluation of different algorithms, a sample with a known spatial distribution is required. In most MSI experiments, biological samples are analysed, which will always have inherent unknowns and so no truly known spatial distribution of molecules is possible. This means that external evaluation of biologically derived MSI data is not possible. As a result of this; 1) an MSI sample with a known spatial distribution is required to allow external evaluation existing and new clustering algorithms against, and 2) a robust internal evaluation metric is required to accurately evaluate clustering performed on data where external evaluation measures are not possible.

LDI of inks produces characteristic mass spectra from associated pigment ions [212], and these inks can easily be printed into well-defined patterns. In this chapter, inkjet printing is presented as a means to generate samples with a known spatial distribution, to act as a reference for external evaluation of clustering algorithms. Where this is not possible, such as exploratory analysis, multivariate normality testing using chi squared quantile plotting is presented as a robust internal evaluation measure of algorithms that

assume normality within the data. The shape of these quantile plots can also be used to more deeply understand the distribution of these MSI data.

4.2 Experimental

4.2.1 Materials and methods

Ink patterns (Figures 4.1 and 4.2) were printed using an HP Z75200 PS printer, on Coala, Matt coated 180 gsm paper (Antalis, UK). These were fixed onto stainless steel MALDI target plates (AB Sciex, Warrington, UK) for LDI, and glass slides for DESI using double sided tape.

Mixed polymer samples were created by Naresh Kumar (NPL), briefly, 20 mg/ml solution of poly(methyl methacrylate) (PMMA) in chloroform was mixed with 20 mg/ml solution of polystyrene in chloroform in equal amounts. 50 μl of the mixed solution was spin-coated on clean silicon substrates at 3000 revolutions per minutes for 3 minutes to obtain the phase-separated polymer blends.

Coronal mouse brains were sectioned to 12 μm thickness, and thaw mounted onto, glass slide (Superfrost, Thermo Fisher Scientific, Waltham, MA USA), and sagittal rat brain, and mouse lungs were sectioned to 12 μm thickness and thaw mounted onto stainless steel MALDI target plates. These were then coated in MALDI matrix CHCA (5 mg/mL, 80 % MeOH 0.1 % TFA) using an automated pneumatic sprayer (TM-sprayer, HTX imaging, Chapel Hill, NC, USA).

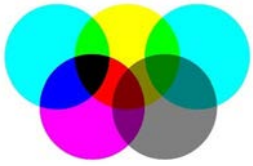

Figure 4.1: Overlapping circles ink pattern used for LDI imaging

Figure 4.2: Individual ink squares pattern used for LDI, and DESI imaging

4.2.2 Mass spectrometry imaging

Ink image acquisition

LDI images were acquired on a QSTAR XL QqTOF (AB Sciex, Warrington, UK), operating in raster imaging mode on medium speed (1 mm s^{-1}), with a pixel size of 200 μm^2. Data were acquired in positive ion reflectron mode with a mass range of 50-1000 Da. An Nd:YAG laser (355 nm, Elforlight, UK) was used to acquire the images, operating at a repetition rate of 1 kHz and attenuated to 10% of maximum power. DESI imaging was performed on an Orbitrap Velos (Thermo Fisher Scientific, Bremen, Germany), using a 2D Omni Spray Ion Source stage (Prosolia, Indianapolis, IN, U.S.A) set to a pixel size of 200 μm. The solvent system used was 50:50:0.1 (Acetonitrile/dimethylformamide (DMF)/Formic acid). Gas pressure was set to 150 psi, voltage 5 kv, a flow rate of 2 μl / min was used and a mass range of 200-2000 Da was used, at a resolution of 100000 FWHM at m/z 400.

Mixed polystyrene PMMA sample

SIMS images were acquired by Jean Luc Vörng (NPL) using an Ion-TOF IV instrument (Ion-TOF, Münster Germany). Experiments were performed in dual beam mode, using a 25 keV Bi_3^+ analysis gun and an argon cluster gun as the sputter gun (5 keV, Ar_{2000}). Depth profile were performed using the non-interlaced mode with (15 frames analysis, sputter 60 seconds, 1.0 second pause). To resolve any issues of sample charging, a flood gun was used (filament current 2.45 V). Depth profiles were performed over a $130\mu m \times 130\mu m$

area centered in a $500\mu m \times 500\mu m$ crater with a square raster of 512 pixels by 512 pixels and 1 shot per pixel. The ion beam currents were measured separately before measurement using a Faraday cup. The pulsed Bi_3^+ (12.5 ns pulse) current was 0.115 pA and the gas cluster ion beam current for 5 keV Ar2000 was 0.5 nA. Depth profile acquisition was performed in positive secondary ion mode. Spectra and depth profiles were calibrated using $H^+, C^+, CH_3^+, C_2H_5^+, C_3H_7^+$ for positive secondary ions.

MSI of biological samples

MALDI images of coronal mouse brain were acquired using a Synapt G2Si (Waters, Manchester, UK), using a pixel size of $45 \times 45\mu m$, and an m/z range of 100-1200 Da. Full description of rat brain image acquisition is detailed elsewhere [10]. Briefly, data were acquired on a QSTAR Elite QqTOF (AB Sciex, Warrington, UK) with a pixel size of $100 \times 100\mu m^2$, and a mass range of 50-1000 Da. Mouse lung data were acquired on a QSTAR XL QqTOF (AB Sciex, Warrington, UK), operating in raster imaging mode with a pixel size of $50 \times 50\mu m^2$ and a m/z range of 50-1000 Da.

4.2.3 Data processing and analysis

Preprocessing

Data processing was performed on an Intel Xeon quad core CPU E5-2637 v2 (3.50 GHz) with 64 GB of RAM. All data were converted from proprietary format to the mzML format using msconvert as part of ProteoWizard [197] software then into imzML using imzML-Converter [118].These were then imported into MATLAB (version R2014a and statistics toolbox, The Math-Works, Inc., Natick, MA, USA) using the SpectralAnalysis software package [191]. For the LDI data, and MALDI data acquired on the QSTAR instrument, a QSTAR specific zero filling was applied, followed by three applications of smoothing with a Savitzky-Golay filter with a window of seven and second order polynomial over the data. Any negative values created by the smoothing were removed, and the summed

spectrum was then generated from these data, and this was peak picked using a gradient method to give a total of 6,130 peaks. The ink regions of the data were then segmented from the background based on PCA scores images (Figure 4.3), and prior knowledge of the sample(Figure 4.4). Of note, there is some uncertainty in the assignment in some of the boundary regions between some inks and as such these were not given an assignment to prevent potential erroneous results. This uncertainty is most likely due to the fact that unlike in MALDI, LDI of inks does not fully ablate the material at a given pixel and as such, there is a large amount of overlap signal in a given pixel from the previous one. Spectra from the simple squares pattern DESI data were zero filled using the orbitrap specific zero filling from SpectralAnalysis [191], and the basepeak spectrum was peak picked using the Matlab "mspeaks" function from the bioinformatics toolbox. Since this gave $80,000$ peaks which would be too large to calculate a covariance matrix from, this was further reduced to 10,000 dimensions using random projection orthogonalised by QR decomposition [133]. The basis for external evaluation was segmented from ion images (Figure 4.5), and prior knowledge of the sample (Figure 4.2) to give a final segmentation (Figure 4.6). For data acquired on the Synapt instrument, spectra were zero filled using interpolated rebinning between m/z 50 and 1000 with a bin width of 0.01 Da. A total spectrum was then generated, and peak picked using a gradient approach. The area under each peak was then extracted from the unprocessed data to give the final dataset for clustering and normality testing.

Clustering

Following preprocessing, k-means clustering was performed on the data using the MATLAB "kmeans" function (version R2014a and statistics toolbox, The Math-Works, Inc., Natick, MA, USA), with $k = 2 \rightarrow 20$ with the Euclidean, cosine and correlation distance metrics (Figure 1.11), three replicates and random starting clusters. The results were then evaluated using four internal evaluation metrics (Calinski-Harabasz, Davies-Bouldin, Dunn and Silhouette from section 12) using both standard Euclidean measures

of similarity, and using the measures of similarity used in the clustering itself. Alongside internal evaluation, two external metrics (Rand and Jaccard indices as seen in table 1.5) were also used as a gold standard to compare the results [188]. Normality testing was performed on the data assigned to each cluster by plotting the squared Mahalanobis distance from each pixel to the distribution within its cluster against a chi-square distribution with a number of degrees of freedom equal to the dimensions of the data [213]. The Mahalanobis distance for the data within each cluster was calculated by first performing PCA, then in cases where the number of dimensions exceeds the number of pixels, any resulting fully zero components that would have zero variance are removed, and all remaining components are scaled such that each component has a standard deviation of 1. The squared Euclidean distance of each pixel to the mean of its assigned cluster is then calculated to give the Mahalanobis distance [214]. For data clustered using the cosine distance metric, data were first converted into a polar coordinate system, comprising of a distance from the origin r, and a series of angles from the origin θ_{n-1} relative to each of the coordinate axes where n is the dimensionality of the data [215]. The angles from the co-ordinate axes were then used to determine normality of the angular distribution. For creating the plots the Mahalanobis distance and chi-squared values were all rescaled to between 0 and 1 in order to plot them all on a common axis.

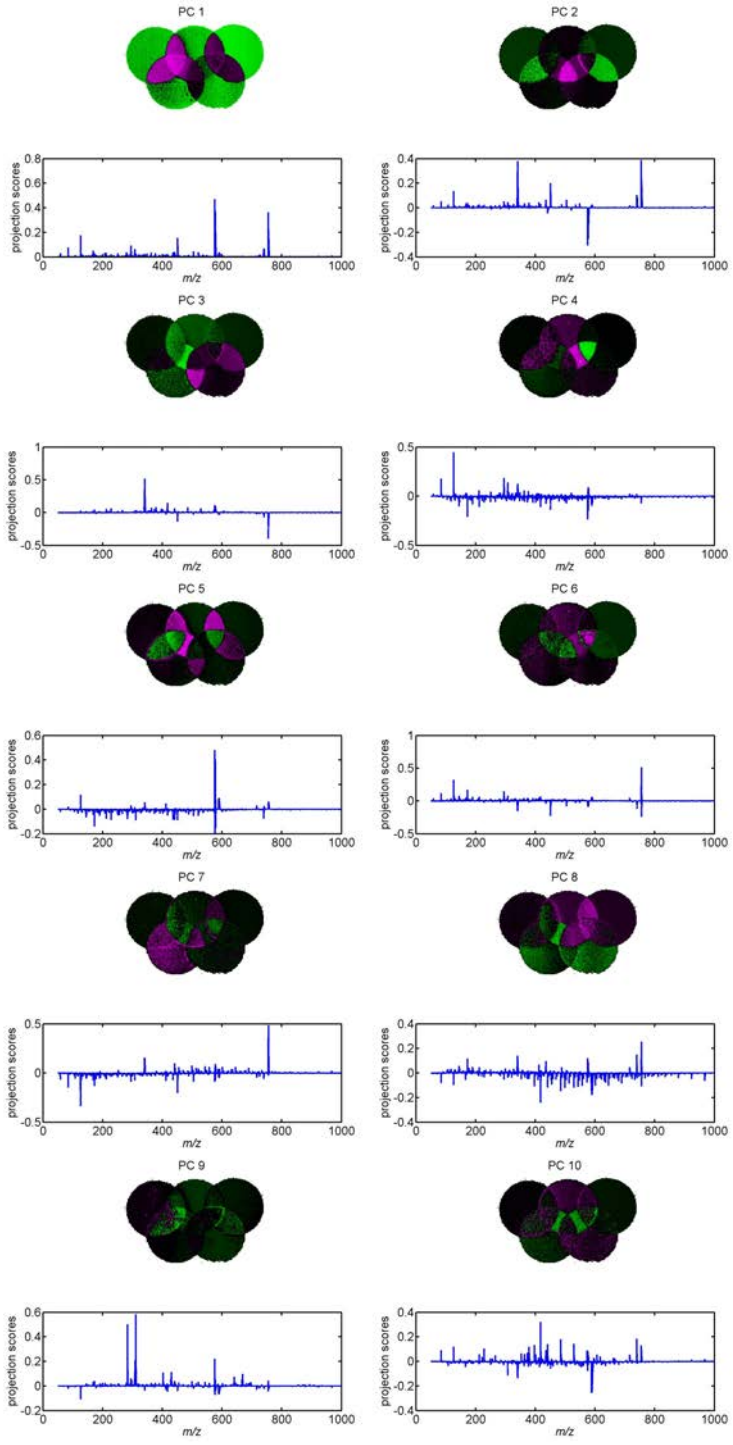

Figure 4.3: First 10 components from PCA on the LDI image of the circular pattern showing distinction between the different ink regions. The regions identified by this are then used as the basis for the ground truth of spatial distribution of these inks.

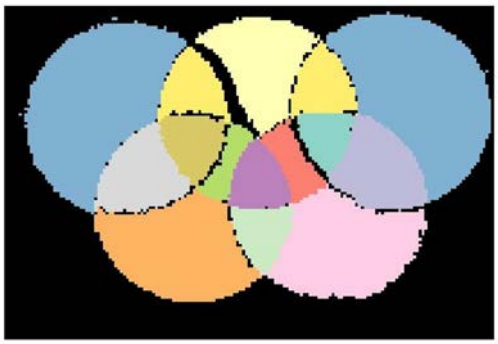

Figure 4.4: Final segmentation for the 13 different ink regions analysed by LDI used as a basis for external evaluation.

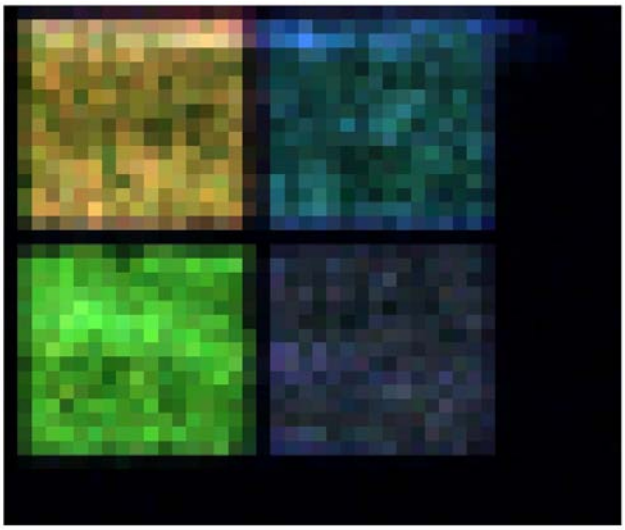

Figure 4.5: Overlay of ion images from m/z 855.178 ± 0.017 (red), 855.547 ± 0.025 (green), and 531.430 ± 0.046 (blue), showing the four different ink regions in figure 4.2.

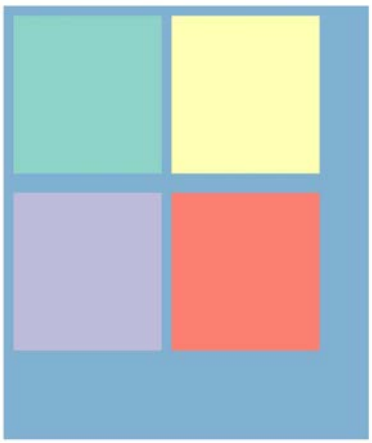

Figure 4.6: Final segmentation for the 4 different ink regions analysed by DESI used as a basis for external evaluation.

4.3 Results and discussion

4.3.1 Printed ink standards

Inkjet printers generally consist of four ink cartridges, cyan, magenta, yellow, and black, each which give distinct spectral patterns when analysed by LDI or DESI (Figure 4.7). Through combination of these inks, a total of 12 different regions can be generated, and in the examples shown there is an additional grey pigment which has a different composition from both pure black and the individual cyan, magenta and yellow (Figure 4.7 e and m) resulting in a total of 13 spectrally distinct regions (Figure 4.1). In some cases the exact composition of the pigments present in the inks may be known [212], but this is not always the case. For the purpose of evaluating spatial clustering however, the spectral identities need not be known, merely that each colour gives a distinct spectral pattern. MALDI-MSI of a simple pattern of the four constituent inks show distinct spectra for each colour, and peaks which are likely to be consistent with pigment ions from the ink. In the case of cyan ink, the peaks at m/z 575-578 (Figure 4.8) can be attributed to the same pigments observed by Donnelly et al., as seen by the distinct copper isotope pattern arising in the

spectrum. LDI-MSI of a simple pattern of these four inks (Figure 4.1) produces good quality images resulting from ions related to these pigments (Figure 4.9).

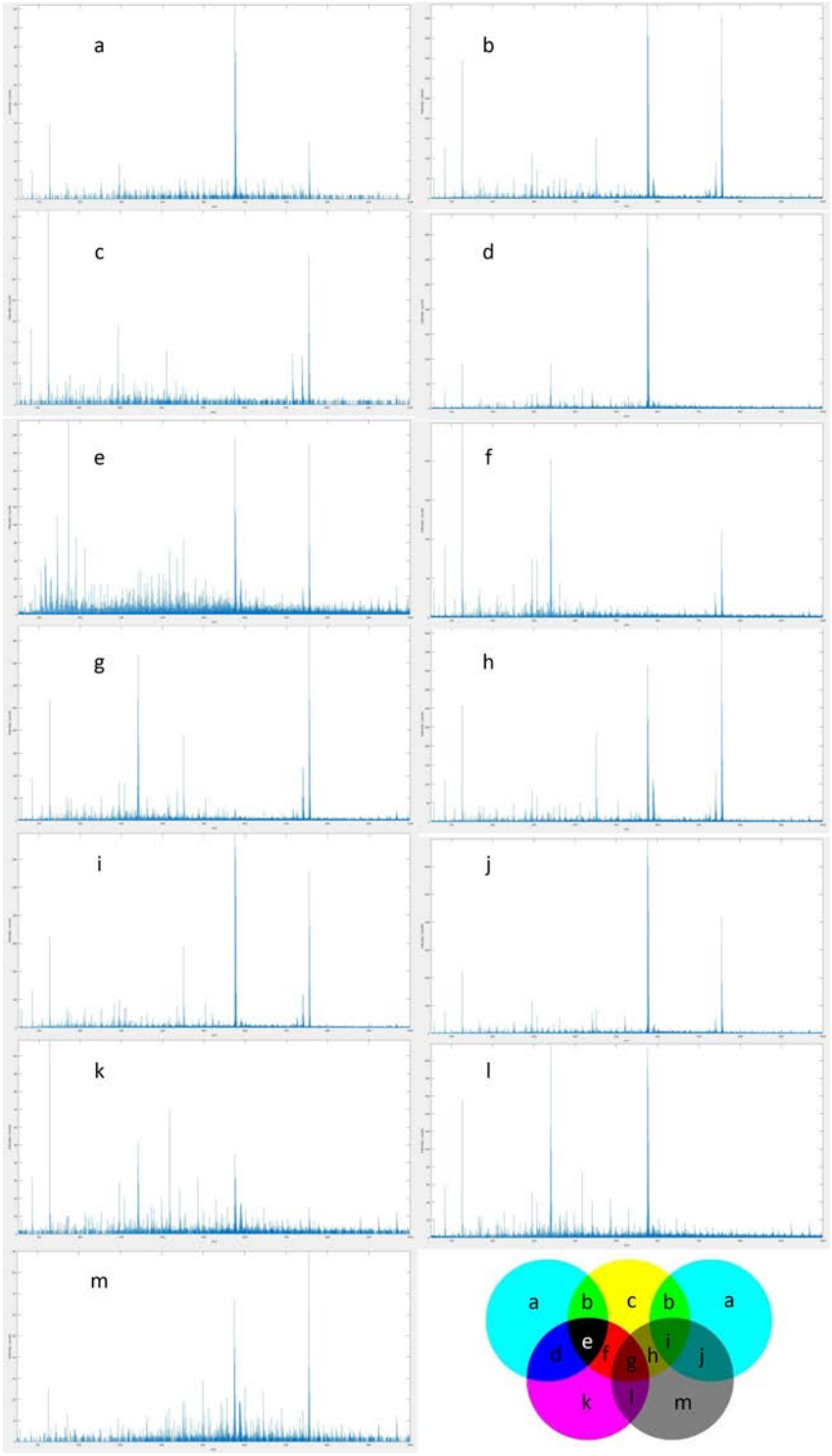

Figure 4.7: Example single pixel spectra from LDI of the thirteen different ink regions (a-m) showing clearly different spectral profiles for all thirteen different ink regions.

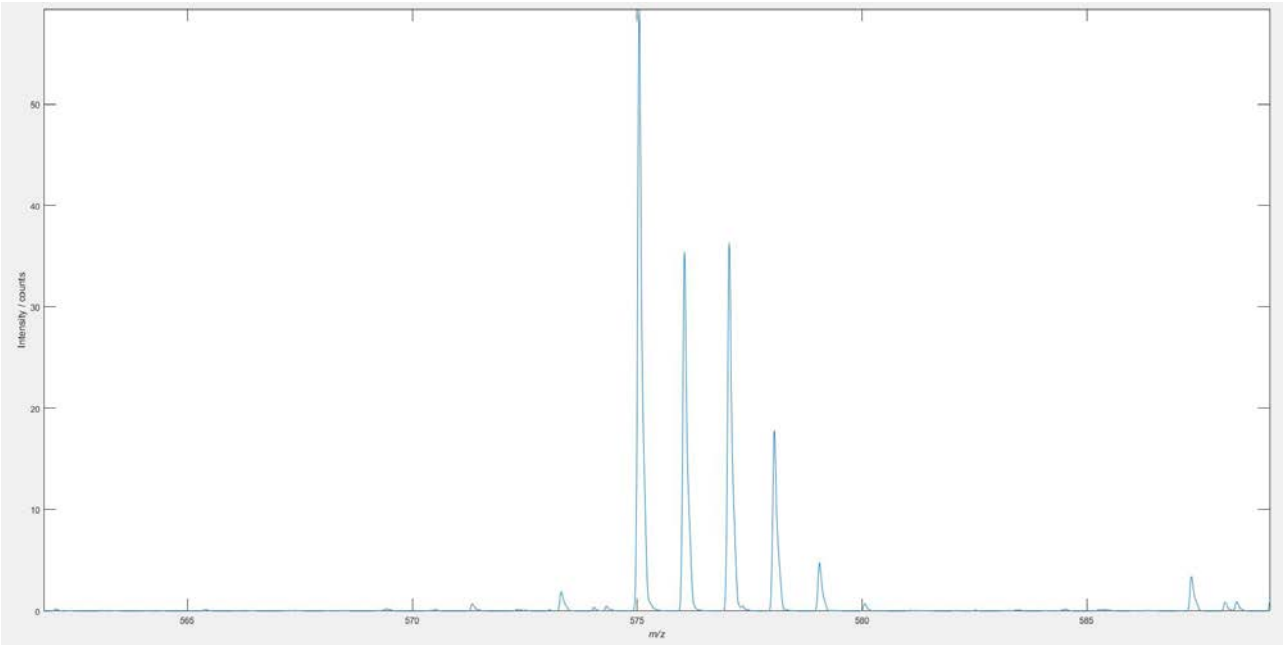

Figure 4.8: Zoom in around m/z 575 from the mean spectrum of the cyan ink data showing the peaks corresponding to copper phthalocyanine [212]

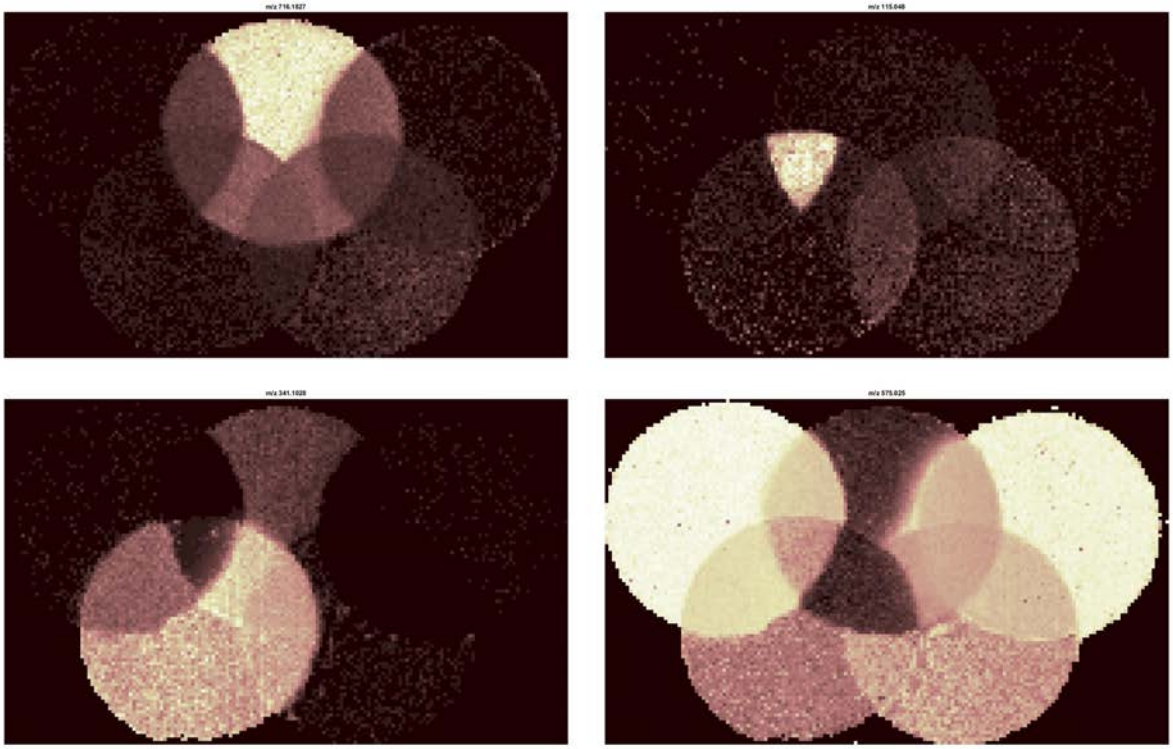

Figure 4.9: Ion images of m/z values which correspond to distinctive peaks for the yellow (top left), black (top right), magenta (bottom left) and cyan (bottom right) inks.

To further confirm the discrete spatial locations of the inks, along with the differences between the spectra, dimensionality reduction by PCA and t-SNE was used to reduce the data to three dimensions. These were then visualised as the three red, green and blue colour channels in an overlay image as described by Fonville *et al.* [122]. As well as this, these three dimensions were viewed as scatter plots to show the distribution in these three dimensions. The t-SNE in particular shows clear distinction between the different inks, while there is less separation of the primary ink spectra observed by PCA (Figure 4.10 to 4.13). Nevertheless, this shows that the spectra from these inks are suitably unique to act as image standards for external evaluation of clustering algorithms.

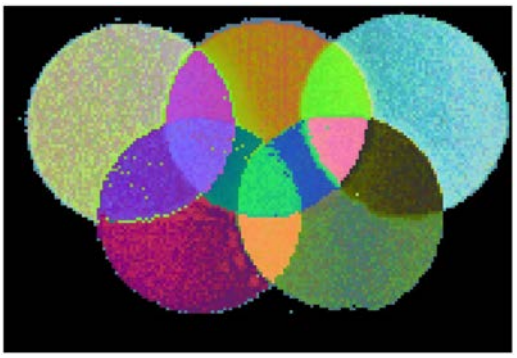

Figure 4.10: Image of the LDI image of the inks reduced to three dimension by t-SNE visualised by the methods described by Fonville *et al.* [122] showing good separation between all the different ink regions.

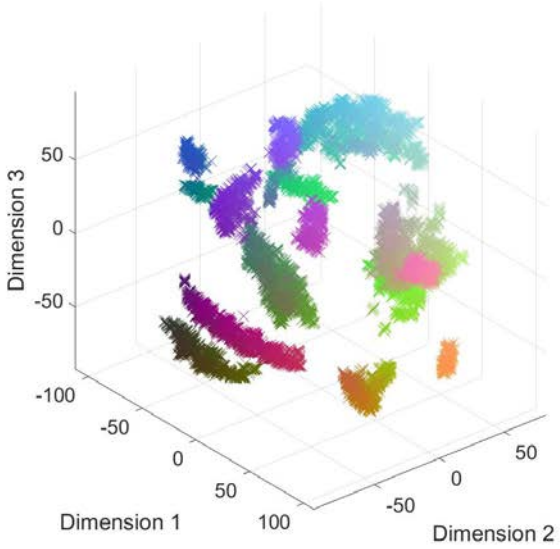

Figure 4.11: Scatter plot of the LDI image of the inks reduced to three dimension by t-SNE showing good separation between the different inks.

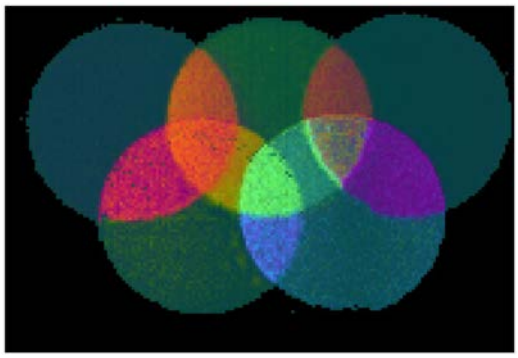

Figure 4.12: Image of the LDI image of the inks reduced to three dimension by PCA visualised by the methods described by Fonville *et al.* [122] showing good separation between the colours from a combination of inks but not distinctly separating the individual ink regions.

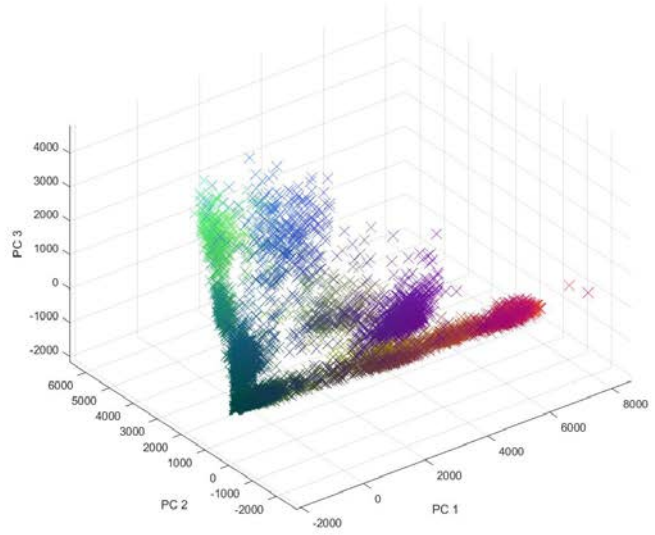

Figure 4.13: Scatter plot of the LDI image of the inks reduced to three dimension by PCA showing good separation between the colours from a combination of inks but not distinctly separating the individual ink regions.

Initial visual evaluation of the result of k-means clustering on the ink data shows that most of the different regions are identified, but in very few cases are all 13 different ink regions segmented accurately (Figure 4.14). This could be occuring because in some of

the mixed in examples (such as the blue region), one of the inks dominate the spectra (cyan) and there is very little intensity from the magenta ink (Figure 4.7 a, d and k). The use of the Rand index to evaluate the clustering results on these data shows a generally increasing accuracy with cluster number up to around $k = 15$, where there is a very slight drop. At low values of k (where $k < 8$) the cosine and correlation distances perform better than the Euclidean, however above this, the Euclidean distance performs better(Figure 4.15). The main limitation of the Rand index in the context of MSI clustering is the very large sample numbers involved (each pixel is one sample), this results in a very high number of true negative results for each cluster, which overshadows the effects of the other parameters. This is the likely reason why there are only very small changes in the Rand index at $k > 10$ despite the clustering results varying dramatically. Unlike the Rand index, the Jaccard index does not consider the true negatives, and therefore is more suitable as an external evaluation metric for MSI data. Like the Rand index, the Jaccard index for the clustering results on the ink data show more accurate clustering using the cosine and correlation distances at low values for k, and more accurate using the Euclidean for higher values of k. Unlike the Rand index however, the Jaccard index drops off much more dramatically at higher values of k, and is much lower than the Rand index, despite both indices returning values between 0 and 1 (Figure 4.16). This reflects the inability to trully segment all thirteen ink regions successfully in most results, and is more representative of the success of the clustering on these data (Figure 4.14).

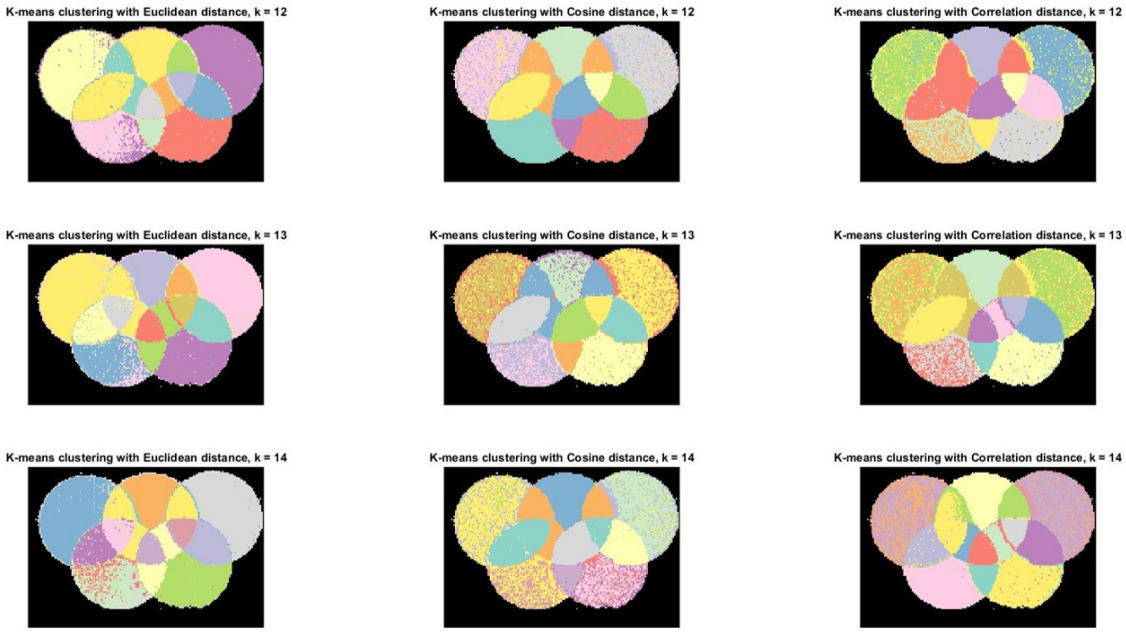

Figure 4.14: Results of k-means clustering on the ink data with Euclidean cosine and correlation distances with k = 12, 13 and 14. All of the images show segmentation of most of the features, but no case accurately segments all thirteen ink regions.

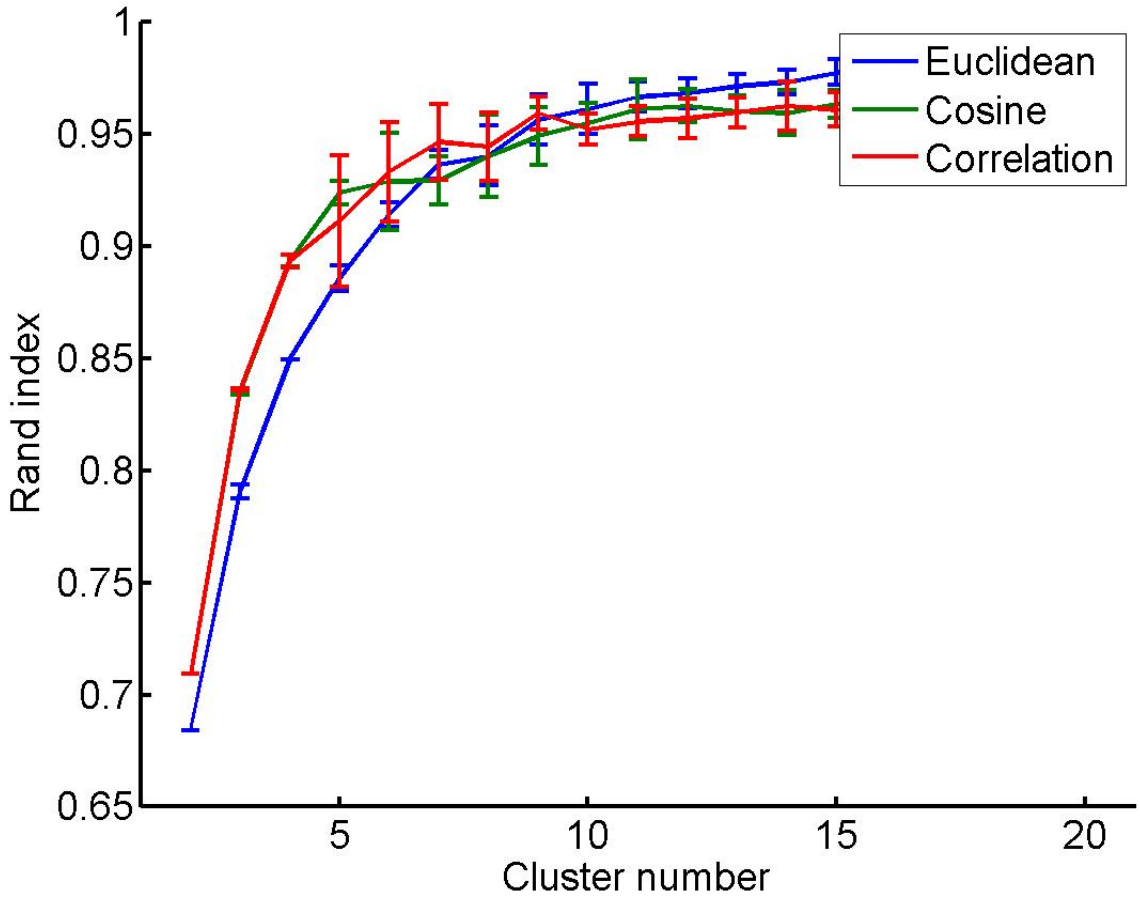

Figure 4.15: Rand index on the results of k-means clustering on the ink data using $k = 2 - 20$ and the Euclidean, cosine and correlation distances.

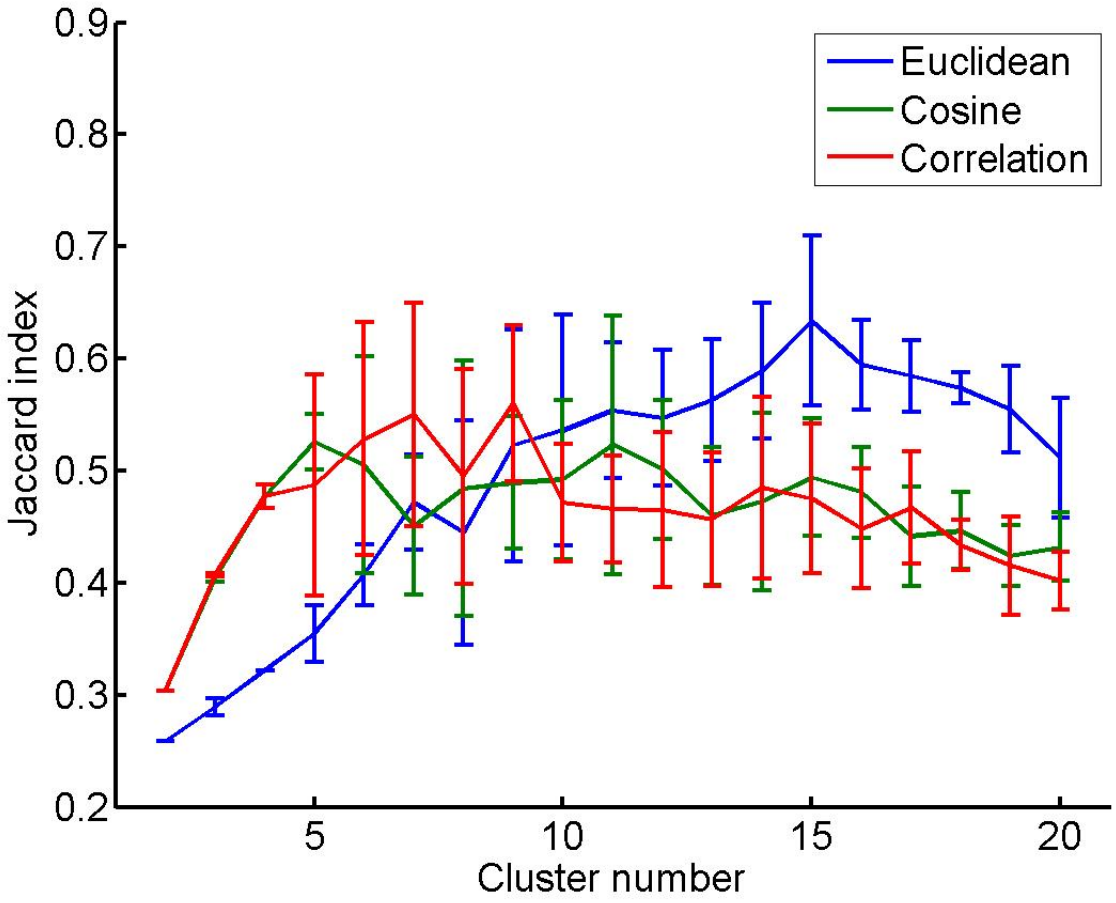

Figure 4.16: Jaccard index on the results of k-means clustering on the ink data using $k = 2 - 20$ and the Euclidean, cosine and correlation distances.

Compared to the data acquired from inks by LDI, imaging inks by DESI produces spectra with many more peaks, which are most likely derived from polymeric species from either the paper or ink themselves [199] (Figure 4.17). The result of this is that unlike the LDI images, the DESI images are not already well separated when reduced to three components via PCA (Figure 4.18 and 4.19) but are by t-SNE (Figure 4.20 and 4.21). In this case, clustering these data by the k-means algorithm produces much more accurate results when the cosine distance metric is chosen rather than the Euclidean, although the Euclidean distance performs better when l_2 normalisation is applied (Figure 4.22). Of note however, the commonly applied TIC normalisation does not improve the clustering results using the Euclidean distance in this dataset (Figure 4.22c).

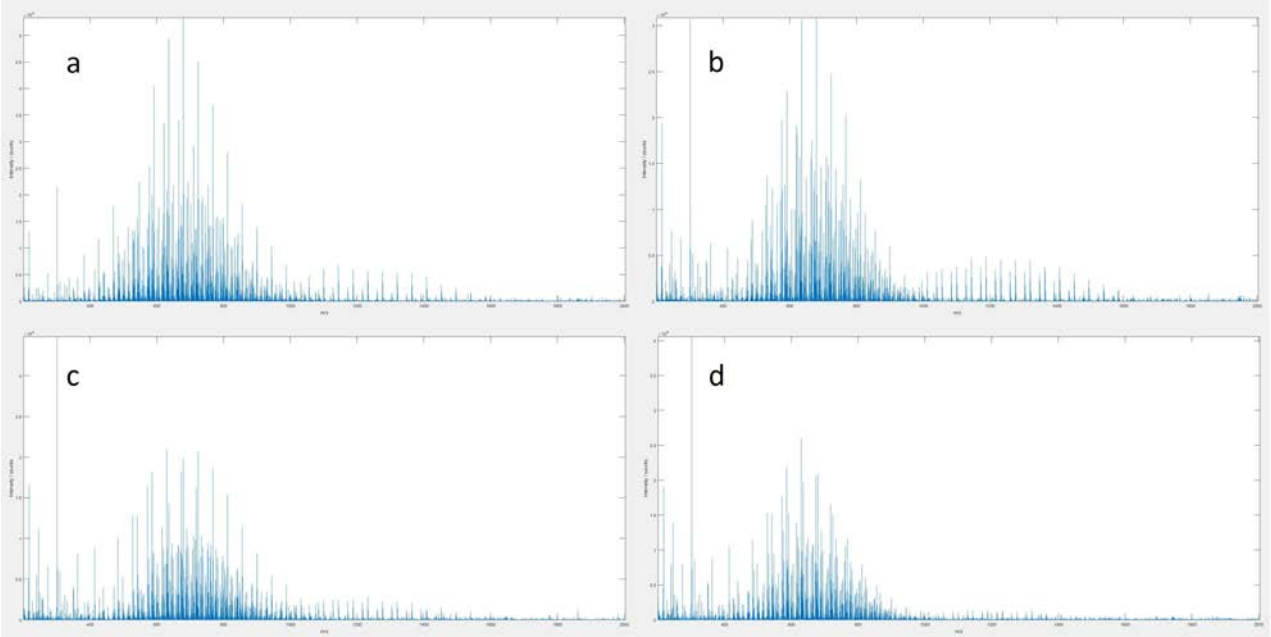

Figure 4.17: Spectra from the yellow (a), cyan (b), magenta(c) and black (d) inks analysed by DESI. Unlike the LDI spectra, these spectra are not dominated by a single peak, but show many regularly spaced peaks, most likely arising from polymeric species.

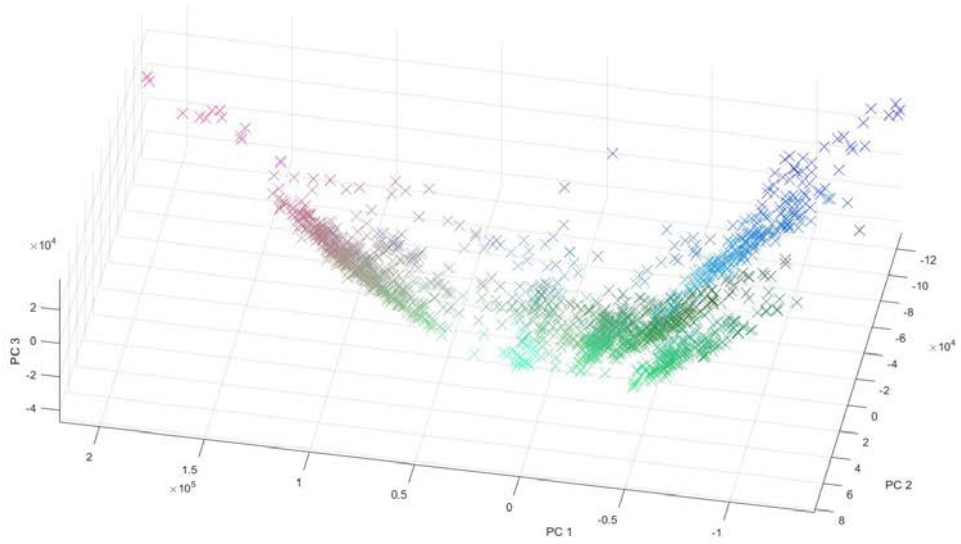

Figure 4.18: Scatter plot of the DESI image of the inks reduced to three dimension by PCA showing only minimal separation between the different inks and background.

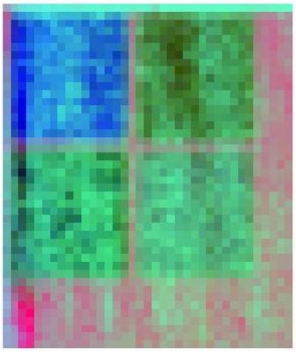

Figure 4.19: Image of the DESI image of the inks reduced to three dimension by PCA visualised by the methods described by Fonville *et al.* [122] showing only minimal separation between the different inks and background.

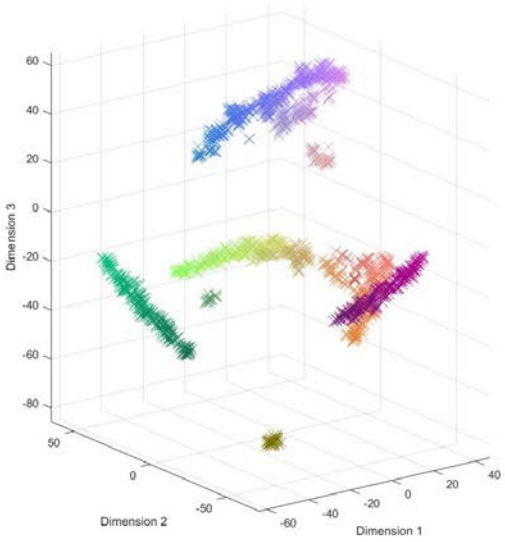

Figure 4.20: Scatter plot of the DESI image of the inks reduced to three dimension by t-SNE showing good separation between the different inks and background

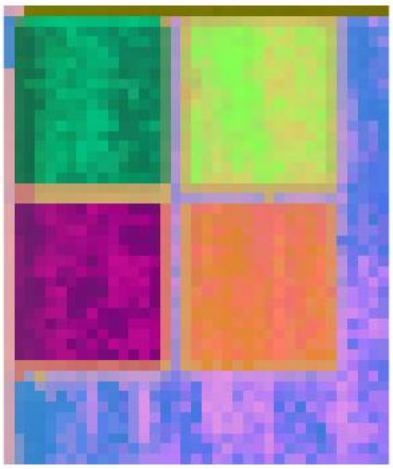

Figure 4.21: Image of the DESI image of the inks reduced to three dimension by t-SNE visualised by the methods described by Fonville *et al.* [122] showing good separation between the different inks and background.

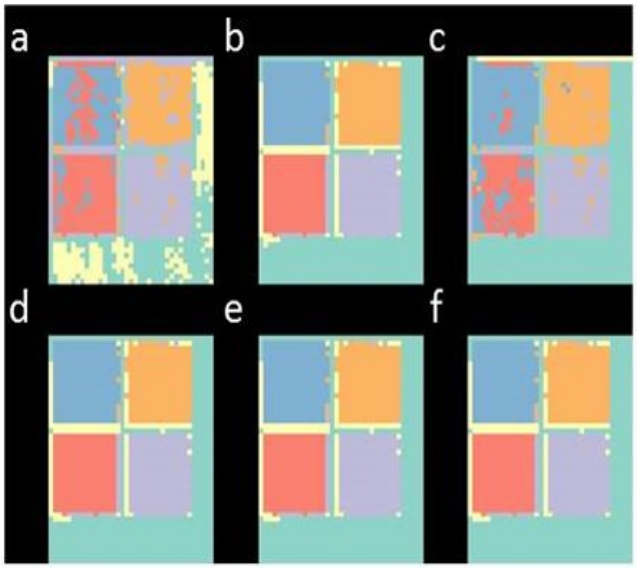

Figure 4.22: k-means clustering results for DESI image ($k = 6$) with Euclidean (top) and cosine (bottom) distances, left to right, un-normalised, l^2, and TIC normalised.

While external evaluation offers a powerful unbiased means to evaluate clustering results, the requirement for a ground truth prevents its use in any MSI applied to biological

samples where a ground truth is not possible. Internal clustering metrics which evaluate the data against itself do not require this, and could be a means to overcome this problem, however their use in MSI to date has proved inconclusive [120, 185, 187]. There are a number of possible factors which could cause this; the high dimensionality of MSI data means that Euclidean based measures of distance will converge to infinity, and the data are extremely sparse. Most evaluation metrics are based on Euclidean measures such as sum of inter-cluster distances, thus are not always appropriate when using other distance metrics to cluster with.

4.3.2 Normality testing

One of the primary assumptions of the k-means clustering and other algorithms is that the data within clusters are normally distributed. Previously, in other fields, methods have been used to evaluate whether the data within clusters is normally distributed to evaluate the clustering performance [216], or to determine whether to continue to divide clusters further [217,218]. By evaluating the degree of normality within the clusters, when clustering with an algorithm that assumes normality, it is possible to evaluate how well the data fits this assumption and thus how appropriate it is. For univariate data, normality testing is relatively straightforward, and there are a number of tests for normality such as Shapiro-Wilks [219], Kolmogorov-Smirnov [220], and Cramer-Von Mises [221] tests. This is more challenging in multivariate data since there will be many dimensions each with different variance and means [222]. It is possible to test for multivariate normality however using quantile-quantile plots [213]. If the data is multivariate normal then the Mahalanobis distance will have a χ^2_p distribution [223]. Therefore plotting the Mahalanobis distance from each pixel to its parent distribution versus a χ^2_p distribution where p is the dimensions of the data will give a straight line if the data is multivariate normal. This however, remains a Euclidean based measure of normality, and as such is still not suitable for clustering performed with the cosine distance. This could be resolved however by converting the data to polar space prior to normality testing. Unlike Cartesian

space, where each data point is represented by a series of distances from the origin along each axis, in polar space, a single distance from the origin r and a series of angles from reference lines.

Synthetic datasets were generated to simulate in two dimensions, data that are normally distributed in Cartesian and polar coordinates respectively (Figures 4.23, 4.24, and 4.25). Clustering was then performed on these datasets using k-means with the Euclidean and cosine distances, and subsequent normality testing of the clusters was performed. Data that are normally distributed in polar coordinates accurately cluster using the cosine distance, and the subsequent chi squared quantile plots have an r^2 that is close to 1 (Figure 4.23). In contrast, the use of the Euclidean distance result in poor segmentation and a lower r^2 in the subsequent chi squared quantile plots (Figure 4.23). When considering Cartesian distributed data, if the angles from the origin of the clusters of data are well separated, as in figure 4.24, then the use of either Euclidean or cosine distances results in accurate clustering and high r^2 values (Figure 4.24). However, if the angles of the clusters are not well separated as in figure 4.24 then the use of the cosine distance results in poor clustering and subsequent lower r^2 values in the chi squared quantile plots (Figure 4.24). This indicates that the use of normality testing, along with conversion to polar coordinates where the cosine distance is used, is a potential means to evaluate k-means clustering for MSI data.

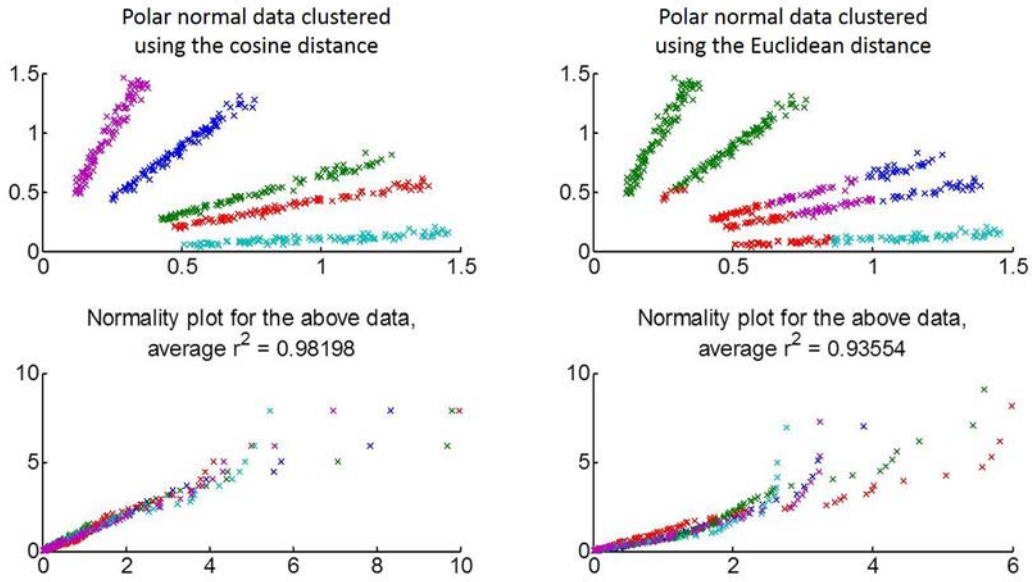

Figure 4.23: Clustering and normality testing using Euclidean and cosine distance on two dimensional simulated data that is normally distributed in polar space.

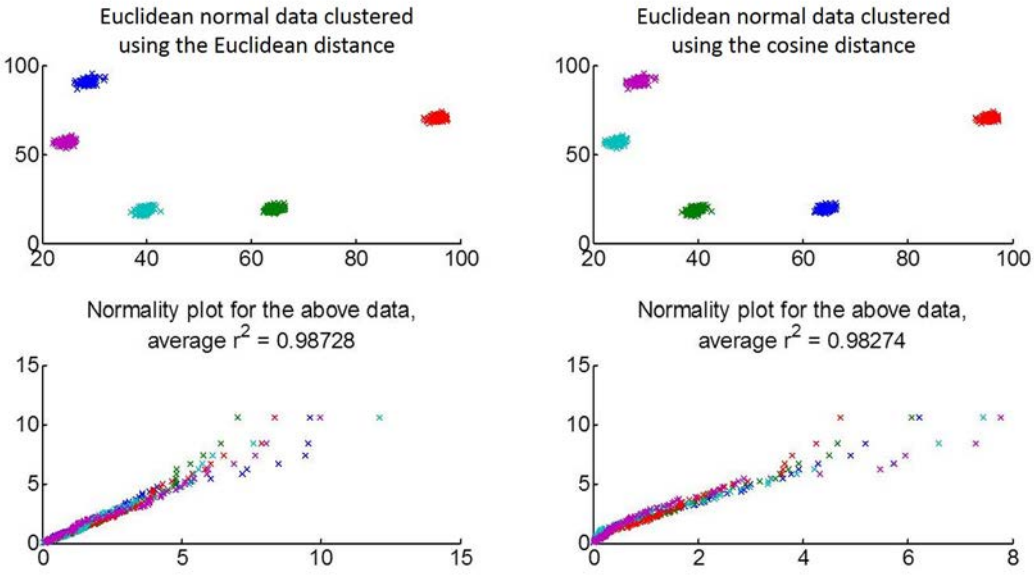

Figure 4.24: Clustering and normality testing using Euclidean and cosine distance on two dimensional simulated data that is normally distributed in Cartesian space where each cluster is also well separated in terms of angles from the origin.

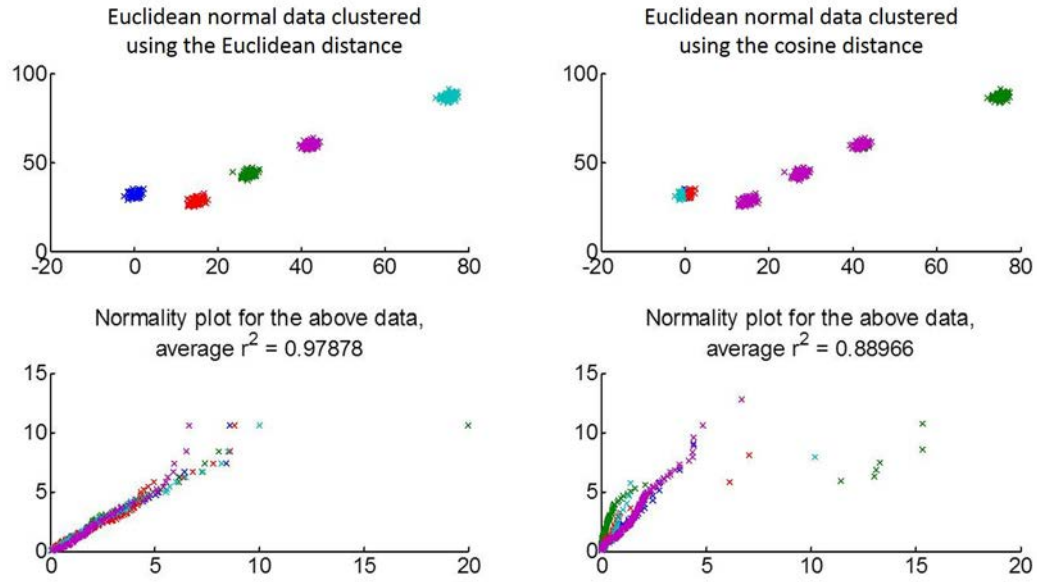

Figure 4.25: Clustering and normality testing using Euclidean and cosine distance on two dimensional simulated data that is normally distributed in Cartesian space where each cluster is not separated in terms of angles from the origin.

Normality on ink datasets

As with the other internal measures of evaluation, multivariate normality testing on the ink clustering results using Euclidean and cosine distances does not produce results consistent with the external evaluation. However, when normality testing was performed on each of the manually segmented region of the data in both Euclidean and polar space, these data can be considered closely approximate to normally distributed in polar space (average $r^2 = 0.9871$) but not so in Euclidean (average $r^2 = 0.7848$)(Figure 4.26). When reduced to three dimensions using PCA, the ink data is much more separated than data from biological MS images such as sagittal rat brain and mouse lung (Figure 4.13 and 4.27). This could be because there are far fewer overlapping peaks within different regions of the ink data, whereas biological samples will generally contain mostly the same peak list but at varying intensities. The resulting ink regions also contain data that appear both normally distributed in Euclidean (light blue regions), and polar space (green, purple and

red regions). This explains why no distance metric successfully segments all 13 regions consistently since algorithms such as k-means are prone to local minima (Figures 1.13). It also explains why the internal evaluation metrics perform poorly on these data. In order to more appropriately evaluate clustering algorithms for MSI, a better model sample is required that can more accurately reflect some of the characteristics of a biological sample.

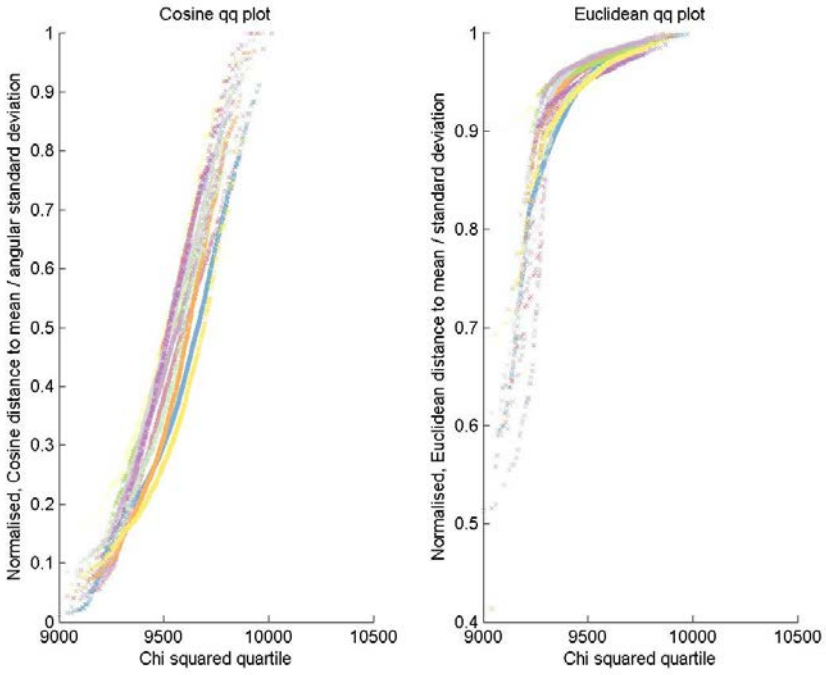

Figure 4.26: Quantile-quantile plot for the thirteen different ink regions in either polar space (left) or Euclidean space (right). The data is much closer to normal in polar than Euclidean space with a polar r^2 of 0.9871, and a Euclidean r^2 of 0.7848.

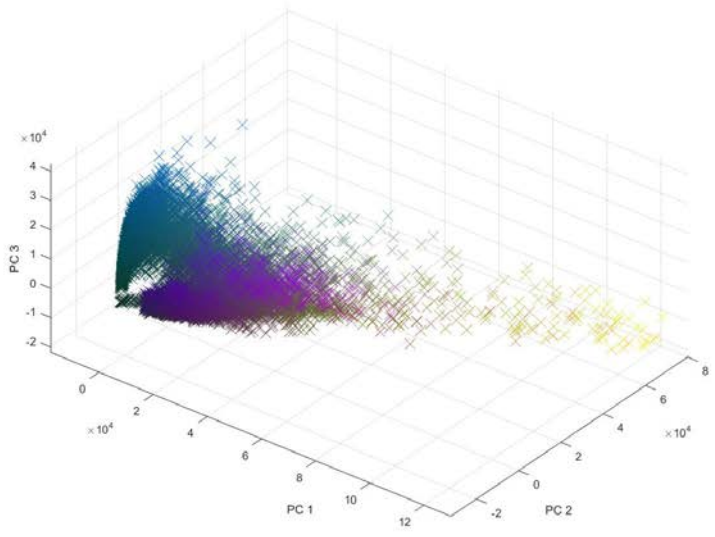

Figure 4.27: Scatter plot from the first three principal components of the sagittal rat brain dataset. This shows a much higher density than the ink data (Figure 4.13, and only really separates matrix from background, and white from grey matter.

Internal evaluation on mixed polymer sample

To analyse a sample with known composition and spatial distribution, a mixed sample of polystyrene and PMMA was analysed using SIMS (Figures 4.28 and 4.29). K-means Clustering was performed with two to ten clusters, using the Euclidean and cosine distances. Unlike the MALDI images from biological samples, the clustering results on this data appear better when using the Euclidean distance rather than the cosine (Figure 4.30). This could be attributed to the more linear response in detected ion intensity in SIMS from non biological samples, and so the Euclidean distance would be more appropriate. The normality testing on these data shows Euclidean distance produces clusters that are closer to being normally distributed than the cosine distance, indicating that the measure of normality is consistent with the accuracy of the clustering results obtained (Figure 4.31).

Since it is known that these data should only contain two regions, they could act as

a means to evaluate the success of evaluation measures that are claimed to be capable of estimating the number of clusters within the data. As with the ink samples however, the main limitation of these data is their imperfect representation of biologically derived MSI data.

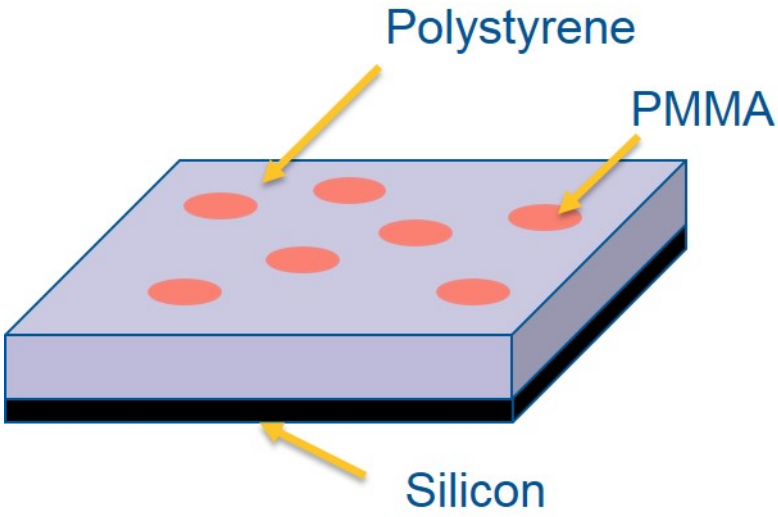

Figure 4.28: Diagram of the mixed polystyrene PMMA sample showing the circular regions of PMMA (red) within the polystyrene (purple).

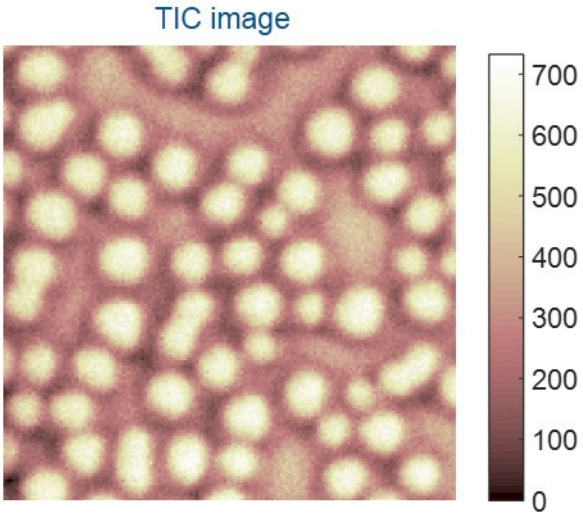

Figure 4.29: TIC from the SIMS image of the mixed polystyrene and PMMA sample. This clearly shows the locations of the PMMA highlighted by regions of higher intensity.

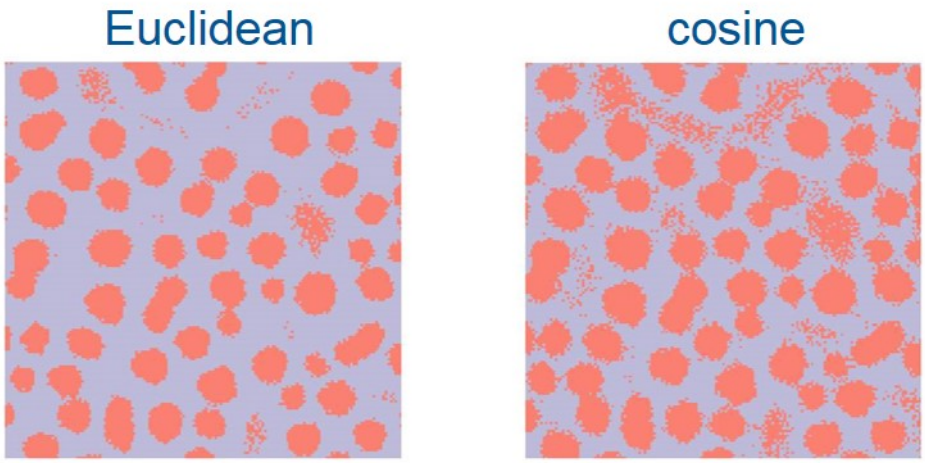

Figure 4.30: k-means clustering results on the mixed ploystyrene PMMA sample, using $k = 2$ and Euclidean and cosine distances. This shows more accurate results using the Euclidean distance as compared to the expected pattern from the polymer mix.

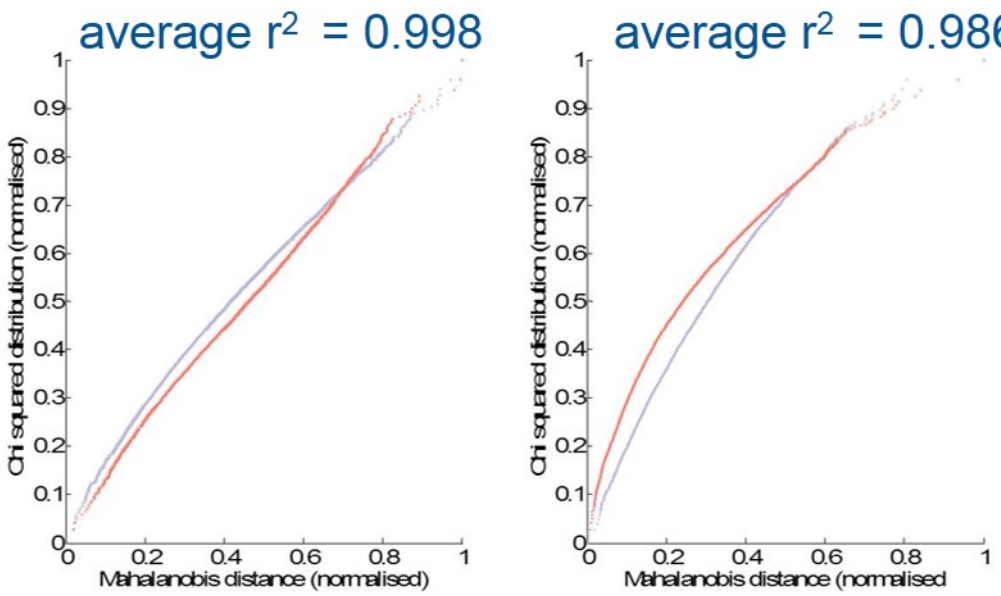

Figure 4.31: Results of normality testing on the individual clusters from the results in 4.30. Unlike the MALDI data, these SIMS data clustered using the Euclidean distance (left) are closer to normal than the data clustered by the cosine distance (right). This can be attributed to the more linear relationship between ion yield and concentration in SIMS compared to MALDI and DESI.

Normality on biological MSI data

When applied to MSI images of a biological systems (coronal mouse brain, saggital rat brain, and mouse lung), the chi squared quantile plots show that the data within clusters obtained using the cosine distance have a higher average r^2 values than the data within the clusters using the Euclidean distance (Figures 4.32 to 4.34). This means that the data in the clusters formed using the cosine distance are closer to normally distributed than the Euclidean distance, indicating that the cosine distance is the more appropriate distance metric for clustering with on these datasets based on the multivariate normal assumption of the k-means algorithm. The inappropriateness of k-means clustering with the Euclidean distance in the case of the coronal mouse brain data mirrors the visually poor results obtained with respect to the anatomical features expected from coronal mouse brain as seen in the Allen brain atlas [224]. In comparison the cosine distance gives visually clearer results, and the distribution of points within clusters are more normally distributed in the appropriate space. Of particular interest, the use of the common TIC normalisation *decreases* the normality of the data, and does not produce visually clearer segmentation images (Figure 4.32 and 4.33). Results obtained using the other tissue samples also show closer to normally distributed data when cosine distance is used to cluster the data, and visually improves segmentation. Without prior knowledge of the sample however it is difficult to determine if the segmentation result is improved, recent work by Palmer *et al.* have shown that experts in the field can, with a high degree of accuracy, evaluate MSI data quality [225]. It is worth noting that the values produced from the r^2 fitting cannot easily be directly interpreted, as they will be dependent on the number of data points, and the dimensionality of the data. Therefore it is recommended as a means to compare results, and caution should be taken when inferring additional information from them.

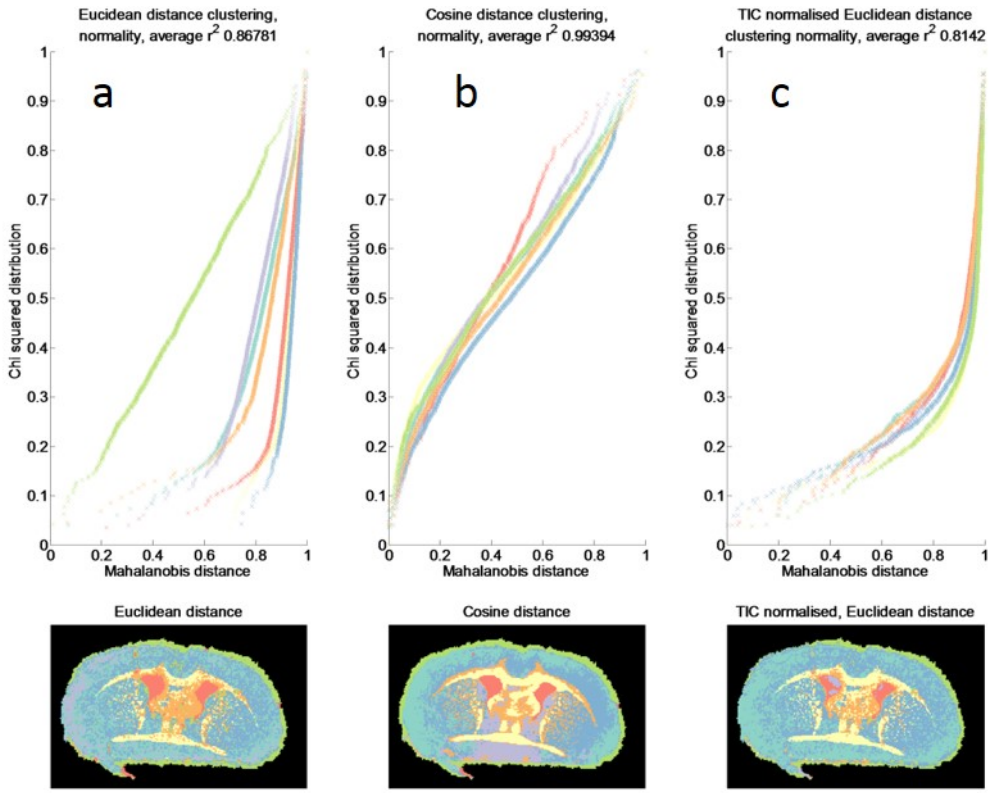

Figure 4.32: Quantile-Quantile plot in a) Euclidean space b) angular space, and c) TIC normalised Euclidean space for the data within each of the 7 clusters of the coronal rat brain image segmented using a) Euclidean distance, b) cosine distance, and c) Euclidean distance with TIC normalisation.

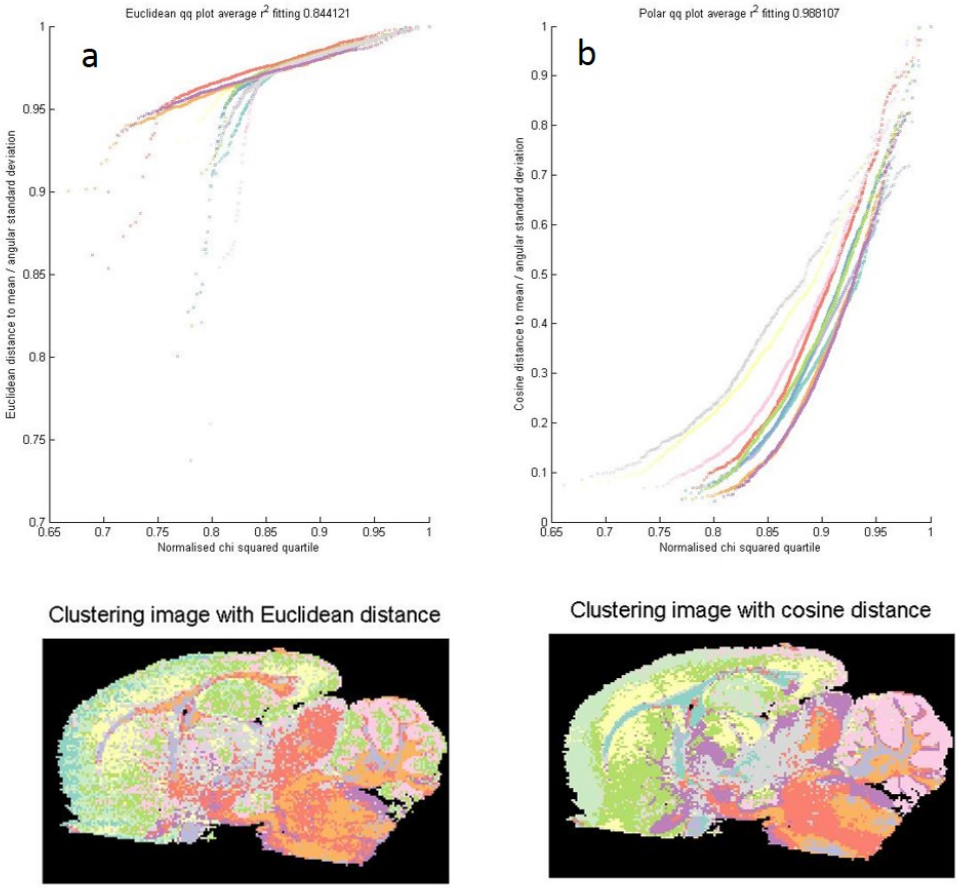

Figure 4.33: Quantile-Quantile plot in a) Euclidean space, and b) angular space for the data within each of the 10 clusters of the sagittal rat brain image segmented using a) Euclidean distance, and b) cosine distance.

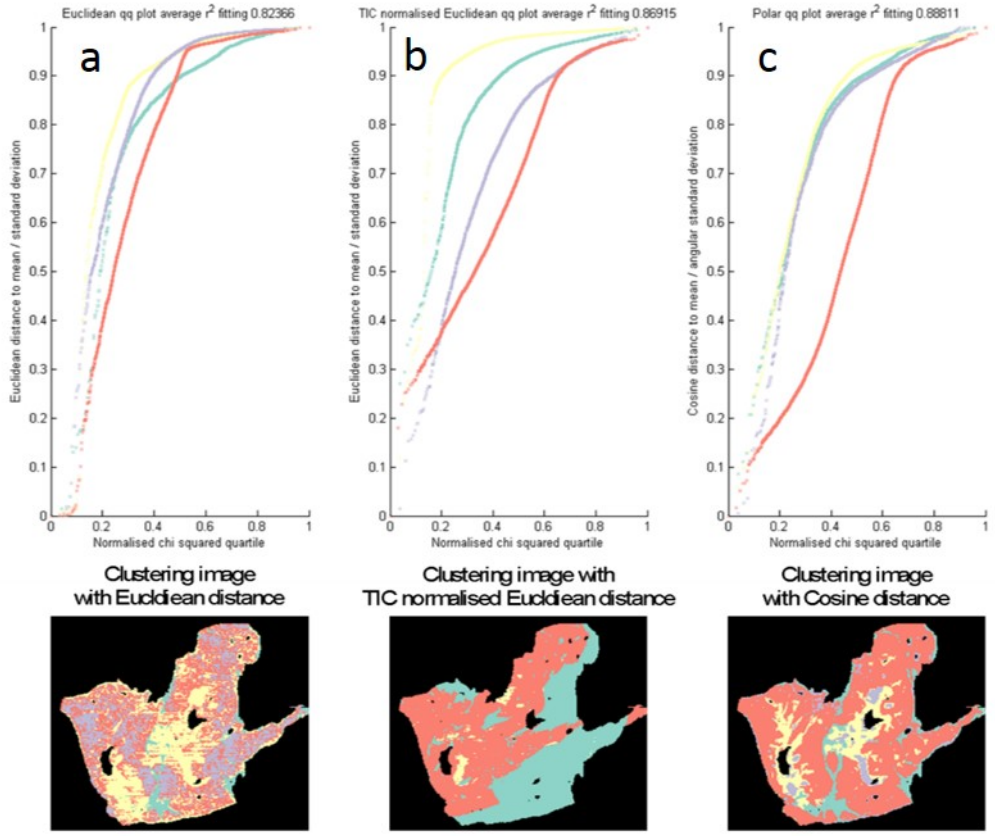

Figure 4.34: Quantile-Quantile plot in a) Euclidean space, b) TIC normalised Euclidean space, and c) angular space for the data within each of the 10 clusters of the mouse lung image segmented using a) Euclidean distance, b) Euclidean distance with TIC normalisation, and c) cosine distance.

Understanding the shape of the quantile-quantile plots

It is also worth noting that the shape of the q-q plots are not completely linear, a feature that can arise form a number of different sources. For example, the presence of a few outliers will skew the distribution towards a sigmoidal shape as is observed when the cosine distance is used (Figures 4.35 and 4.36). This is caused by the outliers skewing the mean of the data, and thus altering the Mahalanobis distance for every point. While this effect is minimised through the variance scaling process, some effect can still be observed. Alternately, a circular distribution of data with a core of normal data within produces

a similar shaped, apparent bilinear plot to those observed when using the Euclidean distance (Figure 4.38). This is not indicative of two normally distributed sets of data however which produces a different shaped plot (Figure 4.39). Further examples of how other distributions of data will affect these plots can be seen in figures 4.40 to 4.43. However, caution is required when generalising from these plots from two dimensional data into the higher dimensional space in which MSI data sits.

As well as providing a measure of the fit of normality to the data, the chi squared quantile-quantile plots can also provide a greater insight into the nature of the deviations from normality based on the shape of the curves. To investigate this, a number of 2D datasets with different properties were generated, and their resulting quantile-quantile plots were created (Figures 4.35 to 4.43). The simulation of a normal distribution with presence of outlier data results in a sigmoidal shaped plot similar to those observed on the clustering of the coronal mouse brain dataset using the cosine distance (Figure 4.32 b), with the gradient of the inflection point indicating the distance of these outliers from the normal distribution of data (Figures 4.35 and 4.36). This is caused by the outliers skewing the mean of the data thereby altering the Mahalanobis distance at each point. A circular distribution of data with a core of normally distributed data (Figure 4.38) produces a bilinear shaped plot similar to that observed in figure 4.32 a and c. A number of different shaped plots can also be observed with differnt distributions of data, and matching the observed quantile plot from MSI data to simulations can give a better understanding into the distribution of these data. While the shape of these plots can give further insights into the distribution of the data, care should be taken when extrapolating from the simulated 2D data into higher dimensions as interpretation of higher dimensional data is challenging at best.

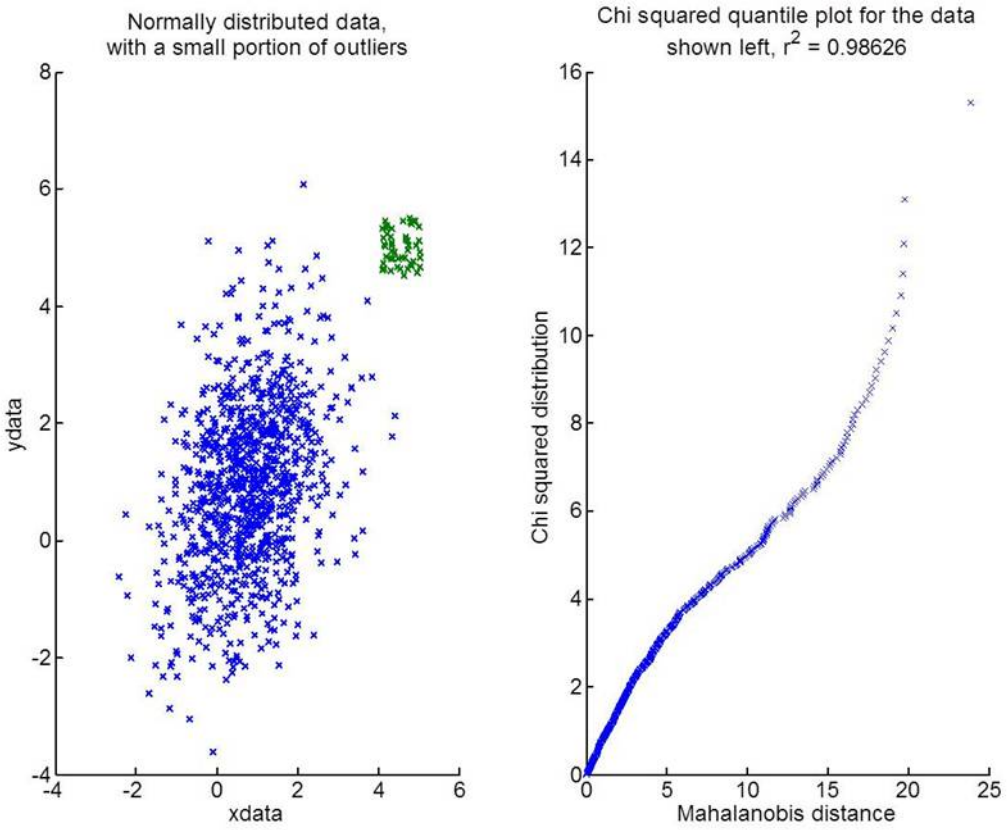

Figure 4.35: Chi squared quantile plot for a normal distribution of data (1000 data points) containing a small portion of outlier data (100 data points).

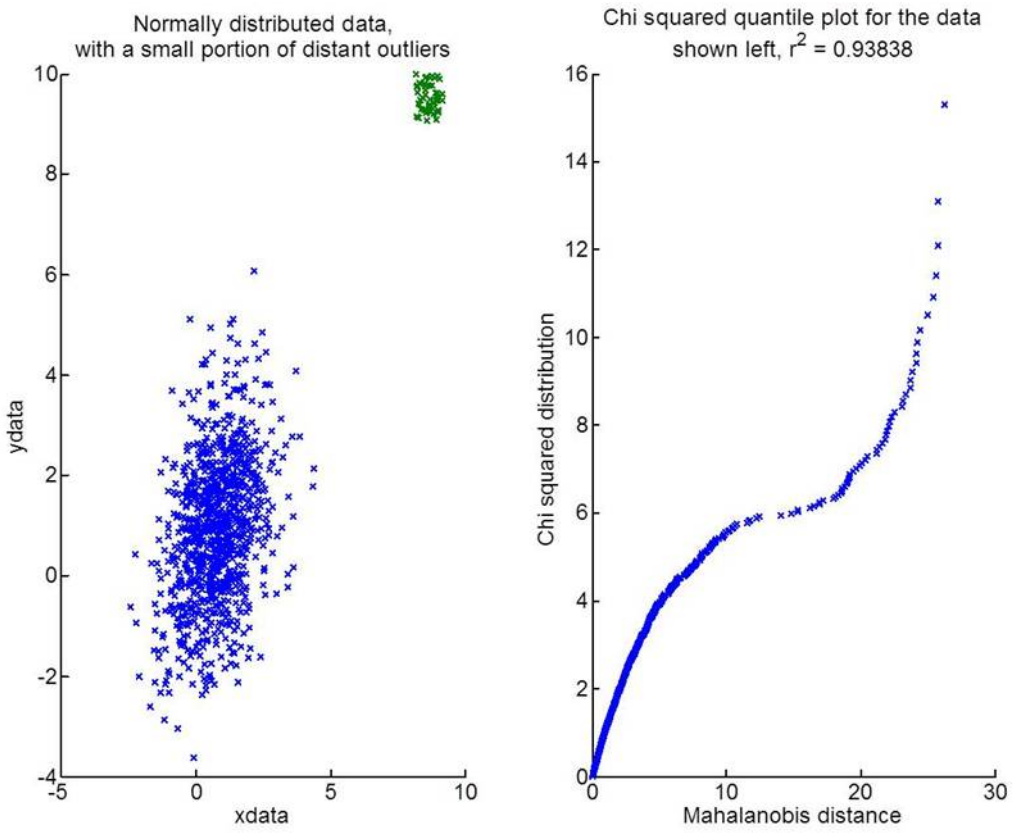

Figure 4.36: Chi squared quantile plot for a normal distribution of data (1000 data points) containing a small portion of distant outlier data (100 data points).

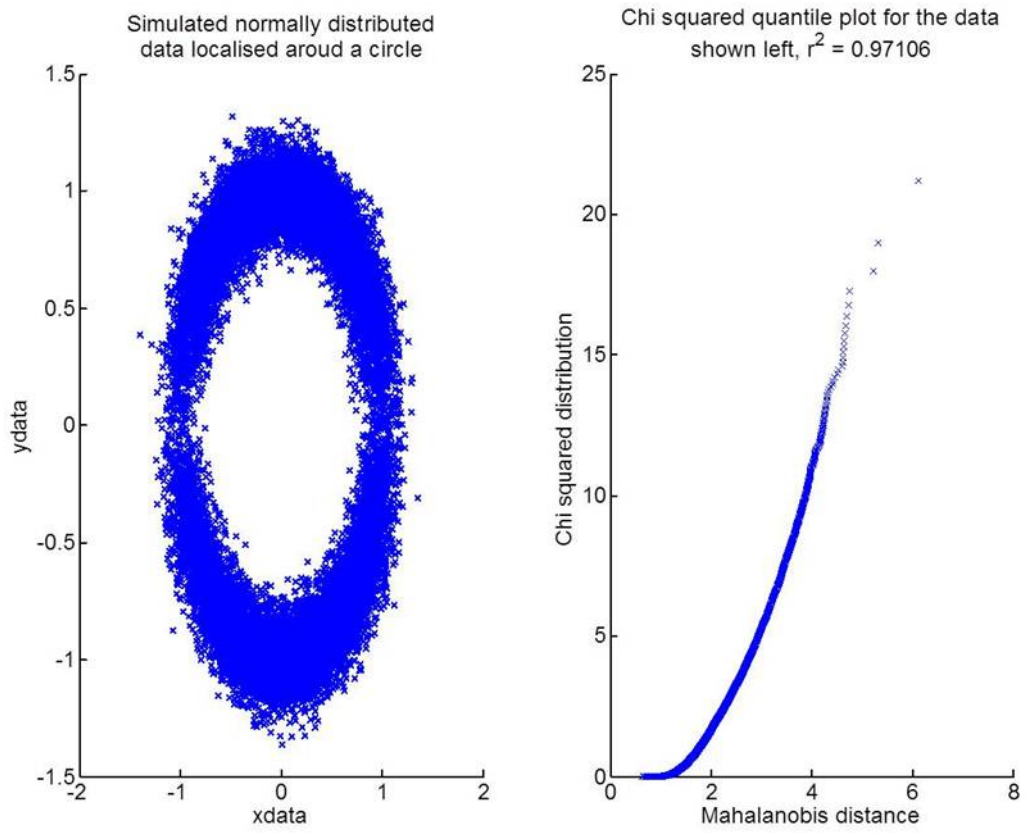

Figure 4.37: Chi squared quantile plot for a circular distribution of data (8000 data points).

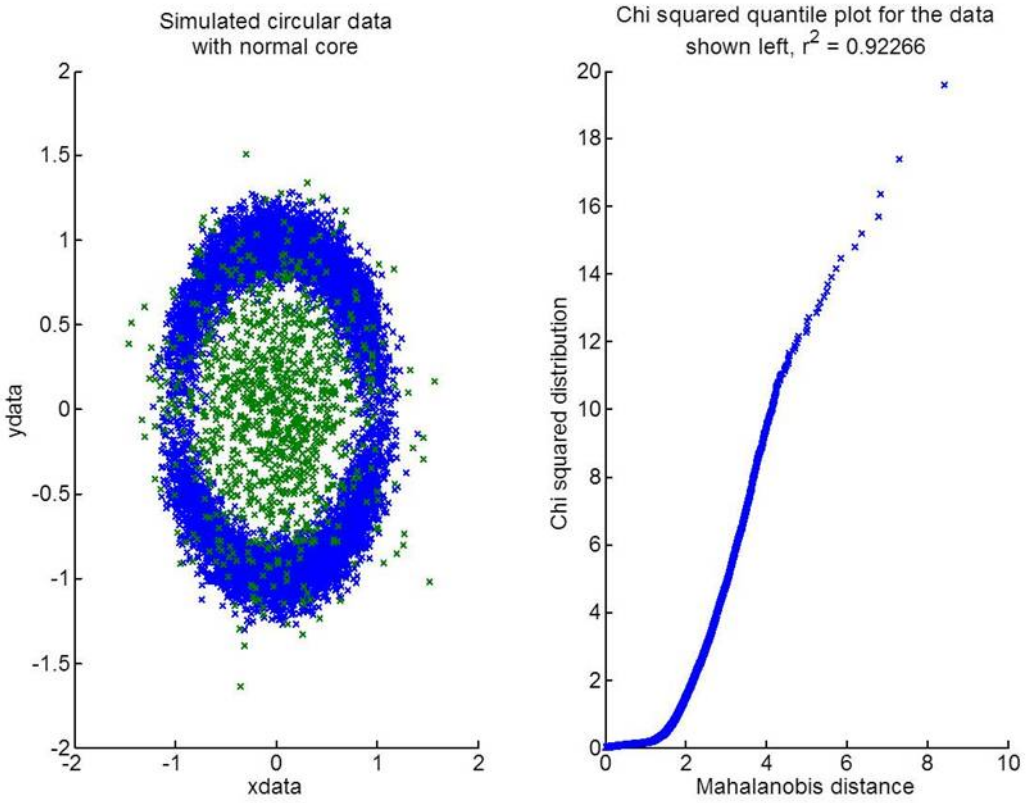

Figure 4.38: Chi squared quantile plot for a circular distribution of data (8000 data points) with a small core of normally distributed data (1000 data points).

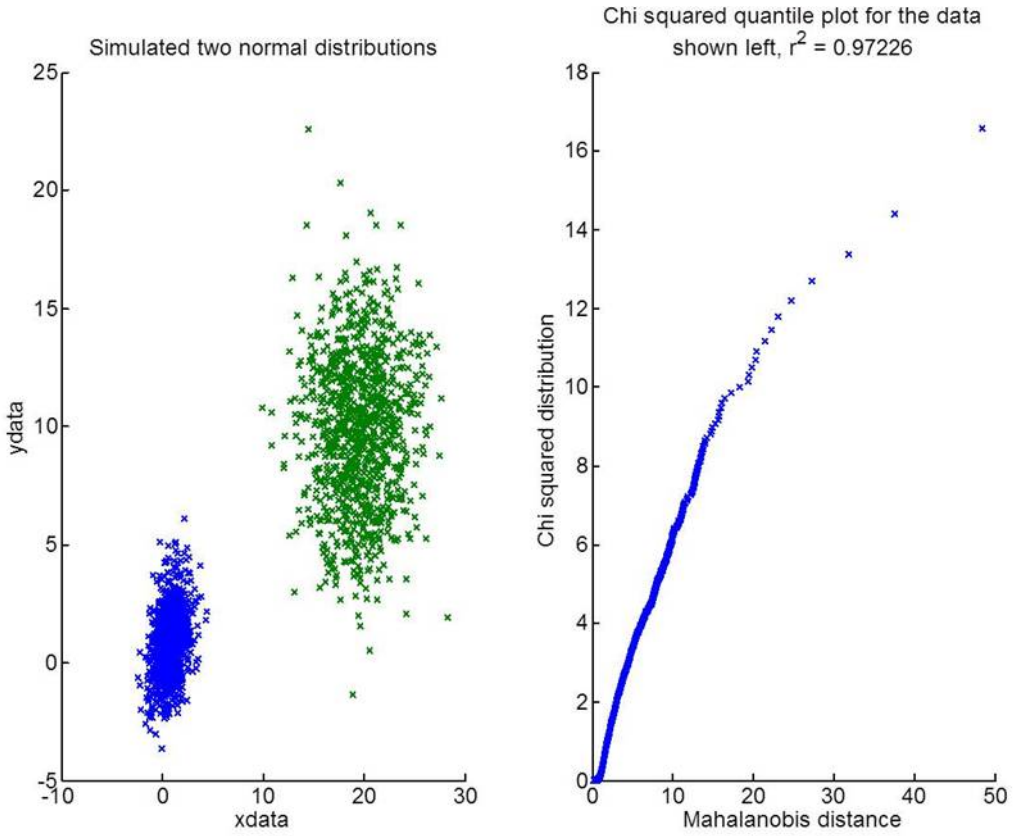

Figure 4.39: Chi squared quantile plot for two sets of normally distributed data (1000 data points each).

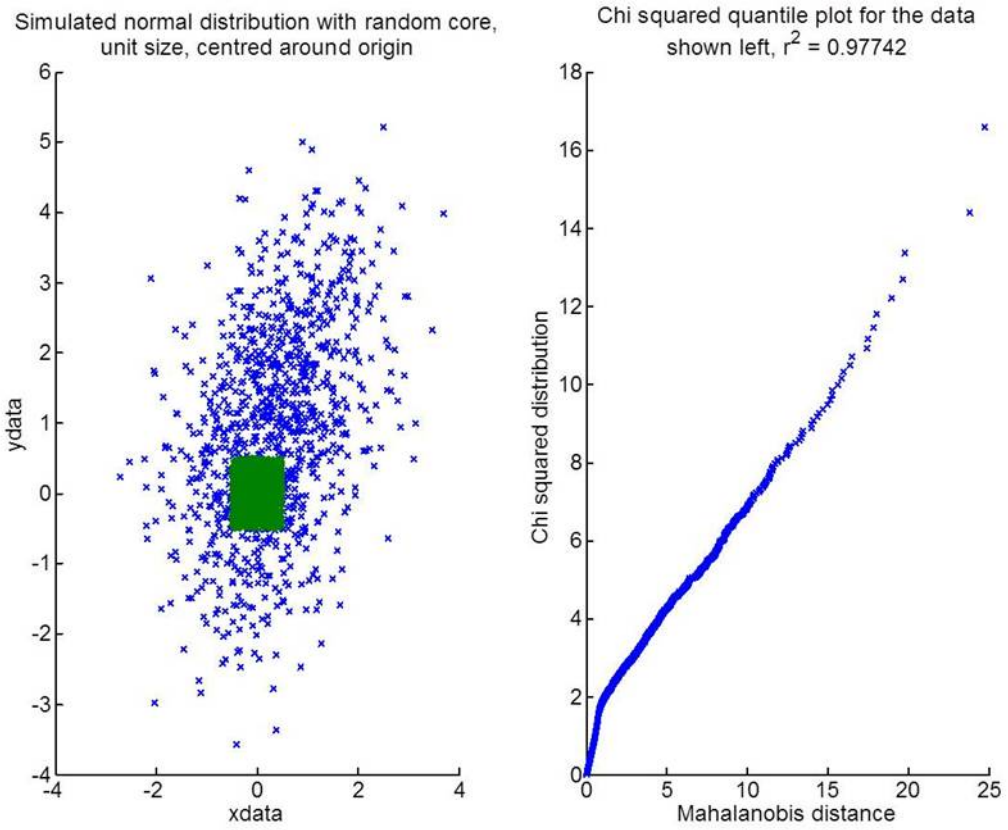

Figure 4.40: Chi squared quantile plot for a normal distribution of data (1000 data points), with additional randomly distributed data (1000 data points) centred at the origin with unit width.

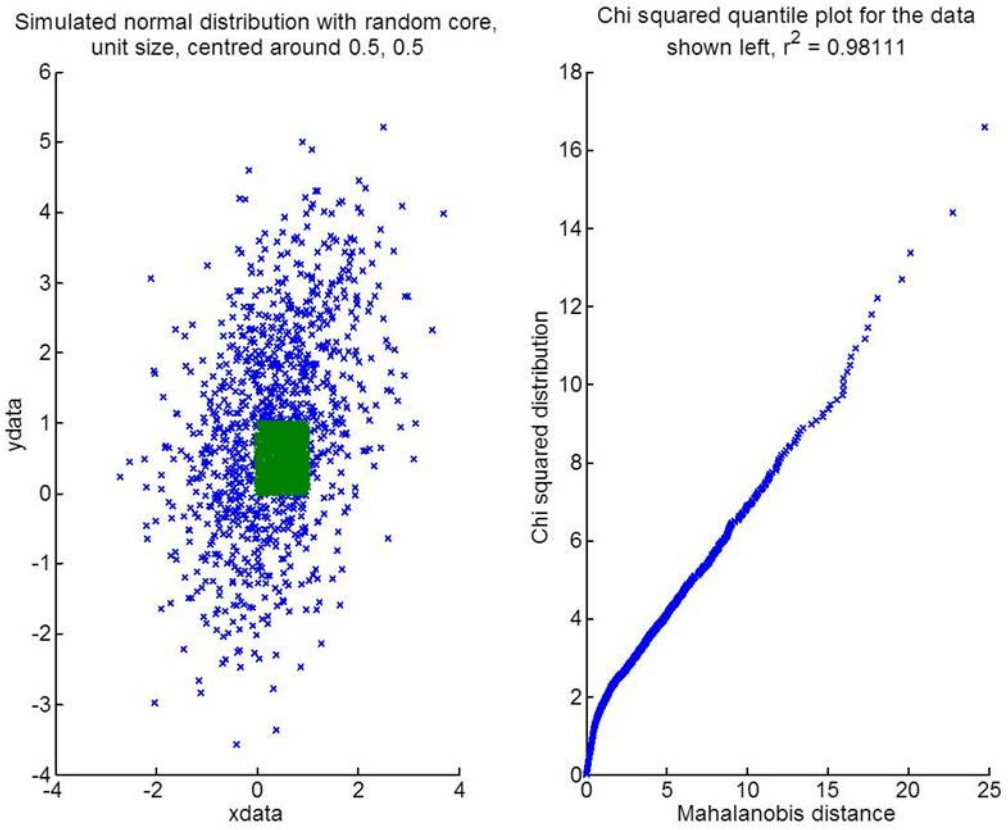

Figure 4.41: Chi squared quantile plot for a normal distribution of data (1000 data points), with additional randomly distributed data (1000 data points) centred at [0.5, 0.5] with unit width.

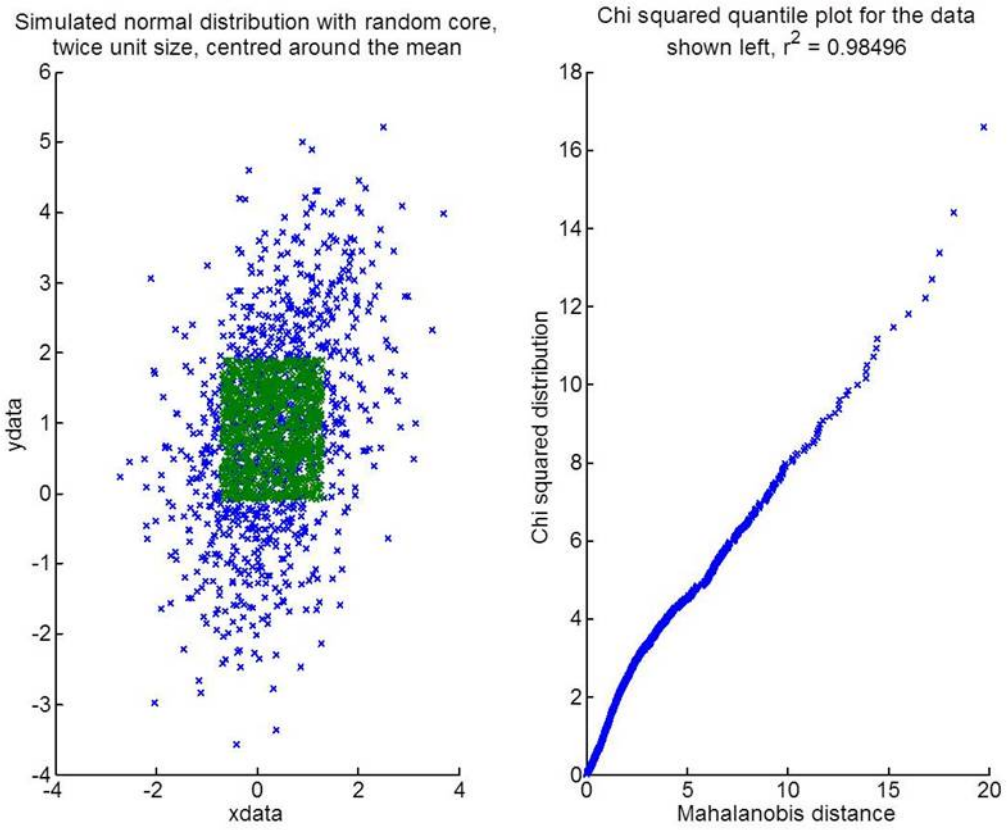

Figure 4.42: Chi squared quantile plot for a normal distribution of data (1000 data points), with additional randomly distributed data (1000 data points) centred on the mean of the normal data with double unit width.

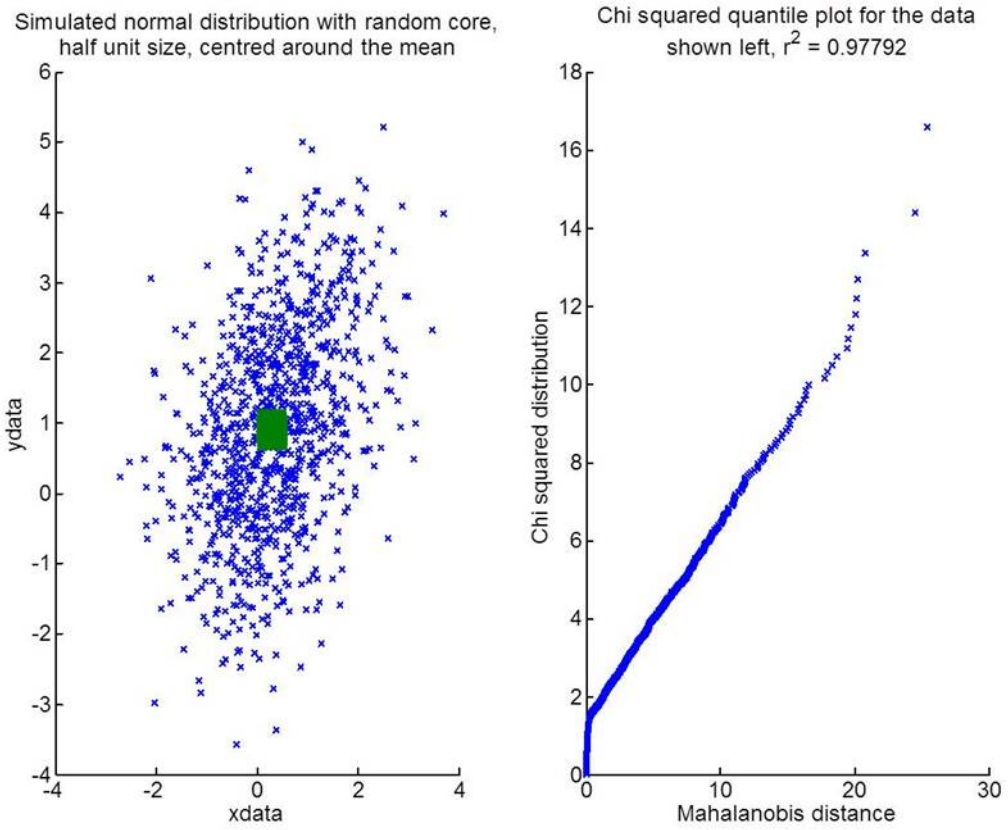

Figure 4.43: Chi squared quantile plot for a normal distribution of data (1000 data points), with additional randomly distributed data (1000 data points) centred on the mean of the normal data with half unit width.

4.4 Conclusions

Accurate and reliable means to evaluate clustering algorithms and parameters are critical to it's development in MSI. Where possible, external evaluation should be performed, and the use of the Jaccard index as an external evaluation metric is suggested over the Rand index due to the high sample numbers in MSI data. Inkjet printed standards can be used as an initial means to provide samples with known spatial distribution to initially test clustering algorithms in MSI however caution must be taken when generalising the results from these ink data to biological MSI datasets. Where external evaluation is not possible,

a robust internal metric for clustering evaluation is required. Currently used metrics suffer from issues of high dimensionality, and inappropriate distance metric selection, resulting in inconsistent evaluation of MSI data. We show the use of multivariate normality testing in MSI to more reliably evaluate clustering results in MSI, and to appropriately select the distance metric to use based on the assumptions of the algorithm involved. The challenge still remains to create model reference data with known spatial distributions that also reflects the complexity of biological MSI data.

CHAPTER 5

SYNTHETIC DATA AND SIMULATION IN MSI

5.1 Introduction

While distance metric determination is a crucial factor in any clustering algorithm, there are still many other parameters which must also be selected, such as the number of clusters, or the method for centroid initiation. In addition to this, there are many other clustering approaches such as density based clustering which do not assume multivariate normality in the data [128, 145]. Therefore, a method to generate datasets with a ground truth is required to assess the suitability of these approaches and to permit a comparison of different clustering approaches. It is also widely recognised that this is vital for a wide number of other areas. Data simulated from first principles is one approach that is used to achieve this in other fields [226, 227]. However, while some aspects of image formation and noise in MSI are well understood, there are still a large number of unknowns in aspects such as sample preparation and ionisation [12, 66]. One approach is to take existing peak lists and to then simulate instrumental variables and apply these to this peak list [177, 228]. A robust method is needed however to generate peak lists that are well controlled, but still representative of the biological rather than instrumental variance expected. A new biological sample could be analysed each time a new set of spectra are required, but using new animal or human tissue each time a different number of regions or pixels is required, is neither practical or ethical. In other areas such as financial

prediction, and geological analysis, statistical modelling is used to convert discrete data into a continuous function, thereby allowing resampling to generate the desired number of data points [229]. Statistical modelling assumes that data from a population are derived from a known probability distribution function. Provided that the model adequately describes the data, the underlying distribution can then be resampled to give a new synthetic dataset with any desired number of data points. This new synthetic dataset will have the same distribution as the original reference dataset that the model was derived from. For large and high dimensional data, model generation and parameter estimation can be challenging, however, the multivariate normal model parameters can be easily estimated even for very large data [230]. As previously demonstrated in chapter 4, the clustered MSI data from coronal mouse brain closely approximates to a multivariate normal distribution when the data is converted to polar coordinates. This means that the multivariate normal distribution can be used as the basis for statistical modelling for MSI data. For new MSI data, normality testing can be performed to determine the appropriateness of this model.

5.2 Experimental

5.2.1 sample preparation

Coronal mouse brain was sectioned to 12 μm thickness and thaw mounted onto glass slides (Superfrost, Thermo Fisher Scientific, Waltham, MA USA), before being coated with CHCA matrix (5 mg/mL, 80% MeOH 0.1% TFA) using an automated pneumatic sprayer (TM-sprayer, HTX imaging, Chapel Hill, NC, USA). Full description of rat brain image acquisition is detailed elsewhere [10]. Briefly, a single formalin fixed rat brain section was coated with -cyano-4-hydroxycinnamic acid using an automated matrix deposition system (TM sprayer from HTX Technologies, NC, U.S.A.).

5.2.2 Image acquisition

MALDI images were acquired using a Synapt G2Si (Waters, Manchester, UK), and QSTAR XL QqTOF (AB Sciex, Warrington, UK) using a pixel size of 45 μm in both x and y. Synapt images were acquired in positive resolution mode with an m/z range of 100-1200 Da. QSTAR images were acquired in raster imaging mode on medium speed ($1mms^{-1}$), using an Nd:YAG laser (355 nm, Elforlight Ltd., Daventry, UK) operating at a repetition rate of 1 kHz. Fixed rat brain data were acquired on a QSTAR Elite QqTOF (AB Sciex, Warrington, UK) with a pixel size of 100 μm^2, and a mass range of 50-1000 Da.

Data for the TOF simulation comparison will be described in full in chapter 6. Briefly, data from mouse colon was collected on a RapiFlex MALDI TOF/TOF (Bruker Daltonics, Germany) in reflectron positive ion mode, using a pixel size of 5 μm in both x and y, over an m/z range of 200-1000 Da.

5.2.3 Synthetic data generation

Synthetic data were generated using the workflow shown in figure 5.1. For a given reference dataset, normality testing in Euclidean and polar space is performed to assess the suitability of this as a model. For polar normal data, the radii and angles from this conversion are then modelled as normally distributed. Normality testing is performed using the chi squared quantile plots described in chapter 4, and provided the data are sufficiently close to normally distributed, the mean spectrum and covariance matrix is calculated and stored. For polar co-ordinate data, a mean and variance for the radii are also stored. Synthetic data are then sampled probabilistically from a multivariate normal distribution using the MATLAB function *mvnrnd*, and synthetic hypotenuses using the MATLAB function *randn*. For polar converted data, following the generation of the synthetic data, these are then converted back to cartesian co-ordinate space using the synthetic multivariate and hypotenuse data.

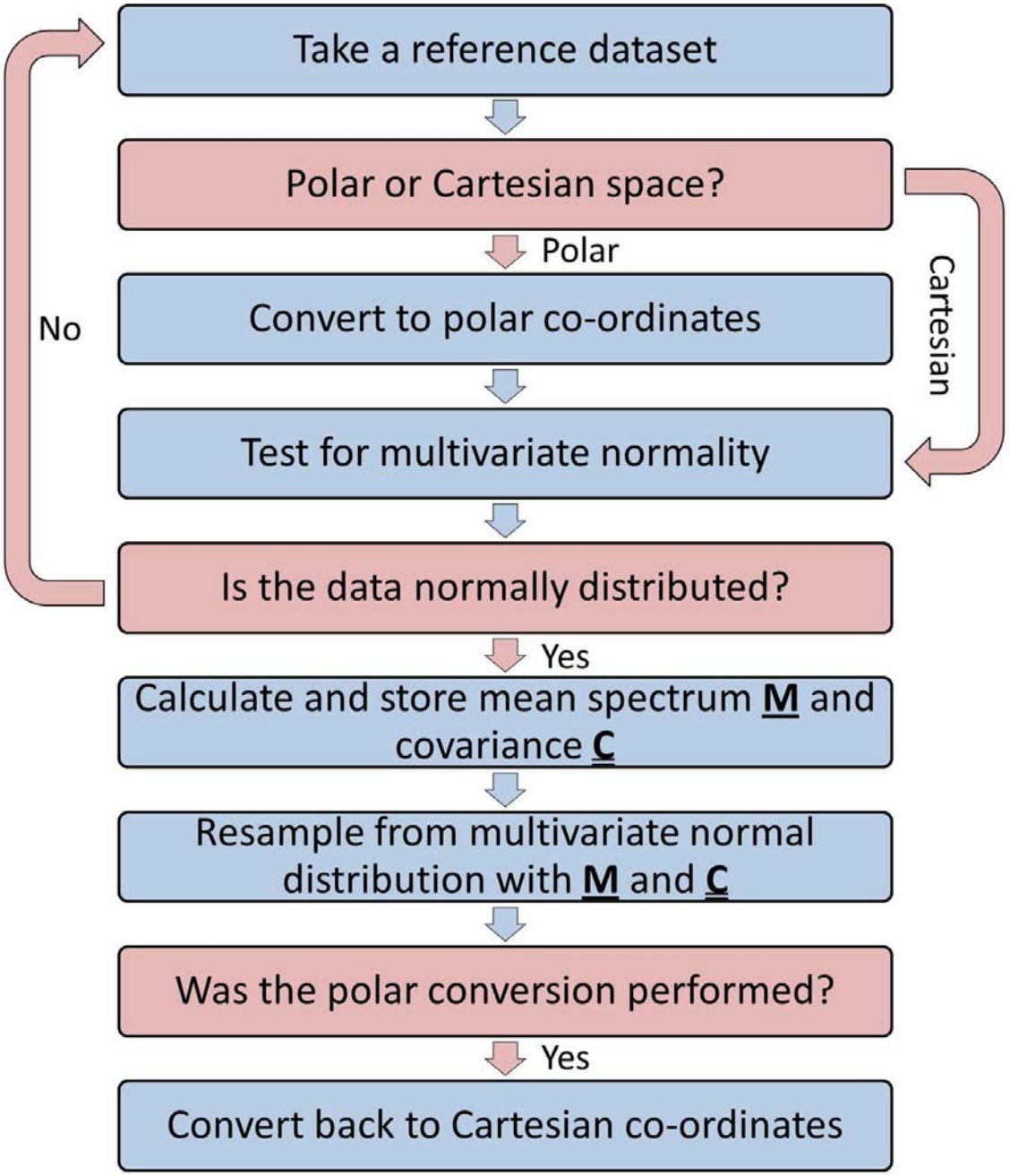

Figure 5.1: Workflow for the generation of synthetic datasets from a set of reference data.

Random projection was performed according to algorithm 1 of [133], with $k = 100$. Instead of recreating the original dataset using from the low dimensonal data matrix $\underline{\underline{\mathbf{A}}}$ and orthonormalised projected data $\underline{\underline{\mathbf{Q}}}$ by performing $\underline{\underline{\mathbf{X}}} = \underline{\underline{\mathbf{Q}}}\,\underline{\underline{\mathbf{A}}}$ as described by Palmer *et al.*, normality testing is performed on the reduced data matrix $\underline{\underline{\mathbf{A}}}$, and synthetic reduced datasets $\underline{\underline{\mathbf{D}}}$ sampled from a probability distribution with a mean and covariance of $\underline{\underline{\mathbf{A}}}$.

These synthetic datasets $\underline{\underline{\mathbf{N}}}$ are then converted back into full spectra using $\underline{\underline{\mathbf{N}}} = \underline{\underline{\mathbf{Q}}}\ \underline{\underline{\mathbf{D}}}$ (Figure 5.2).

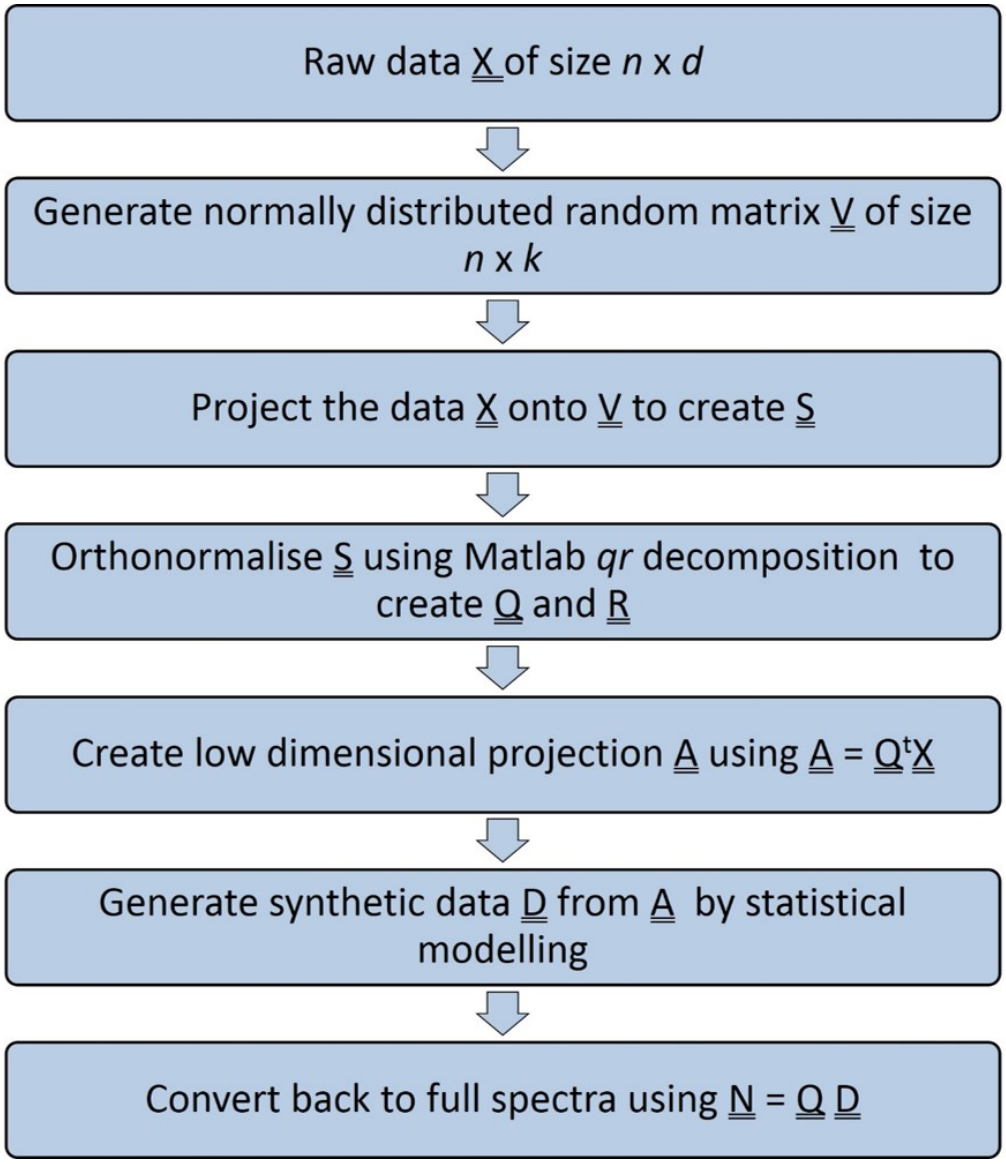

Figure 5.2: Workflow for the use of random projection to be included into the synthetic data generation in figure 5.1.

5.3 Results and Discussion

5.3.1 Statistical modelling for synthetic MSI data

In order to perform statistical modelling of MSI data, a series of seven anatomical features from an MSI image of the previously shown mouse brain were used as a reference dataset (Figure 5.3). These regions were generated based on the analysis of selected ion images and PCA scores (Figures 5.4, and 5.5), in comparison to a high resolution optical image (Figure 5.6), the Allen brain atlas [224], and a serial section stained by hematoxylin and eosin (H&E) segmented and labelled by a pathologist (Figures 5.7, and 5.8). In order to test the suitability of the multivariate normal distribution as a basis for modelling, these data were then tested for normality in polar and Euclidean space using the chi-squared quantile plots from chapter 4, and showed a high degree of normality throughout polar space but not Euclidean (Table 5.1). This means that the multivariate normal model in polar space can be used to summarise the properties of these data. In order to evaluate the effectiveness of this approach, a new synthetic dataset was generated by resampling from the distribution drawing the same number of pixels from the distribution as were in the original reference data. The better the model, the more similar these new data should be to the original reference data. The deviations from normal in the plots appear to match the shapes observed in the case where there are a few outlier data points in the 2D simulations (Figure 4.36), however as discussed, it is difficult to generalise these 2D plots into more dimensions. The synthetic spectra from a number of the different anatomical regions were then visually compared to the original reference spectra (Figure 5.9). The synthetic and real spectra show a high degree of spectral similarity, and expected features such as isotope ratios and fragments are preserved (Figure 5.10), thus ensuring the realism of the synthetic data. Some differences in the spectra are observed, since the synthetic spectra are sampled from a distribution and will therefore contain the same underlying variance as the reference data. This is important since biological samples vary, and so in order to be realistic, the synthetic data must incorporate this variance. Better models

that are more accurate and cannot contain negative values could be developed to improve these synthetic data, however model testing and parameter selection are challenging in high dimensional data, whereas the normal distribution can be tested more easily, and requires only means and covariances [230].

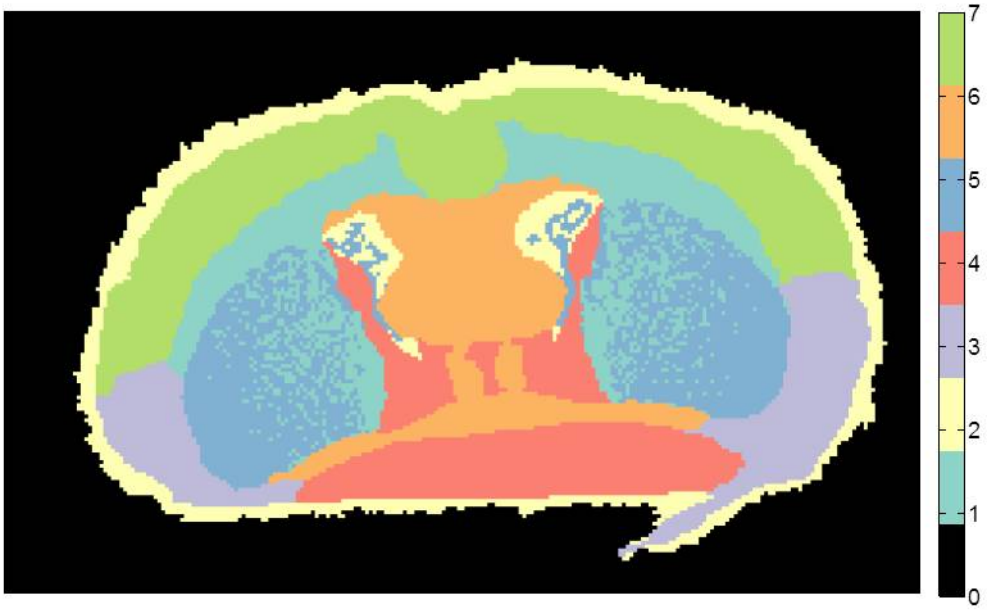

Figure 5.3: Image of the segmentation used to form the basis of the reference dataset for statistical modelling.

Figure 5.4: Selected ion images used to aid segmentation of the anatomical features of the mouse brain tissue.

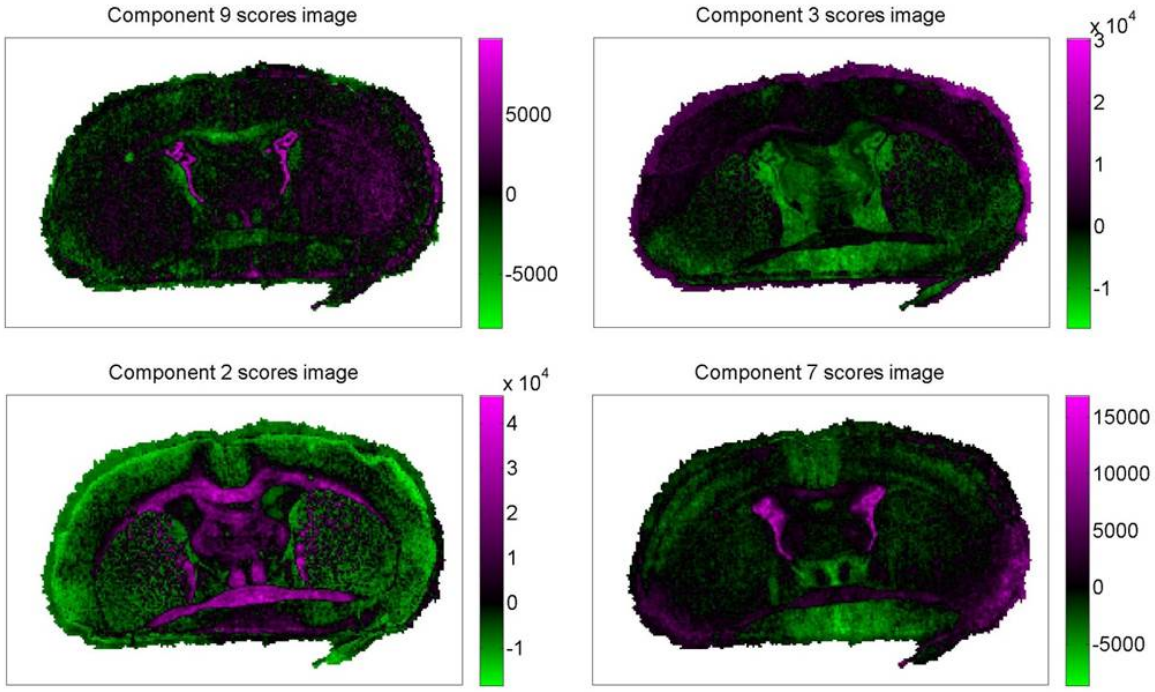

Figure 5.5: Selected PCA scores images used to aid segmentation of segment the anatomical features of the mouse brain tissue.

Figure 5.6: High resolution optical image of the tissue section acquired before MALDI MSI analysis.

Figure 5.7: Coronal brain section analysed by MALDI, haematoxylin and eosin stained and labelled by a pathologist.

Figure 5.8: Coronal brain section analysed by MALDI, haematoxylin and eosin stained, labelled, and segmented by a pathologist.

Antomical region	Segment	Euclidean normality	Polar normality
Corpus callosum	1	0.822	0.983
Outer boundary	2	0.904	0.983
Olfactory areas	3	0.889	0.993
Brain stem	4	0.854	0.988
Caudoputamen	5	0.948	0.994
Lateral septal complex	6	0.916	0.984
Isocortex	7	0.680	0.990

Table 5.1: Normality of the data in the seven anatomical regions of the brain in Eulidean and polar space.

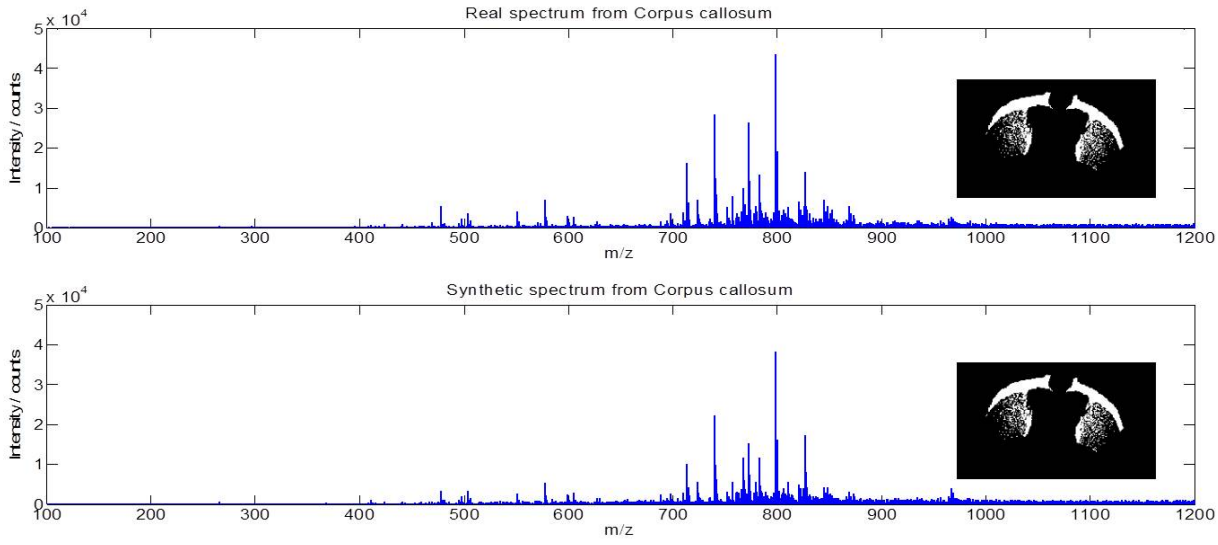

Figure 5.9: Comparison of an example real mass spectrum from the corpus callosum region (top), with a synthetic mass spectrum generated by statistical modelling of all the data from the corpus callosum.

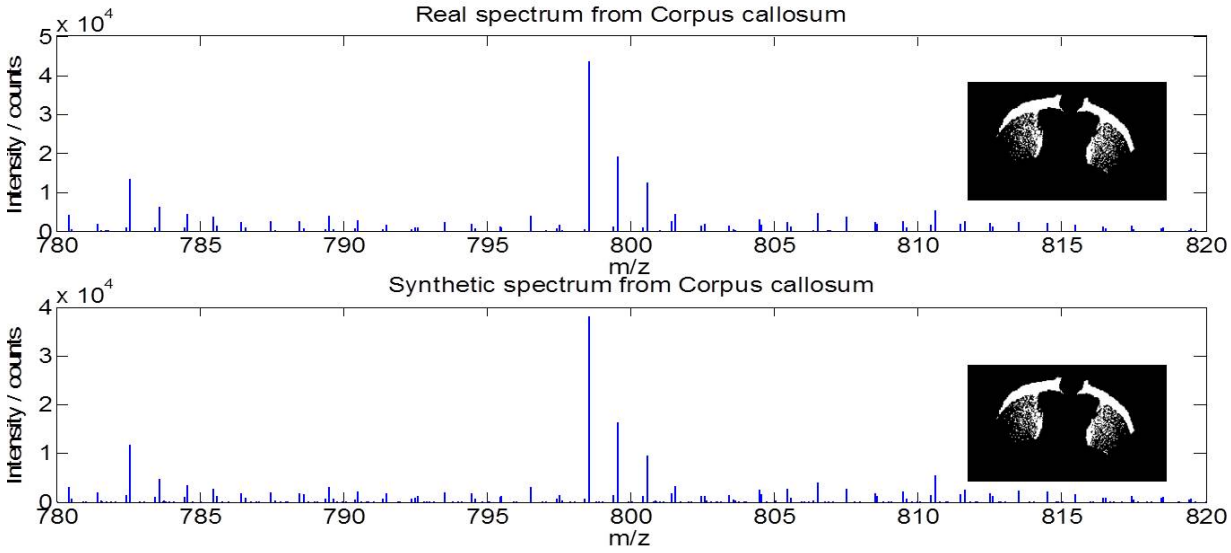

Figure 5.10: Seleccted m/z region from the spectrum in Figure 5.9 showing preservation of expected isotope patterns.

The underlying goal of these synthetic data are to be able to evaluate multivariate methods in MSI. Therefore, visual comparison of the spectra is insufficient to evaluate datasets as they do not capture the full extent of the data. In order to evaluate how closely the synthetic data matches real data, a new dataset, comprising of both synthetic and real spectra was generated. PCA was then performed on this combined dataset to determine if the statistical modelling process introduced any additional observable variance within the data. No principal component scores were found to separate the synthetic from real data (Figures 5.11 to 5.20). This means that even when all mass channels are considered, the difference between the synthetic and real data is smaller than that between different anatomies or the spectral noise within the data and supports the suggestion that the differences from normal are likely to be outlier pixels. As such the statistical modelling of appropriately segmented MSI data using a multivariate normal distribution can generate realistic spectra in order to create new datasets with known ground truth for external evaluation of clustering in mass spectrometry imaging.

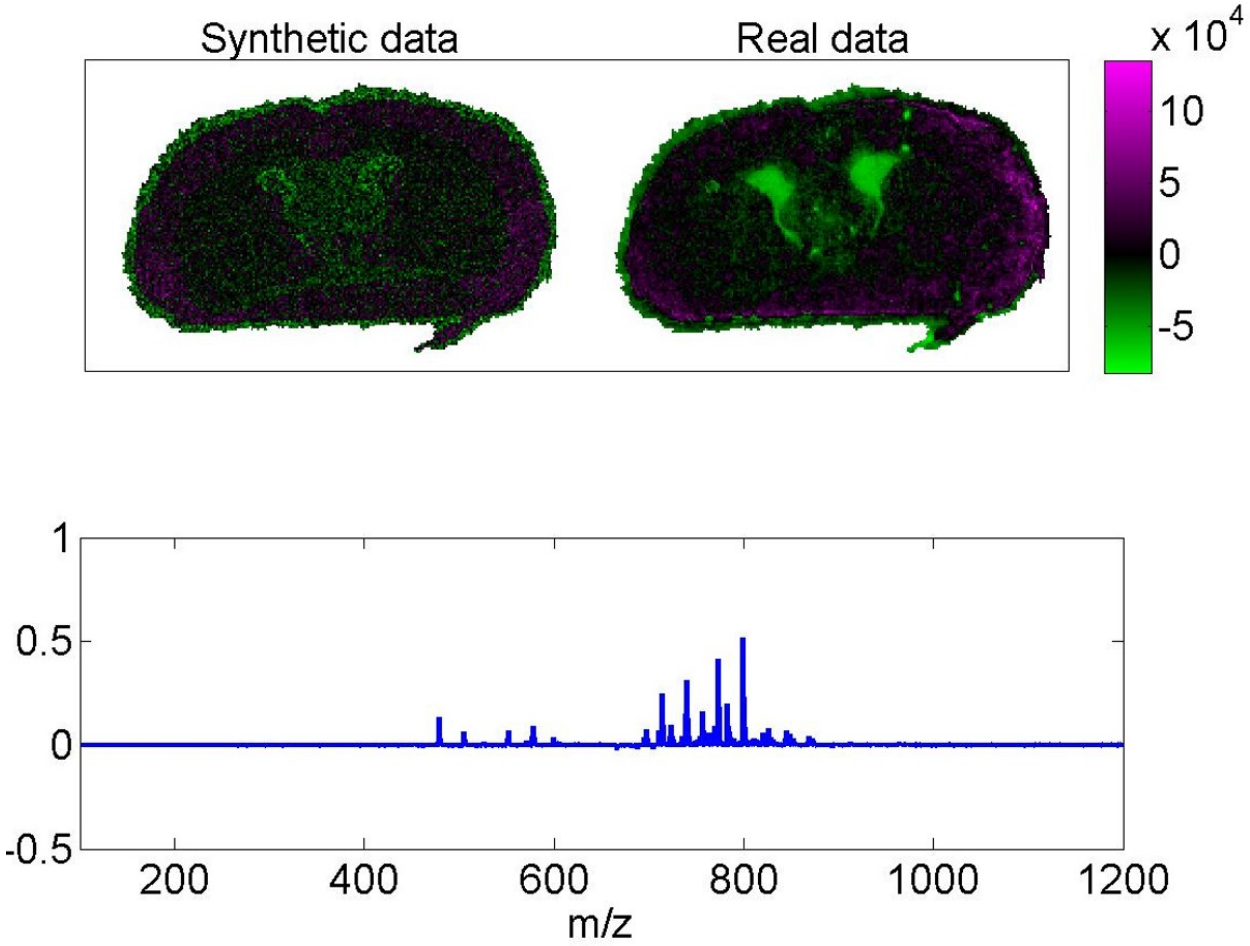

Figure 5.11: Principal component 1 from the combined real and synthetic dataset showing no distinct separation between the two.

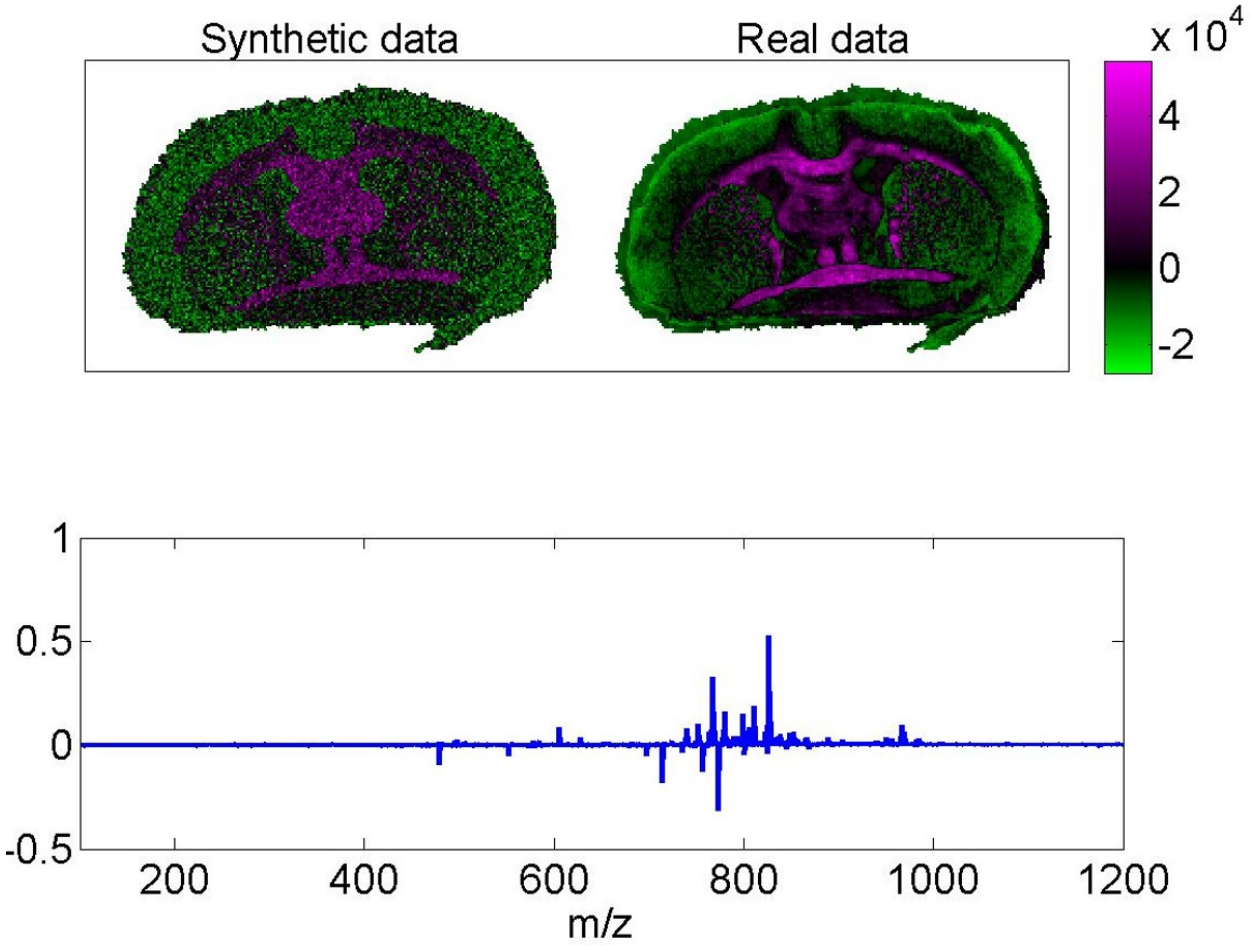

Figure 5.12: Principal component 2 from the combined real and synthetic dataset showing no distinct separation between the two.

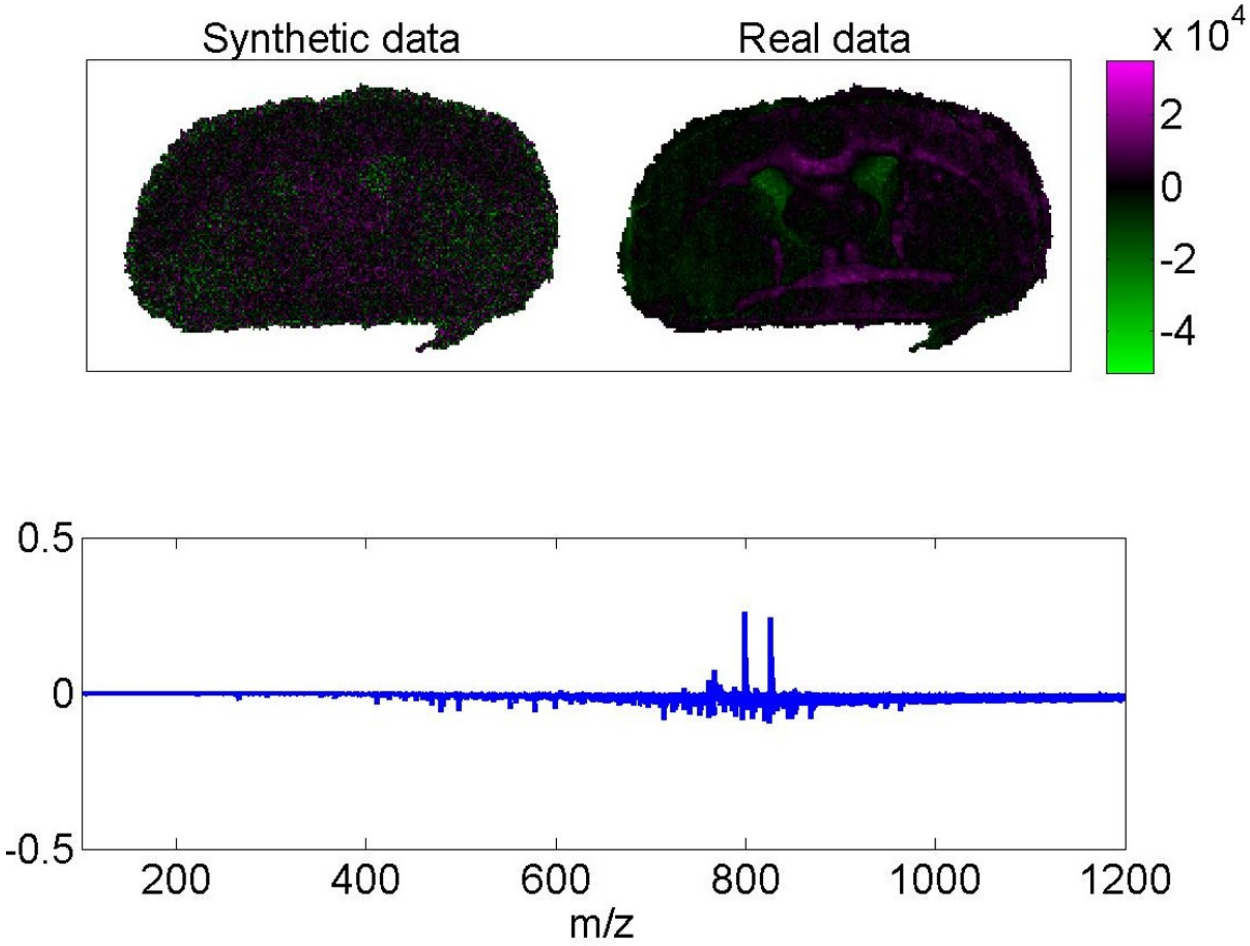

Figure 5.13: Principal component 3 from the combined real and synthetic dataset showing no distinct separation between the two.

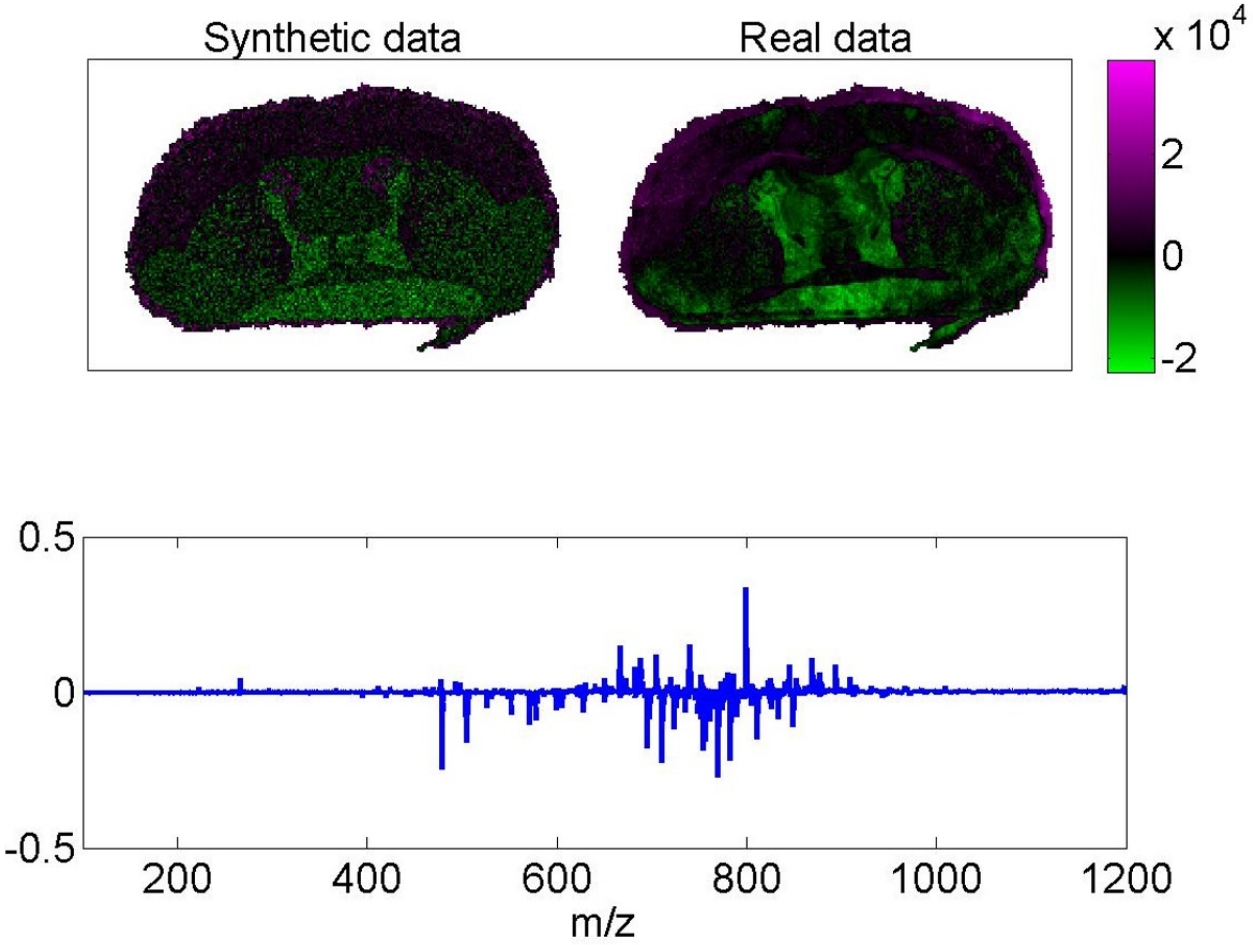

Figure 5.14: Principal component 4 from the combined real and synthetic dataset showing no distinct separation between the two.

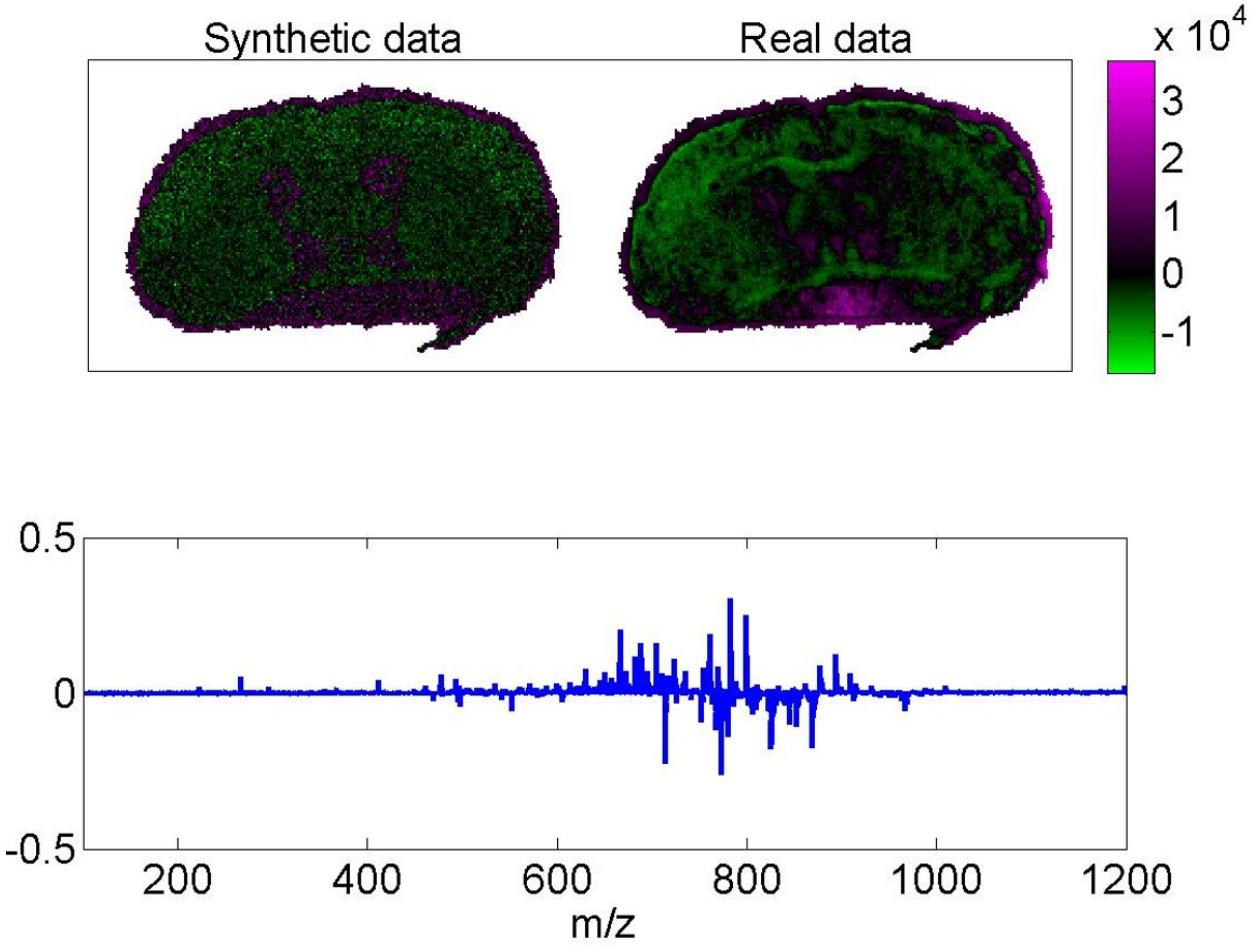

Figure 5.15: Principal component 5 from the combined real and synthetic dataset showing no distinct separation between the two.

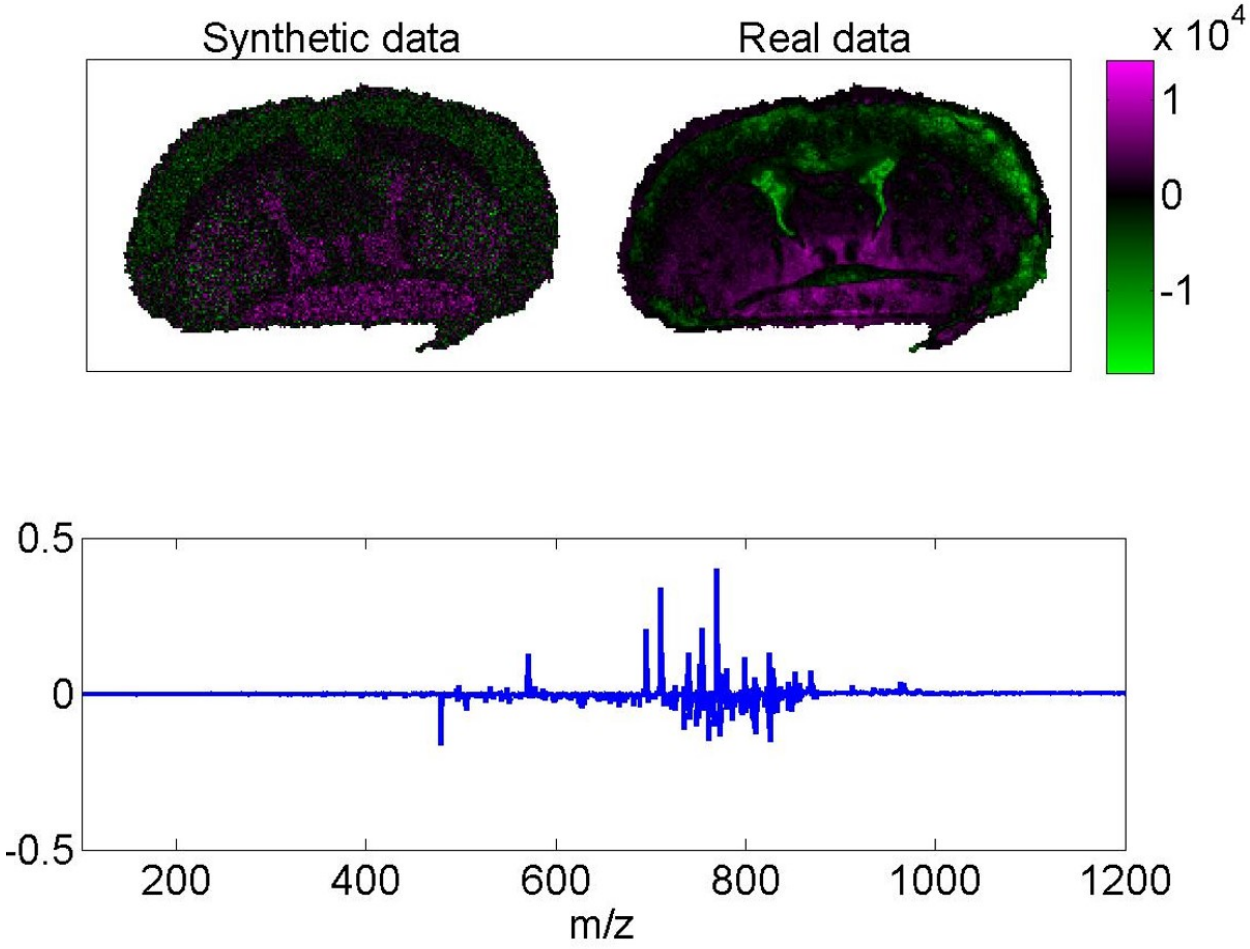

Figure 5.16: Principal component 6 from the combined real and synthetic dataset showing no distinct separation between the two.

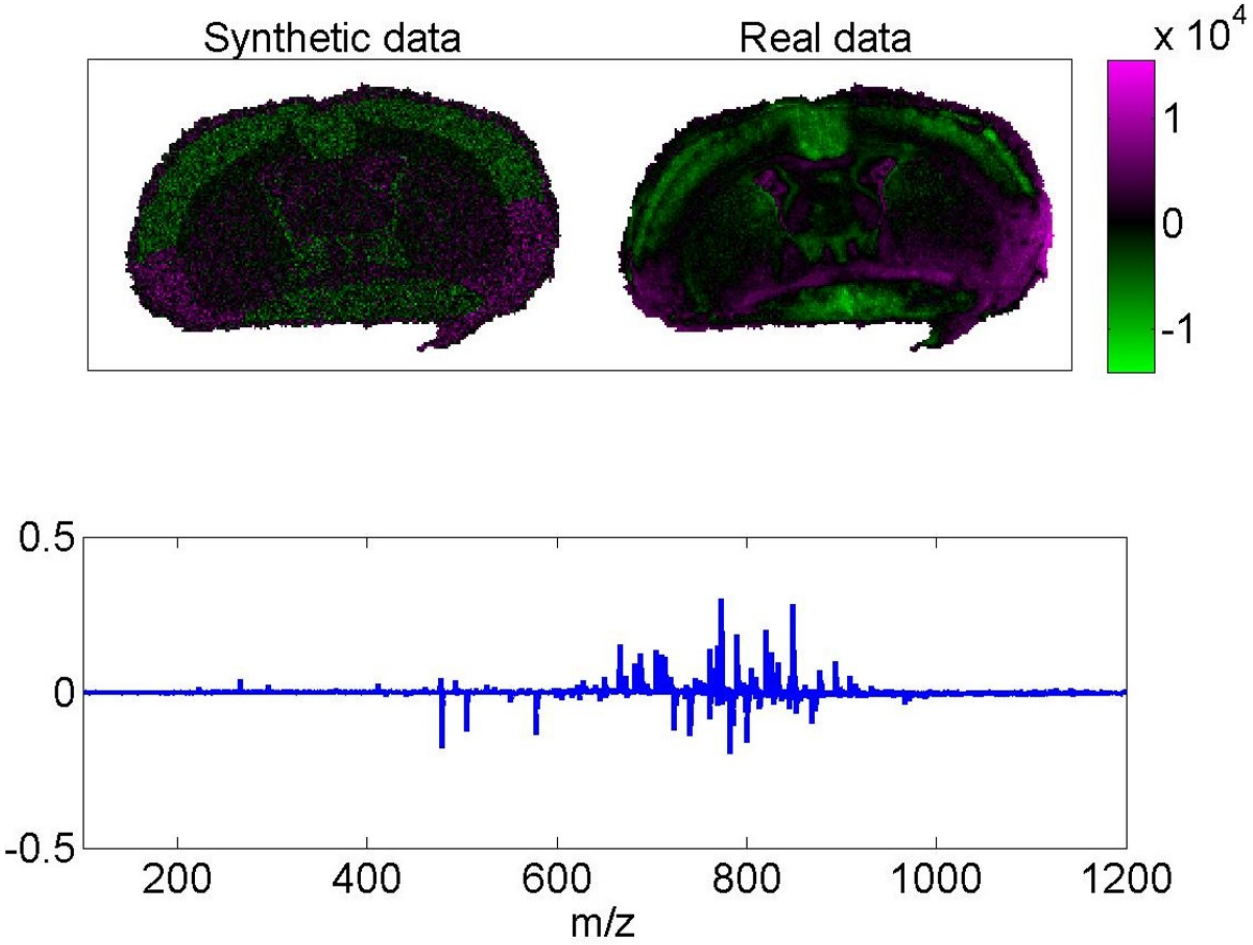

Figure 5.17: Principal component 7 from the combined real and synthetic dataset showing no distinct separation between the two.

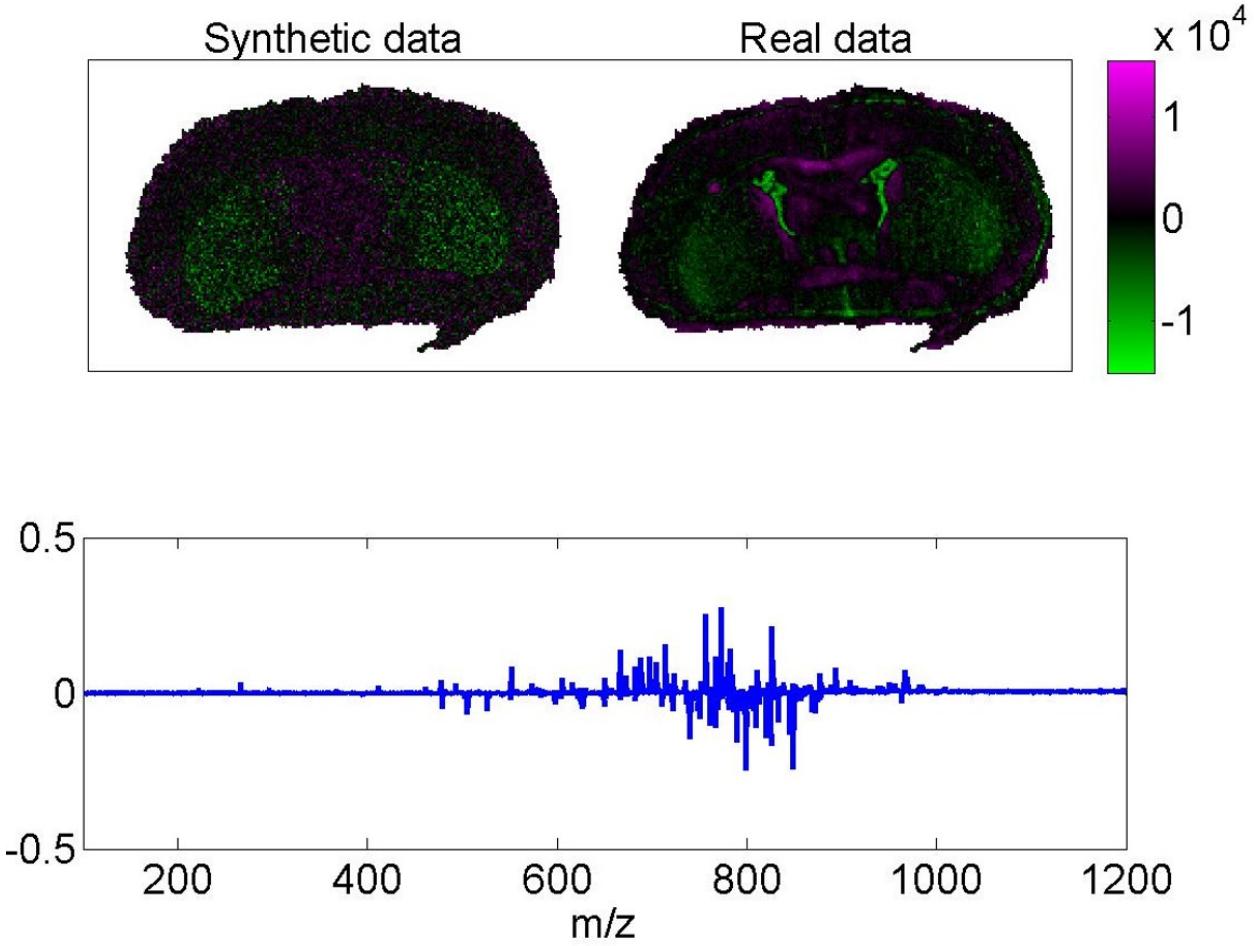

Figure 5.18: Principal component 8 from the combined real and synthetic dataset showing no distinct separation between the two.

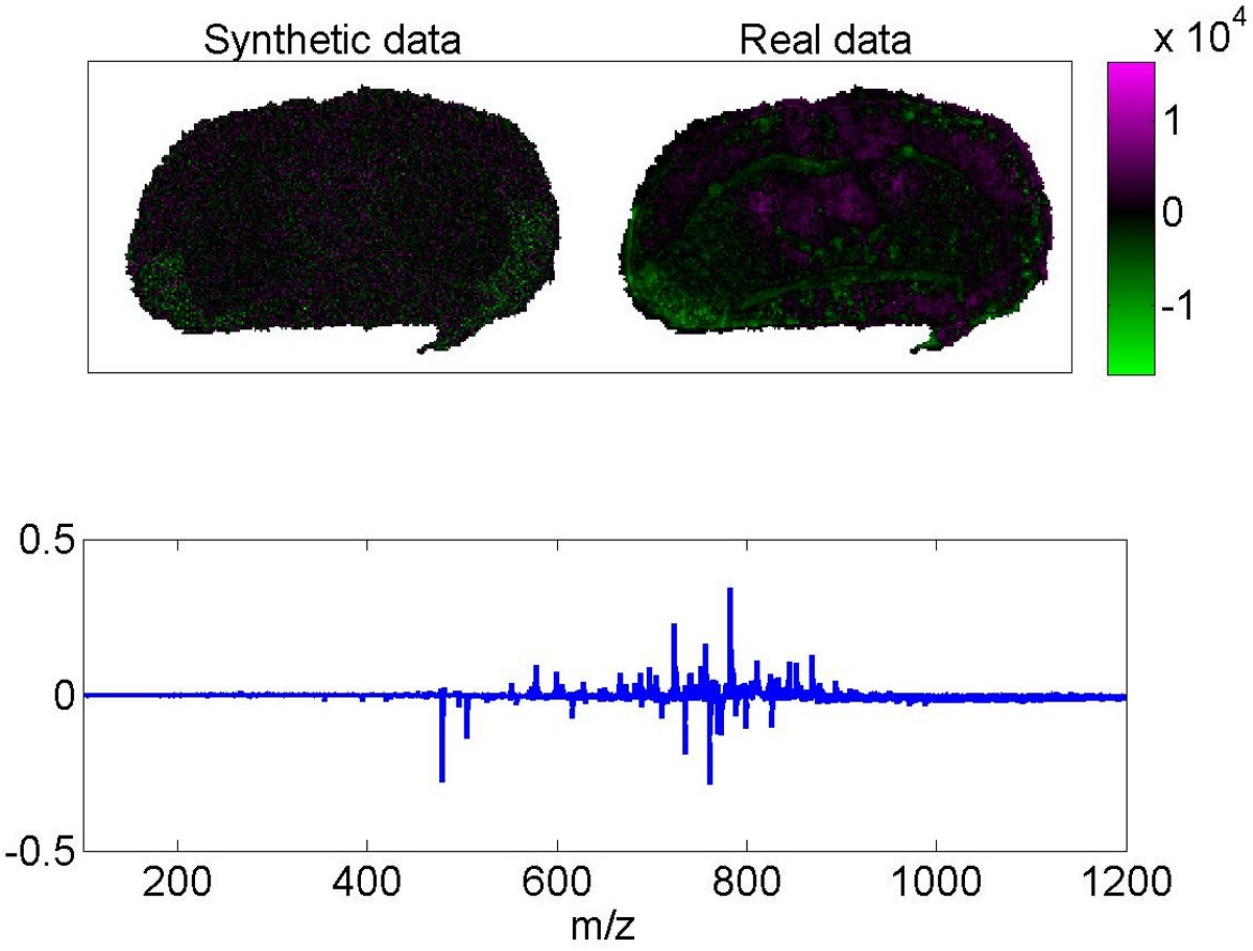

Figure 5.19: Principal component 9 from the combined real and synthetic dataset showing no distinct separation between the two.

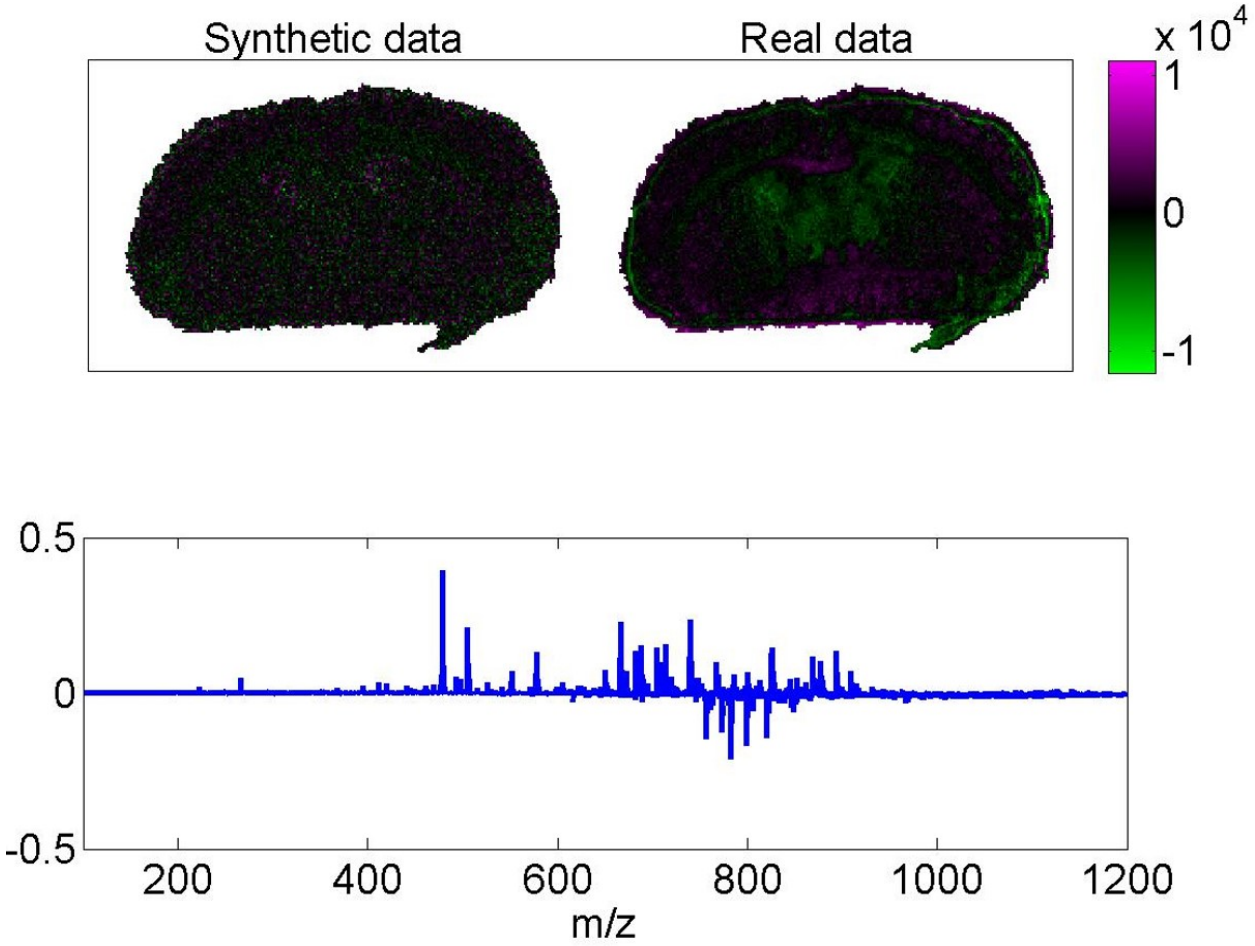

Figure 5.20: Principal component 10 from the combined real and synthetic dataset showing no distinct separation between the two.

As improvements are being made in throughput in MSI, datasets are rapidly increasing in size [95]. This then means that there will be a growing need for more and more efficient algorithms to perform multivariate analysis. Large synthetic datasets can be generated rapidly using the statistical modelling approach, by simply taking more samples from the multivariate normal distribution. To demonstrate this, a dataset containing nine times the number of pixels of the original reference data was generated (187,452 pixels from 20,825 in the original). This represents this size of data from an area three times the size in each dimension, or if the image had been acquired with 15 μm rather than 45 μm pixels. These new data were generated in approximately 5 minutes, but it would have required around 36 hours to acquire the same number of pixels experimentally, and the time limiting step

of covariance calculation need only be performed once, therefore this approach scales well with increasing number of pixels. PCA performed on a combined dataset containing the new larger dataset and the original reference data still shows no separation between the synthetic and real data, demonstrating that this approach scales to large datasets without any statistically detectable changes occurring in the data (Figures 5.21 to 5.30). While in both these cases the full seven regions were used to generate synthetic data, an image containing any desired number of regions can be generated using this approach, provided there is a suitable set of reference data. This means that the performance of different clustering algorithms or multivariate analysis methods can be evaluated with respect to the size and complexity of the data in terms of expected features. In addition, no new tissue sections are required, allowing the potential to minimise animal usage in computational studies in MSI. Of note, the synthetic images appear more speckled than the reference data. This is because when populating the spatial masks with spectra, no spatial smoothing is applied and neighbouring pixels are statistically independent. This could potentially be overcome by also maximising the similarity of neighbouring pixel, but for clustering evaluation this is unnecessary.

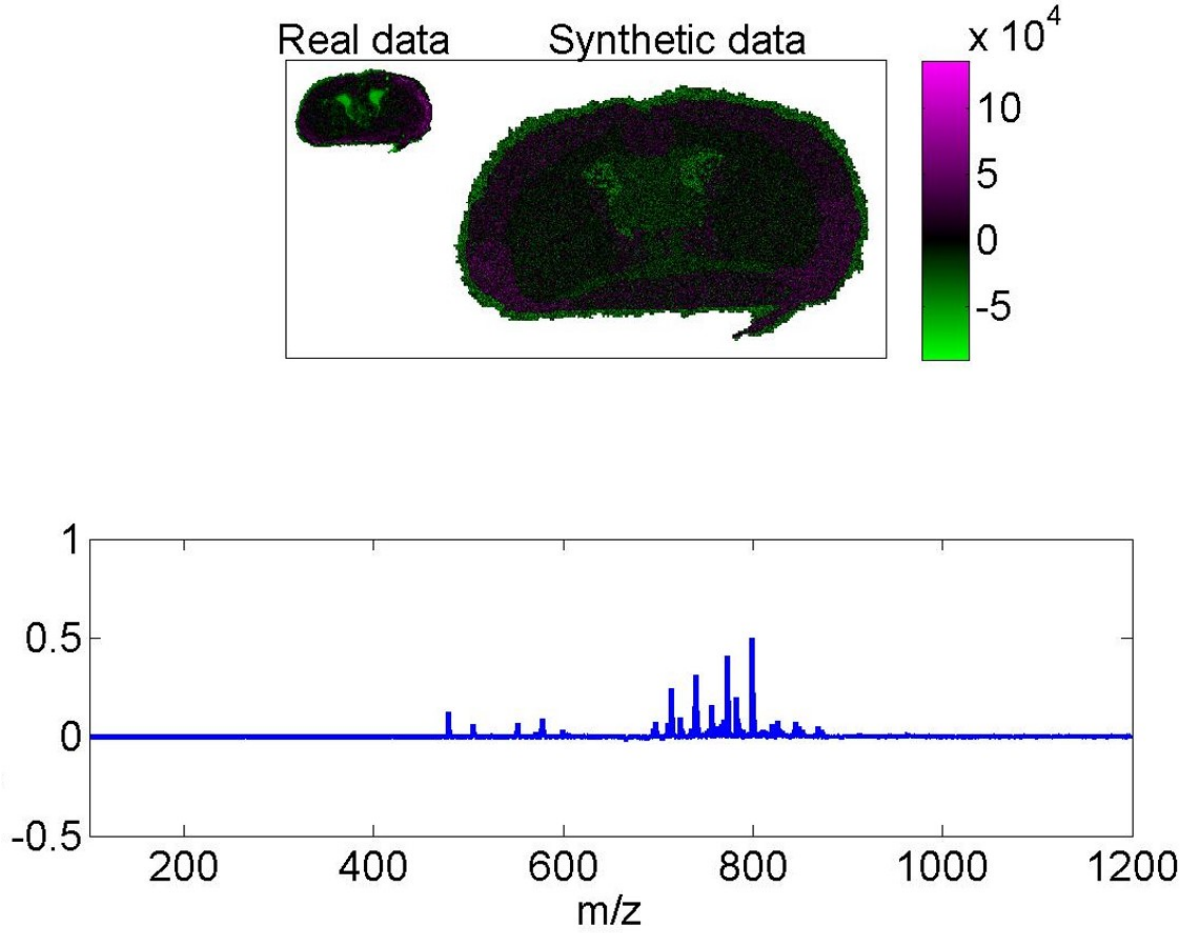

Figure 5.21: Principal component 1 from the combined real and large synthetic dataset, showing the scores image (top) and loadings plot (bottom) showing no distinct separation between the two. As wiith the small synthetic dataset, this indicates that scaling this method to resample larger data does not induce any changes to the data.

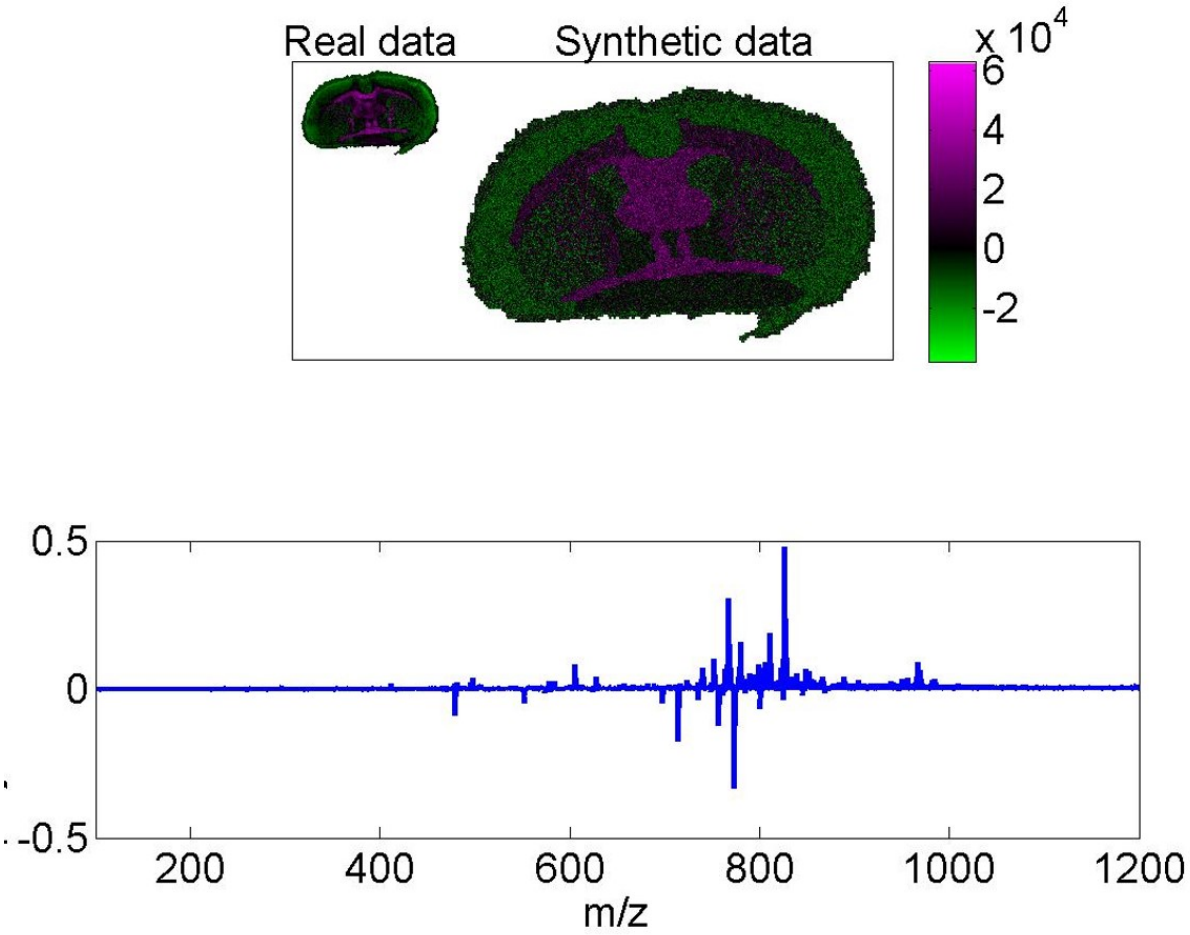

Figure 5.22: Principal component 2 from the combined real and large synthetic dataset, showing the scores image (top) and loadings plot (bottom) showing no distinct separation between the two. As wiith the small synthetic dataset, this indicates that scaling this method to resample larger data does not induce any changes to the data.

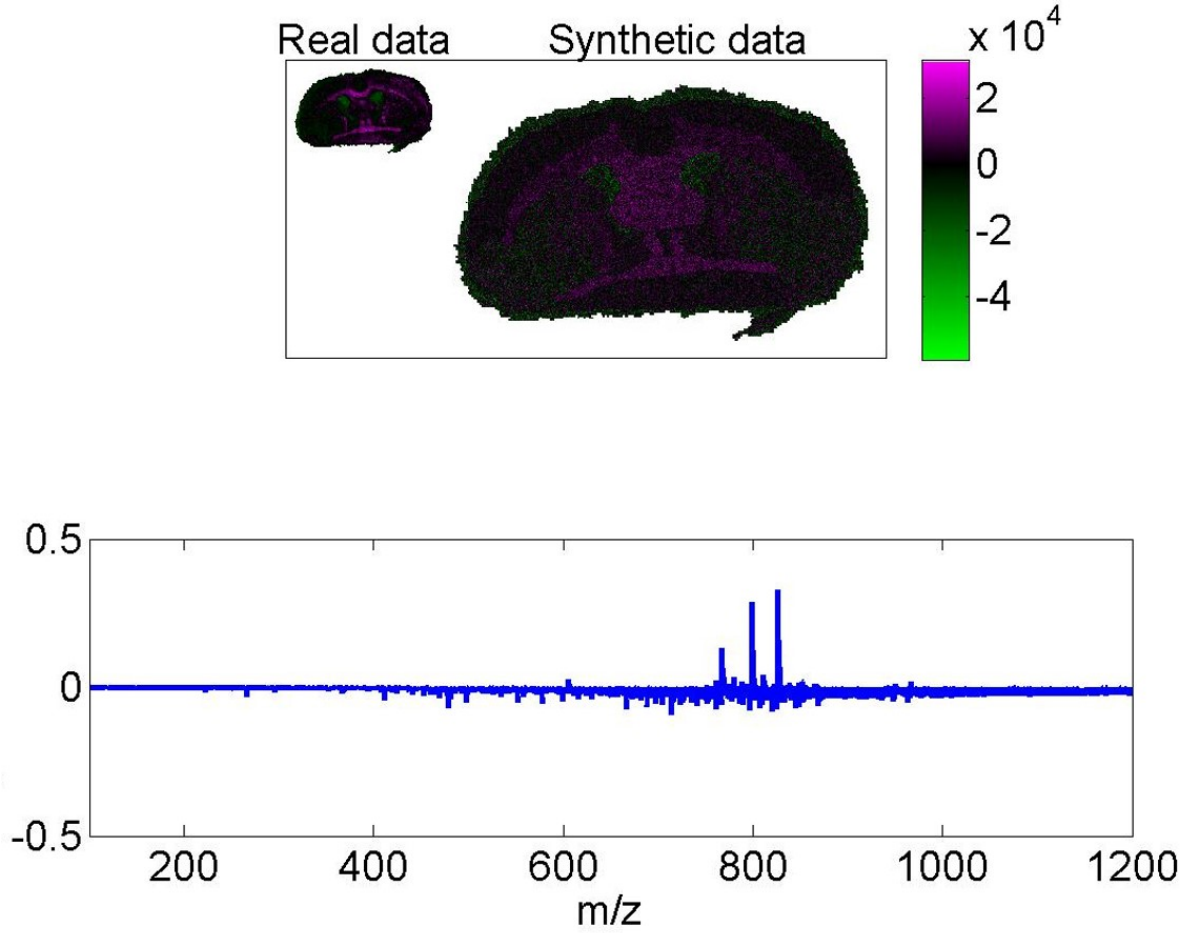

Figure 5.23: Principal component 3 from the combined real and large synthetic dataset, showing the scores image (top) and loadings plot (bottom) showing no distinct separation between the two. As wiith the small synthetic dataset, this indicates that scaling this method to resample larger data does not induce any changes to the data.

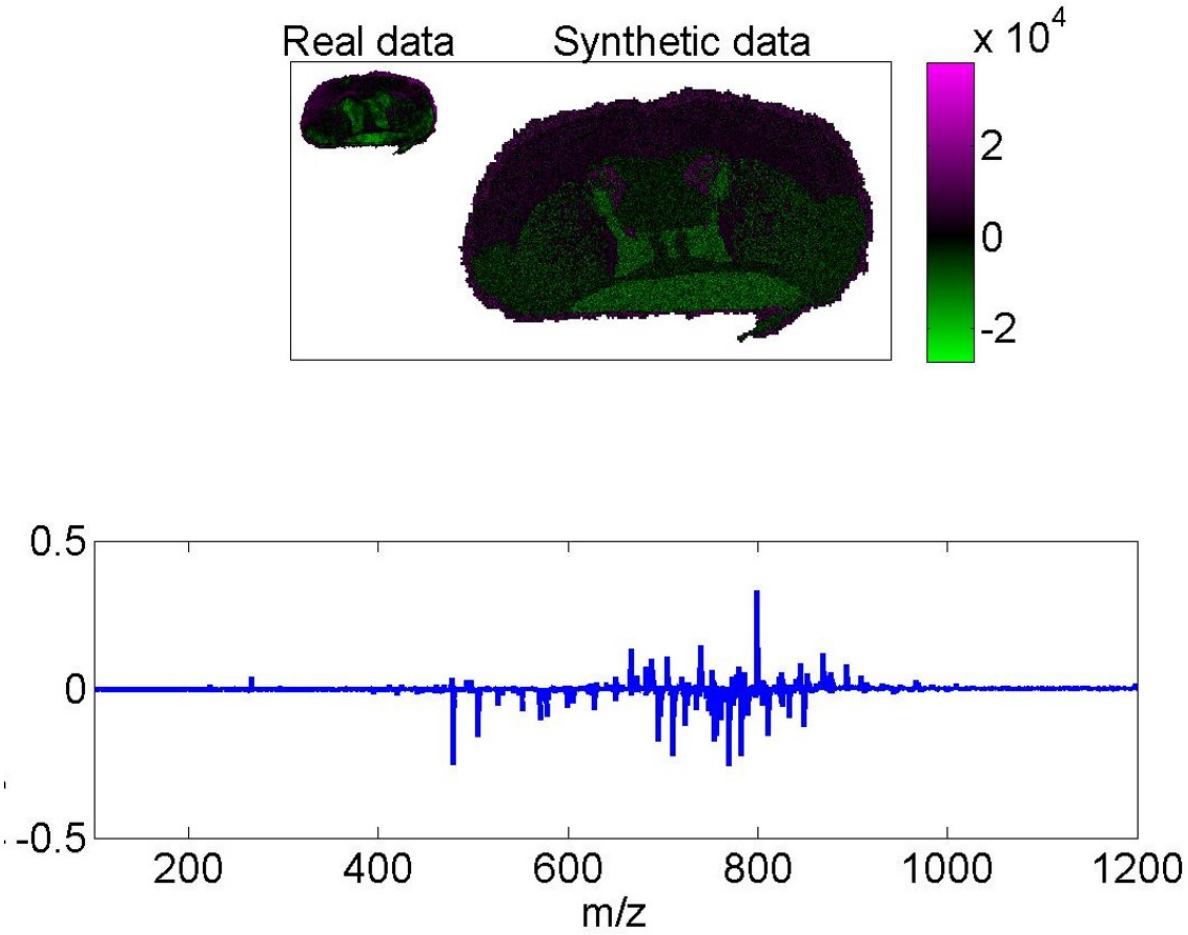

Figure 5.24: Principal component 4 from the combined real and large synthetic dataset, showing the scores image (top) and loadings plot (bottom) showing no distinct separation between the two. As wiith the small synthetic dataset, this indicates that scaling this method to resample larger data does not induce any changes to the data.

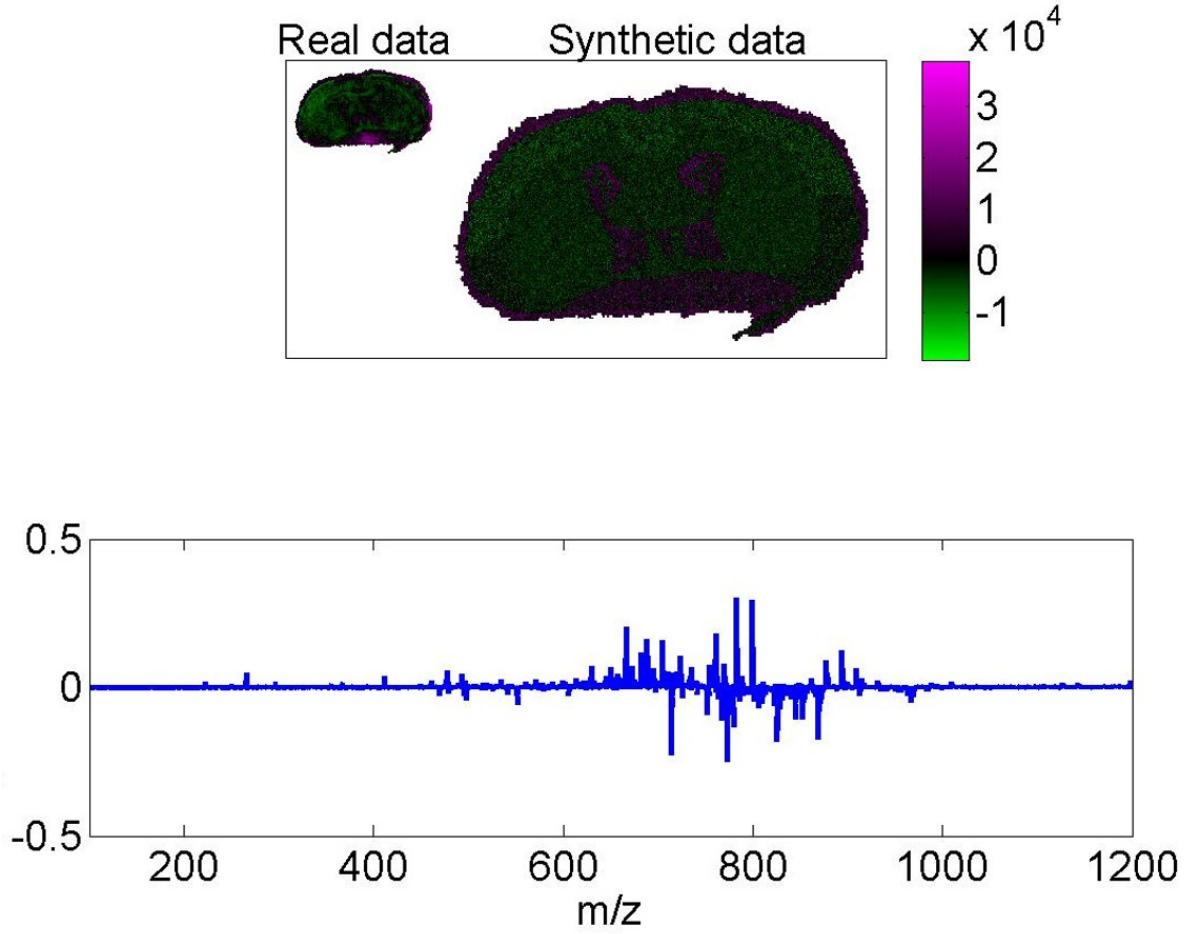

Figure 5.25: Principal component 5 from the combined real and large synthetic dataset, showing the scores image (top) and loadings plot (bottom) showing no distinct separation between the two. As wiith the small synthetic dataset, this indicates that scaling this method to resample larger data does not induce any changes to the data.

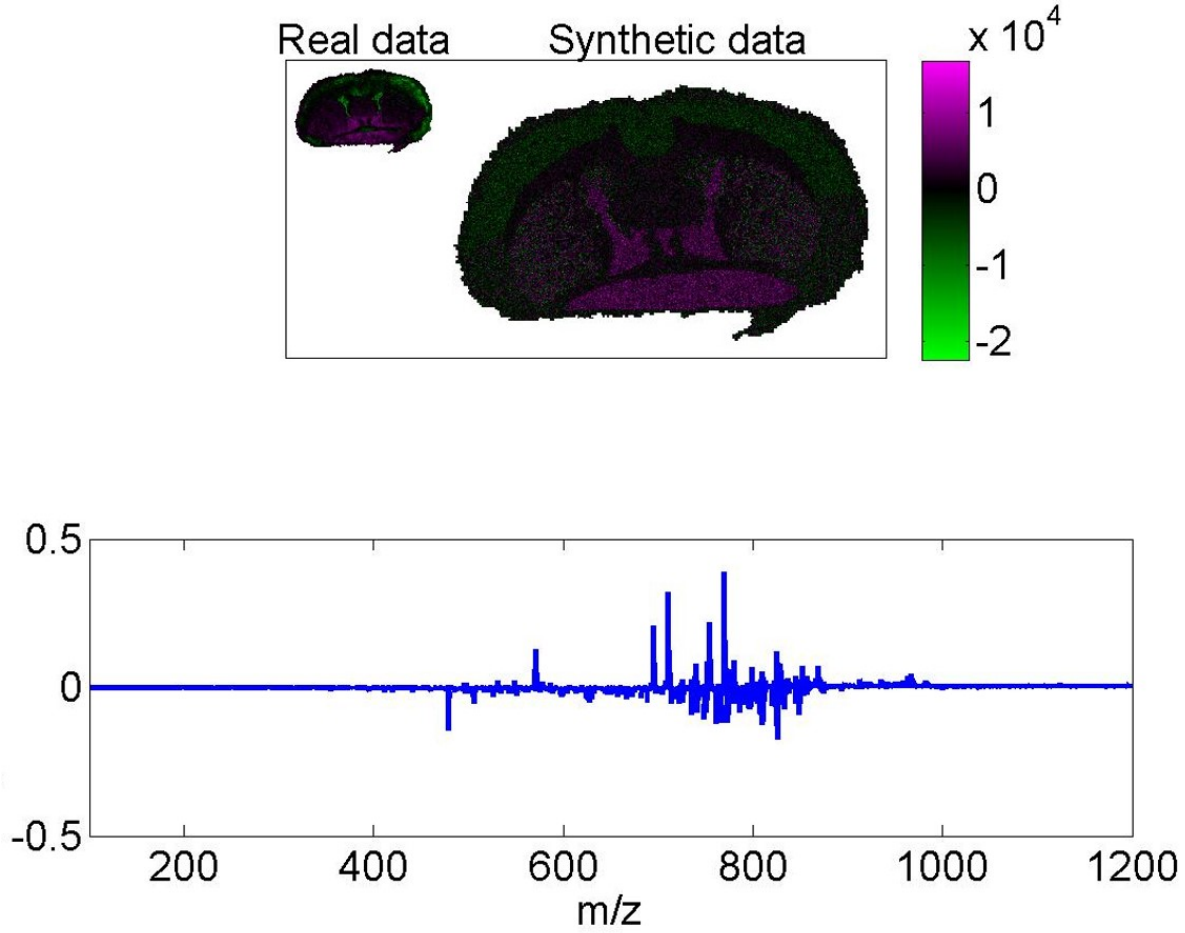

Figure 5.26: Principal component 6 from the combined real and large synthetic dataset, showing the scores image (top) and loadings plot (bottom) showing no distinct separation between the two. As wiith the small synthetic dataset, this indicates that scaling this method to resample larger data does not induce any changes to the data.

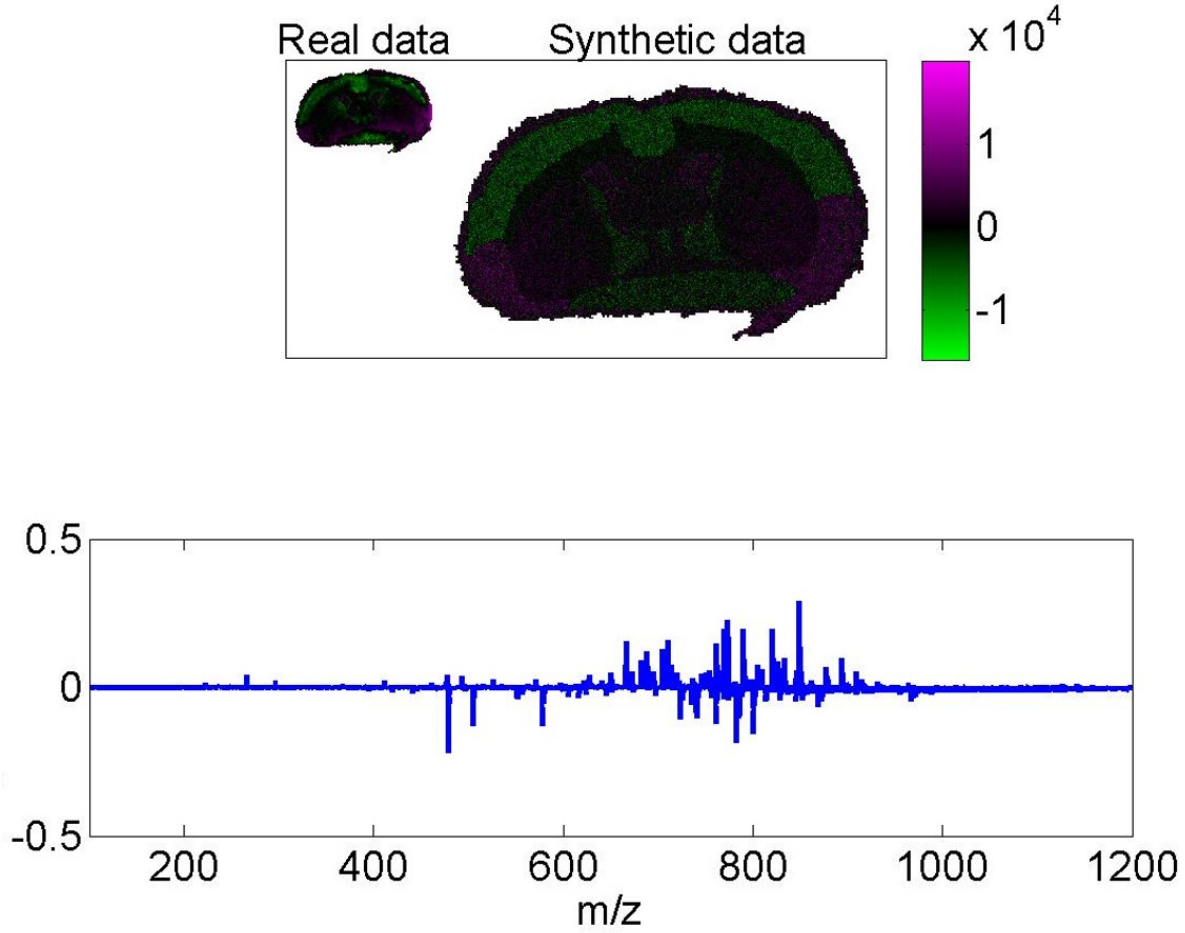

Figure 5.27: Principal component 7 from the combined real and large synthetic dataset, showing the scores image (top) and loadings plot (bottom) showing no distinct separation between the two. As wiith the small synthetic dataset, this indicates that scaling this method to resample larger data does not induce any changes to the data.

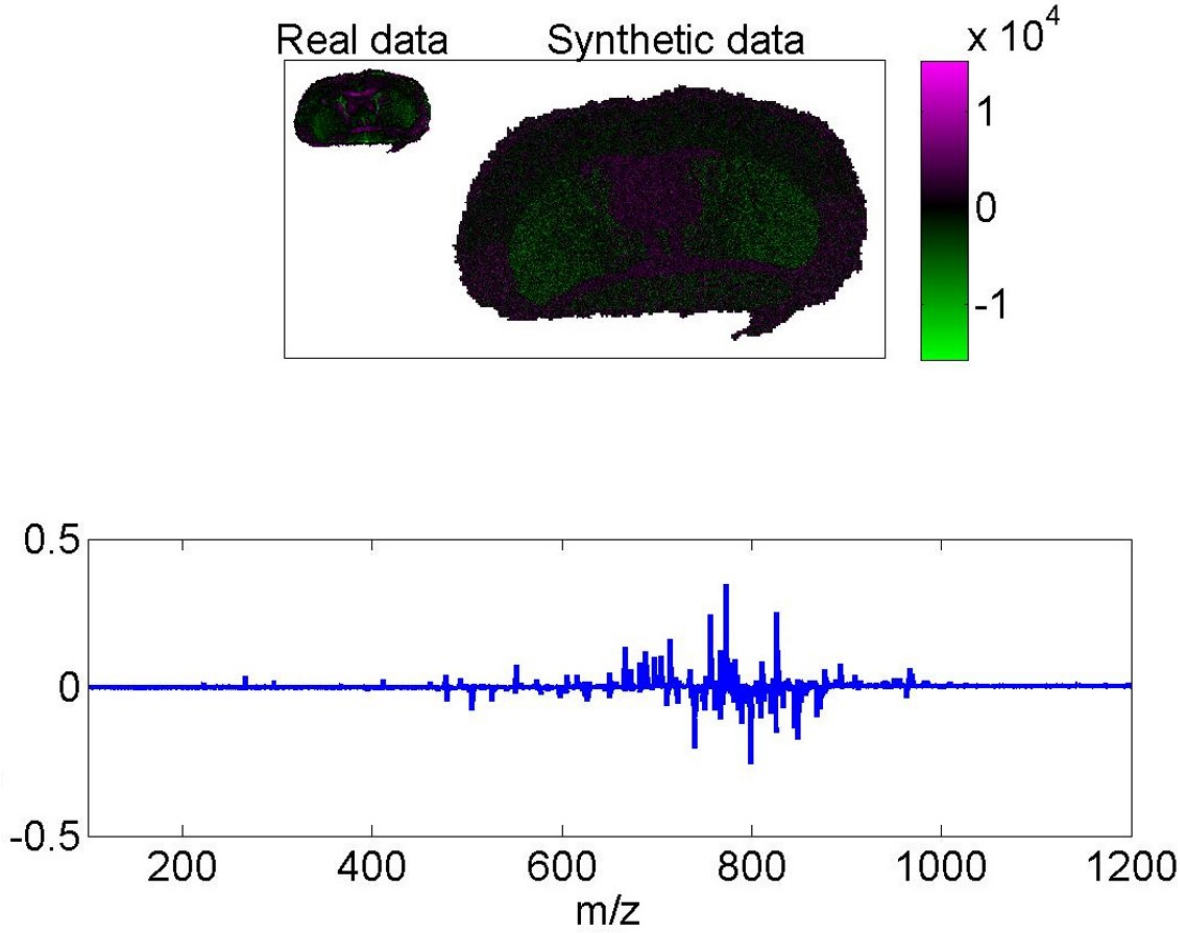

Figure 5.28: Principal component 8 from the combined real and large synthetic dataset, showing the scores image (top) and loadings plot (bottom) showing no distinct separation between the two. As wiith the small synthetic dataset, this indicates that scaling this method to resample larger data does not induce any changes to the data.

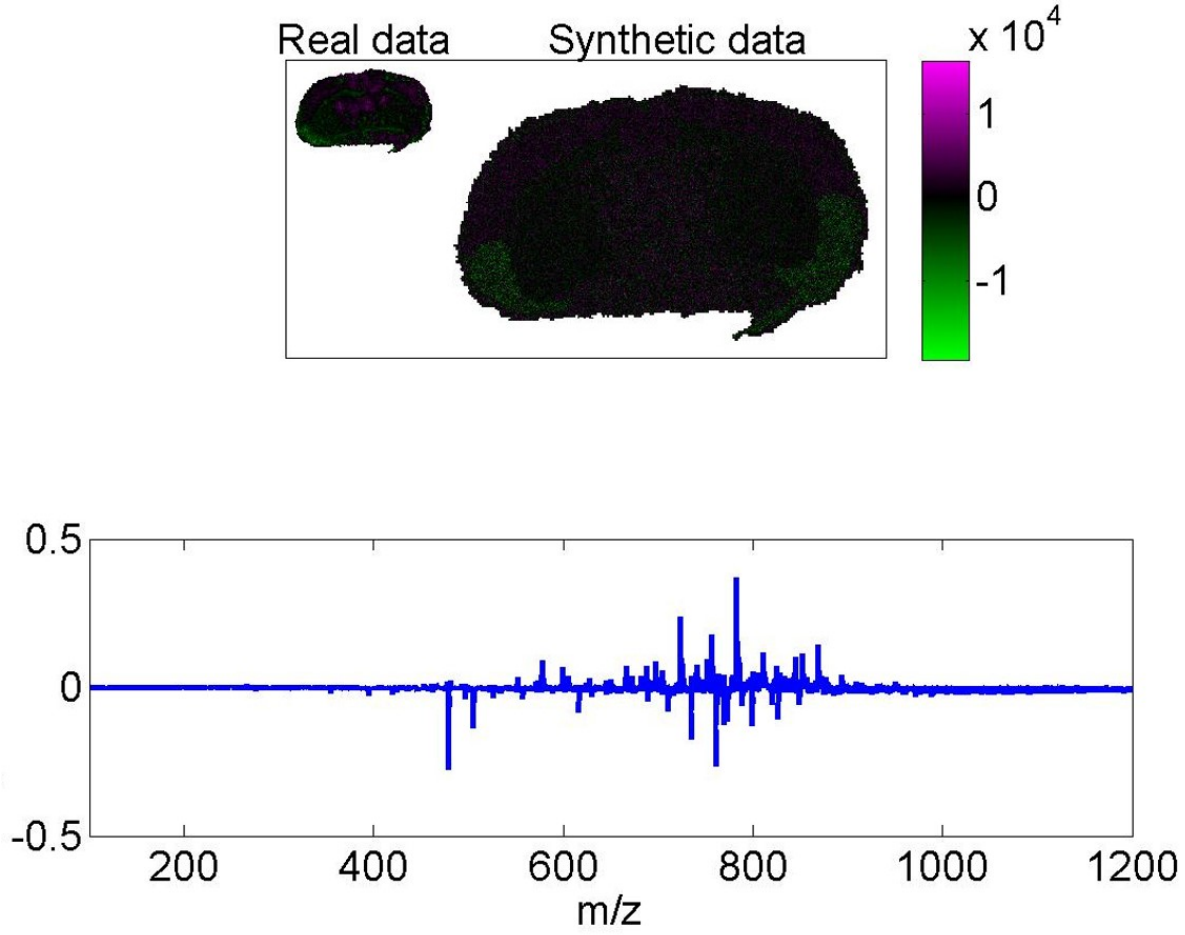

Figure 5.29: Principal component 9 from the combined real and large synthetic dataset, showing the scores image (top) and loadings plot (bottom) showing no distinct separation between the two. As wiith the small synthetic dataset, this indicates that scaling this method to resample larger data does not induce any changes to the data.

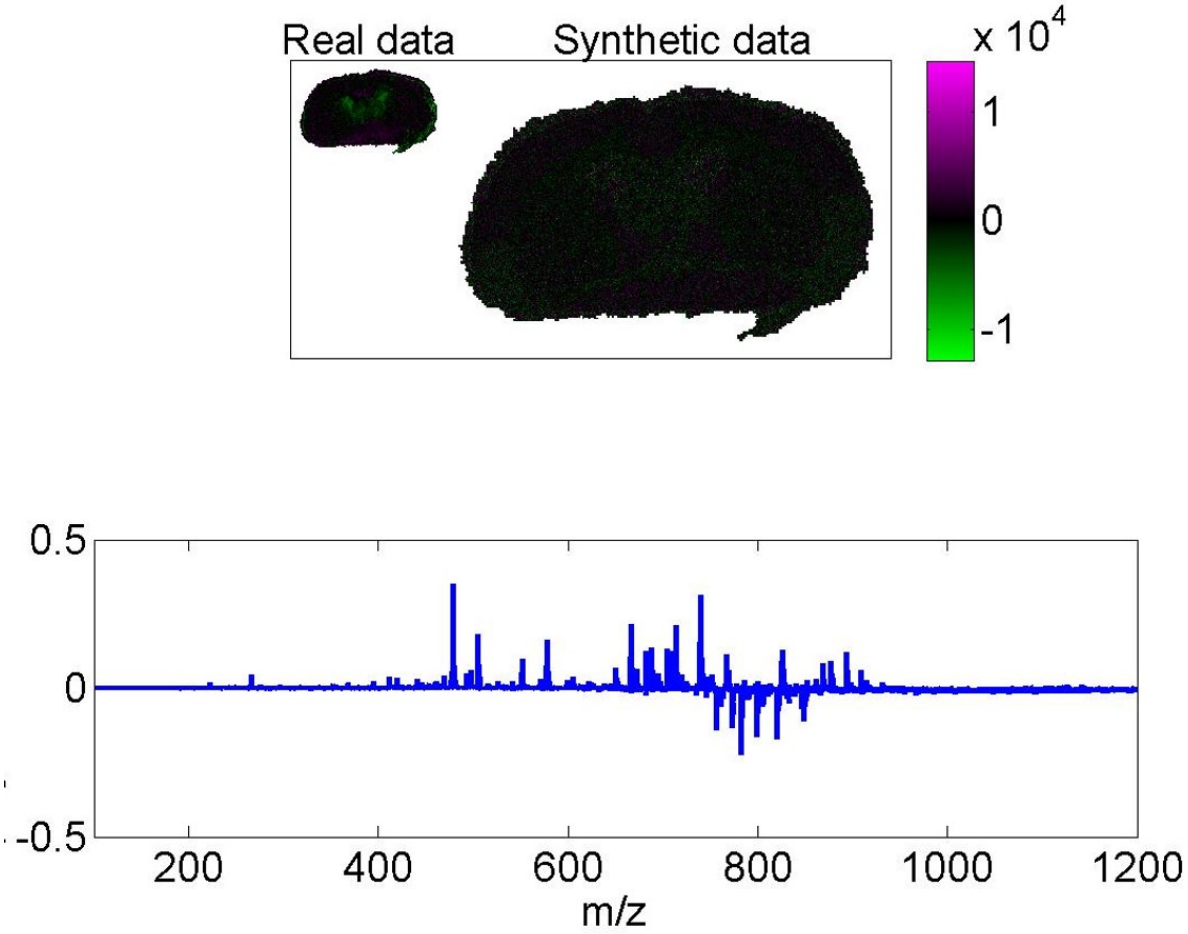

Figure 5.30: Principal component 10 from the combined real and large synthetic dataset, showing the scores image (top) and loadings plot (bottom) showing no distinct separation between the two. As wiith the small synthetic dataset, this indicates that scaling this method to resample larger data does not induce any changes to the data.

In order to generate multivariate normally distributed synthetic data, the m/z covariance matrix $\underline{\underline{\mathbf{C}}}$ is required. The size of this matrix scales by the number of mass channels squared, and in the case of high mass resolution MS data (greater than 50,000 m/z bins or peaks) would be unrealistic to store in memory [129]. Efficient, and reversable dimensionality reduction can be performed on MSI data by the use of random projection [133]. In random projection, the MSI image with n pixels and d mass channels is projected by a k by d matrix of normally distributed random numbers. This can then be orthonormalised by QR decomposition, and a low dimensional projection of the data created (Figure 5.2).

Statistical modelling can then be performed on this reduced data, and synthetic datasets generated. These synthetic reduced datasets can then be projected back into the original space to recreate the appropriate mass spectra. As with the original workflow, datasets created when incorporating random projection into the workflow show no visible difference by PCA from either the original data or the original synthetic data (Figures 5.32 to 5.41). Generating synthetic datasets with the addition of random projection into the workflow allows high mass resolution datasets to be generated where previously the covariance matrix would be too large to calculate. This means that large datasets in both number of pixels and mass channels can be created allowing users to quantitatively evaluate novel algorithms for future instrumentation. This also significantly speeds up the synthetic data generation primarily by reducing the time taken to calculate the covariance matrices. Currently, 50 pixels per second is the highest recorded MALDI imaging acquisition rate [95], whereas synthetic data generated by these means can produces data at a rate of 500 pixels per second without the use of random projection (Figure 5.42), and 12,000 pixels per second when random projection is used to reduce the data to 100 dimensions (Figure 5.43). Comparing data acquired at this rate with synthetic data generation, with and without the inclusion of random projection highlights the dramatic increase in speed (Figure 5.44) without considering time taken for sample preparation, or the benefits of not requiring instrument time, or further precious tissue samples. While there will always be the need for acquiring real data, these synthetic data provide a means for evaluating computational methods in a robust and efficient manner.

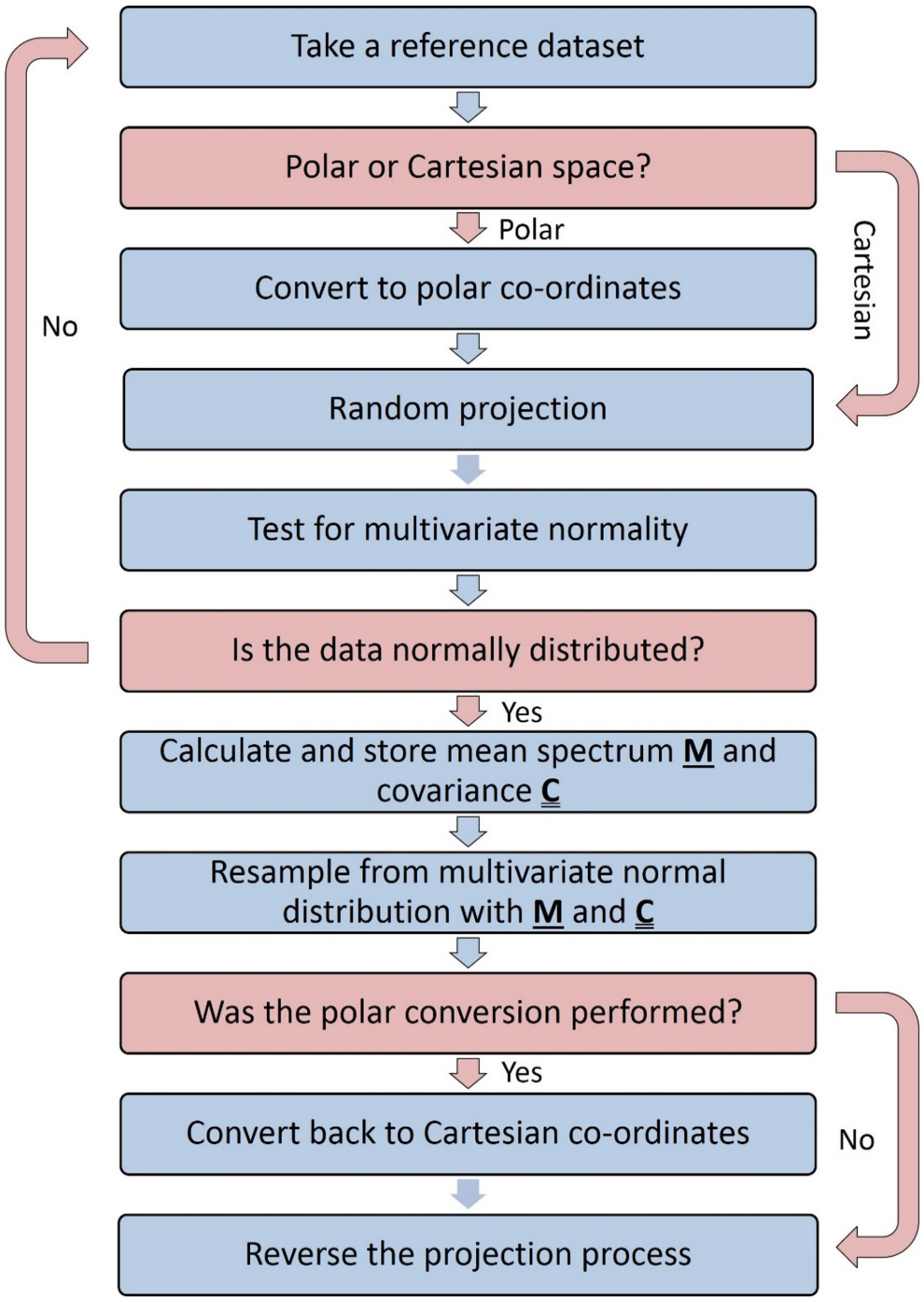

Figure 5.31: Workflow for the generation of synthetic datasets with the inclusion of random projection.

input : data matrix D of n pixels by d mass channels

input : desired number of pixel p

output: Synthetic data with p pixels and d mass channels

```
if using polar normal data then
    perform polar co-ordninate conversion ;
end
test D for multivariate normality;
if D is normal then
    calculate mean spectrum M and covariance C of D ;
    generate synthetic data Z by resampling randomly from a multivariate normal
     distribution with mean M covariance C;
    if polar conversion performed then
        convert Z back to cartesian co-ordinates;
    end
end
```

Algorithm 10: Algorithm for synthetic data generation

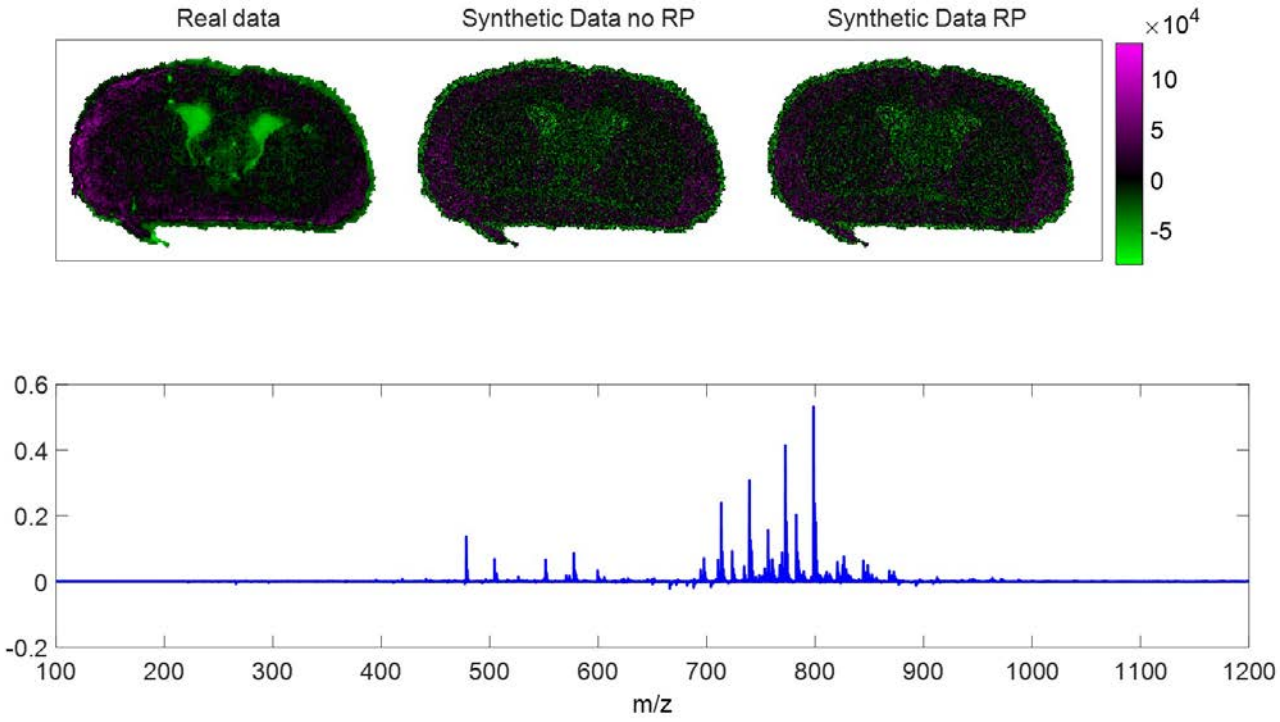

Figure 5.32: Principal component 1 from the combined real and synthetic dataset with and without random projection showing no distinct separation between any of these data.

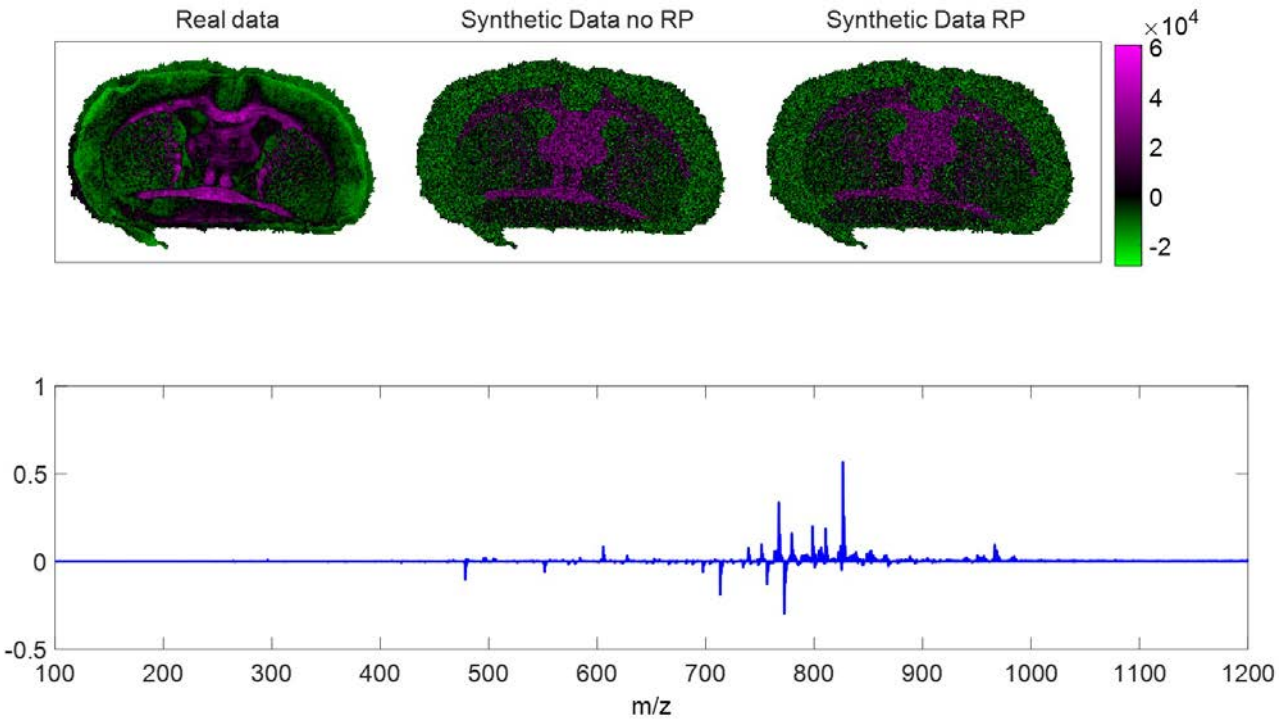

Figure 5.33: Principal component 2 from the combined real and synthetic dataset with and without random projection showing no distinct separation between any of these data.

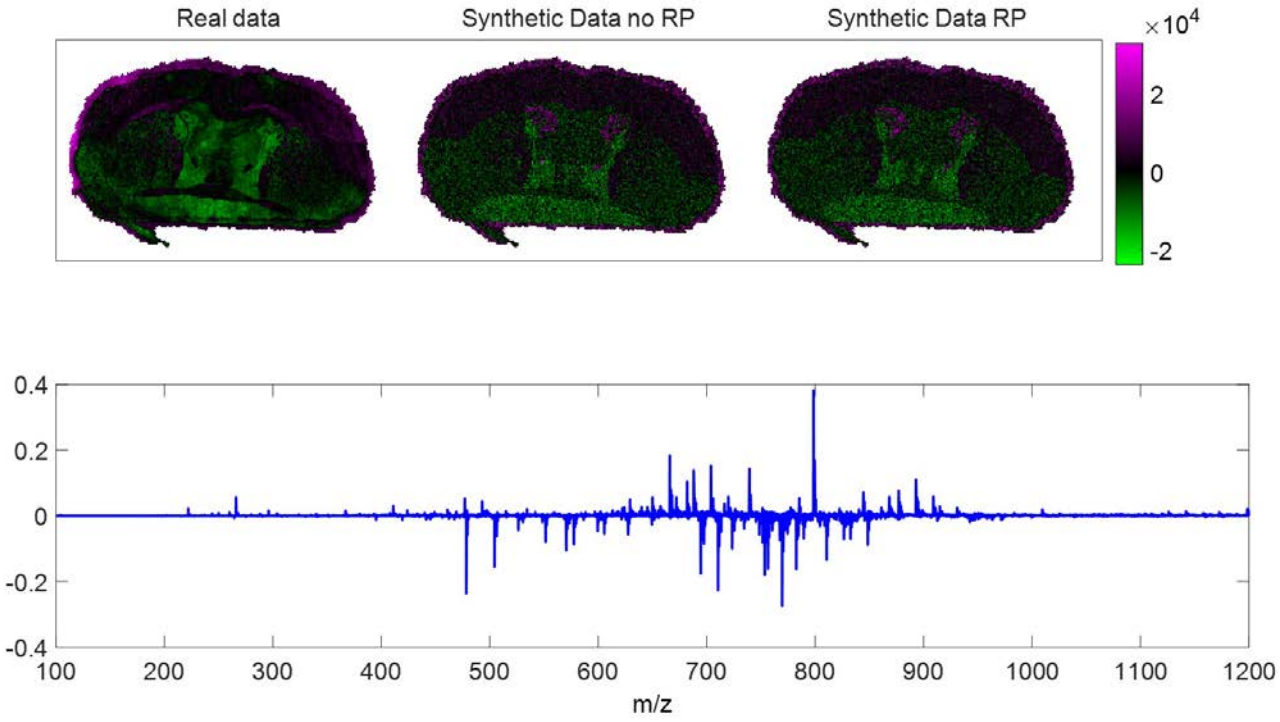

Figure 5.34: Principal component 3 from the combined real and synthetic dataset with and without random projection showing no distinct separation between any of these data.

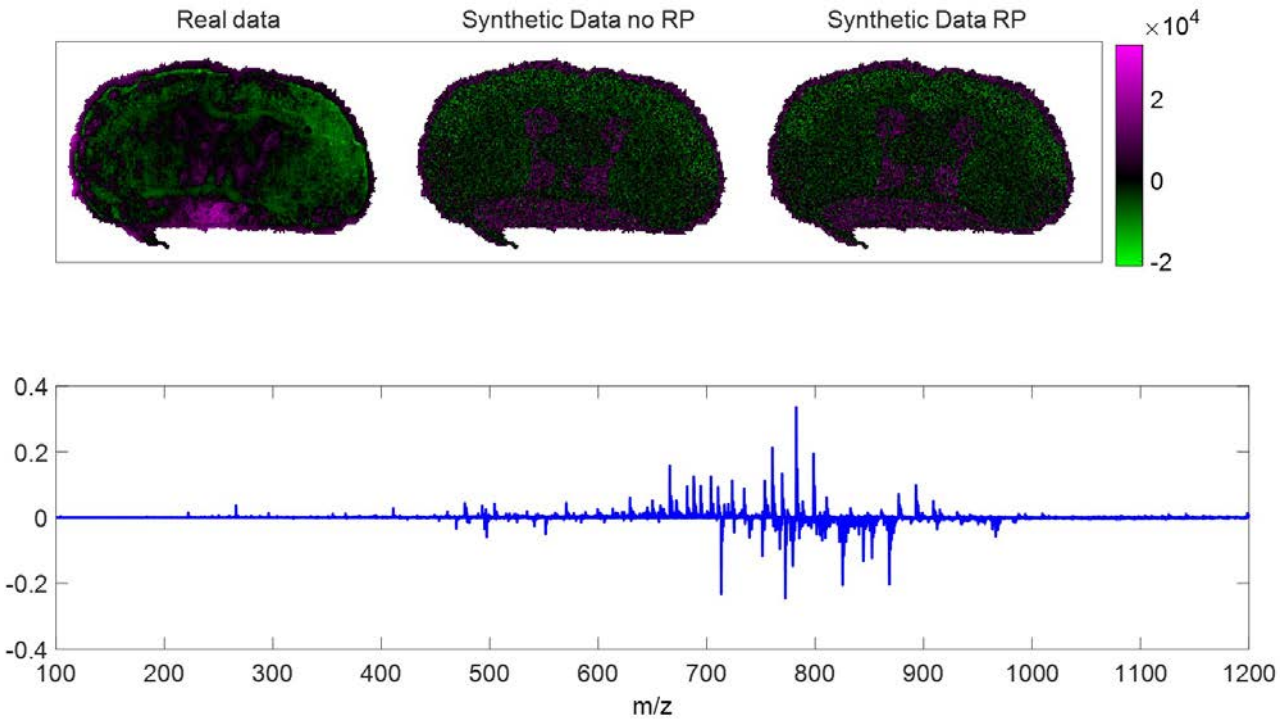

Figure 5.35: Principal component 4 from the combined real and synthetic dataset with and without random projection showing no distinct separation between any of these data.

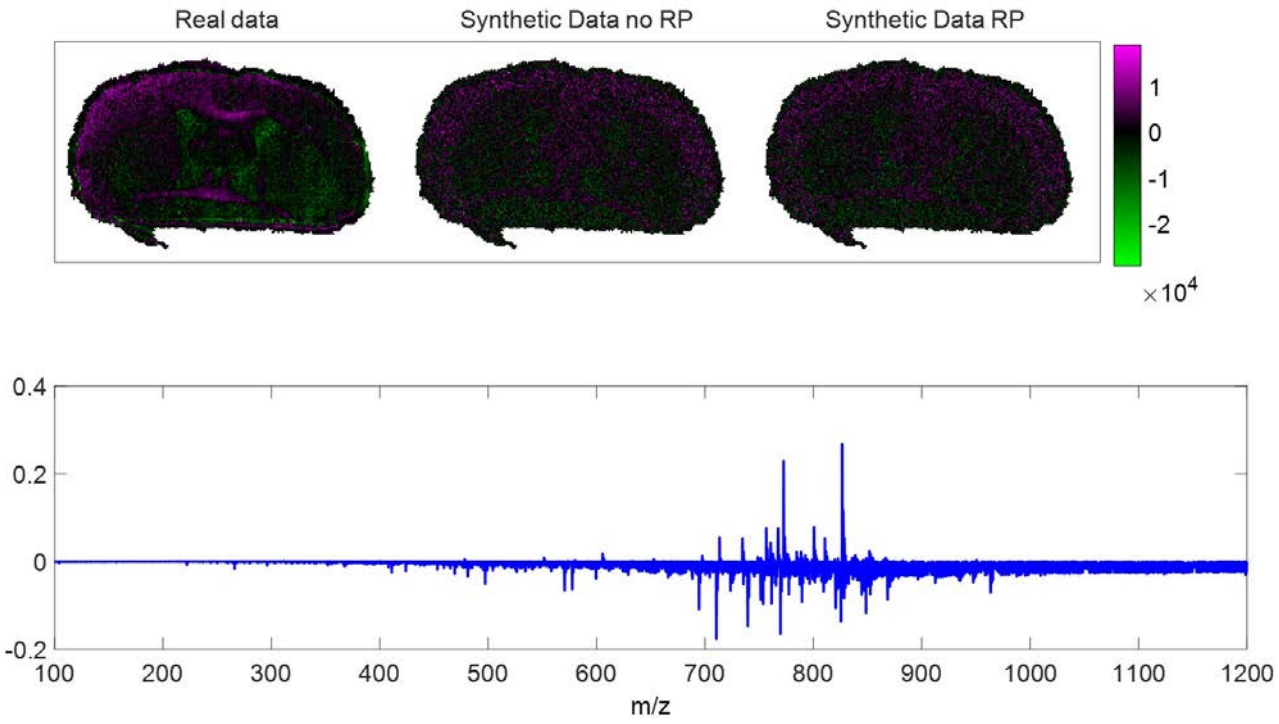

Figure 5.36: Principal component 5 from the combined real and synthetic dataset with and without random projection showing no distinct separation between any of these data.

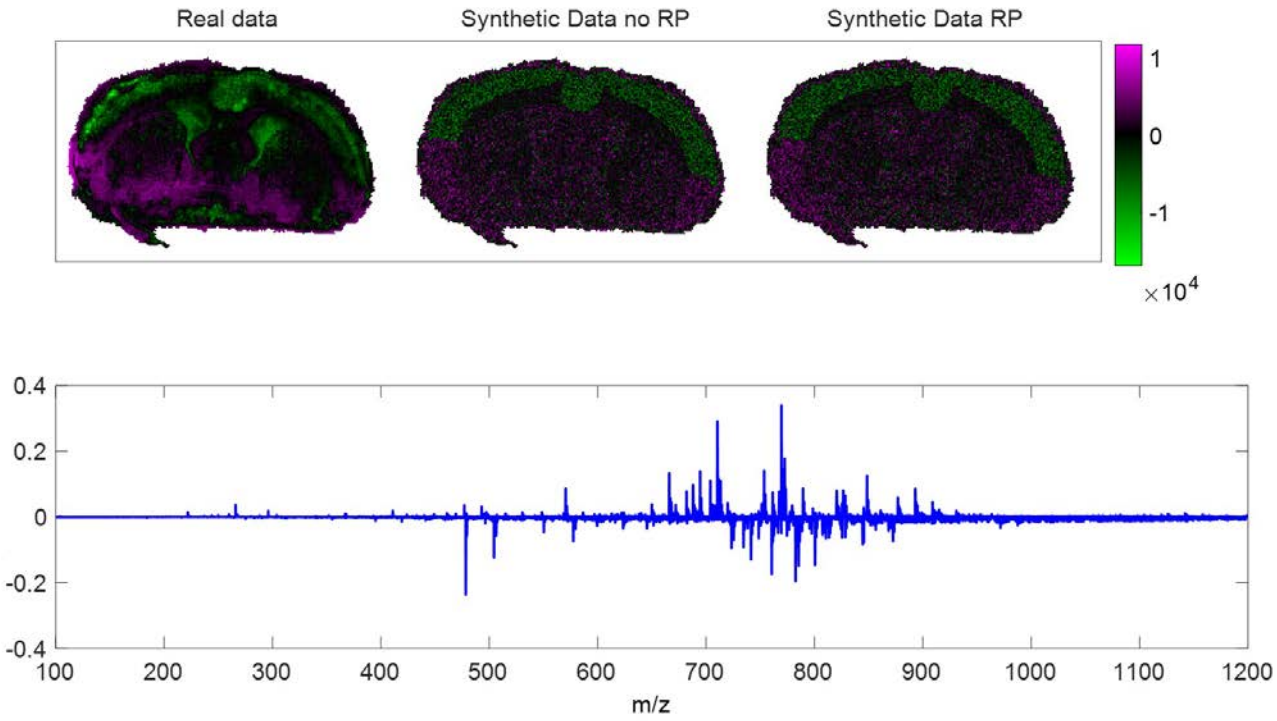

Figure 5.37: Principal component 6 from the combined real and synthetic dataset with and without random projection showing no distinct separation between any of these data.

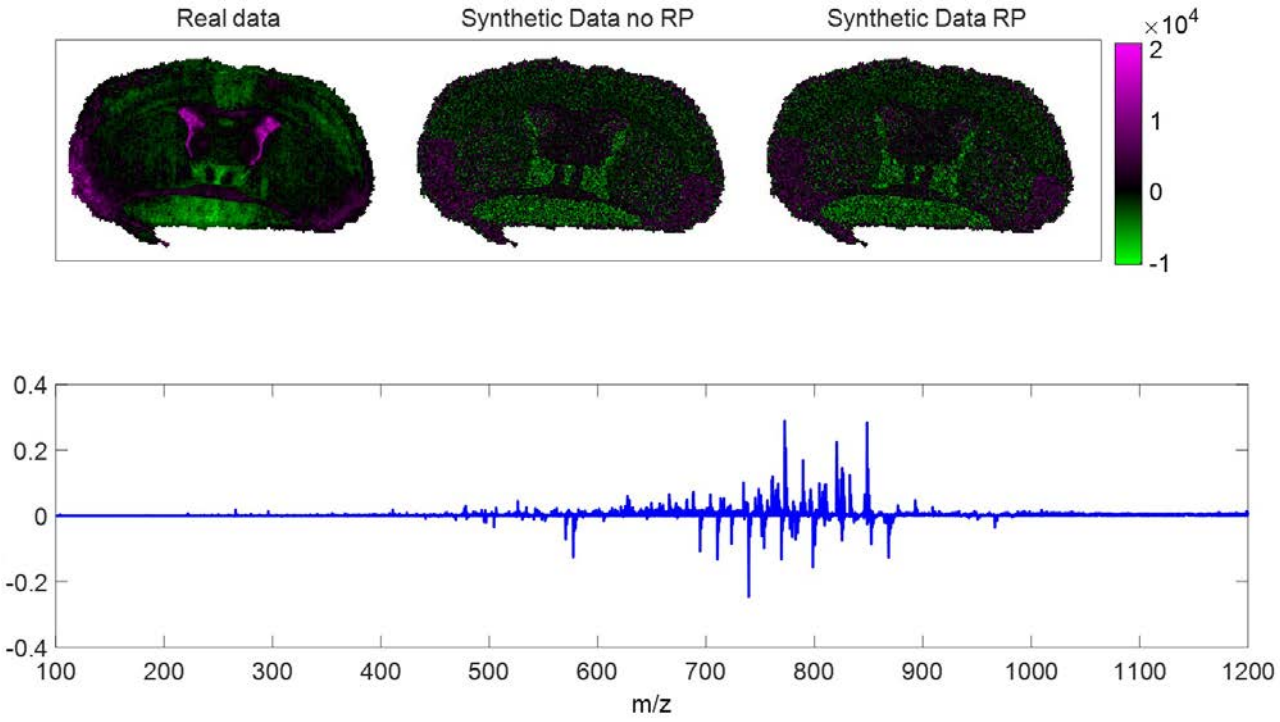

Figure 5.38: Principal component 7 from the combined real and synthetic dataset with and without random projection showing no distinct separation between any of these data.

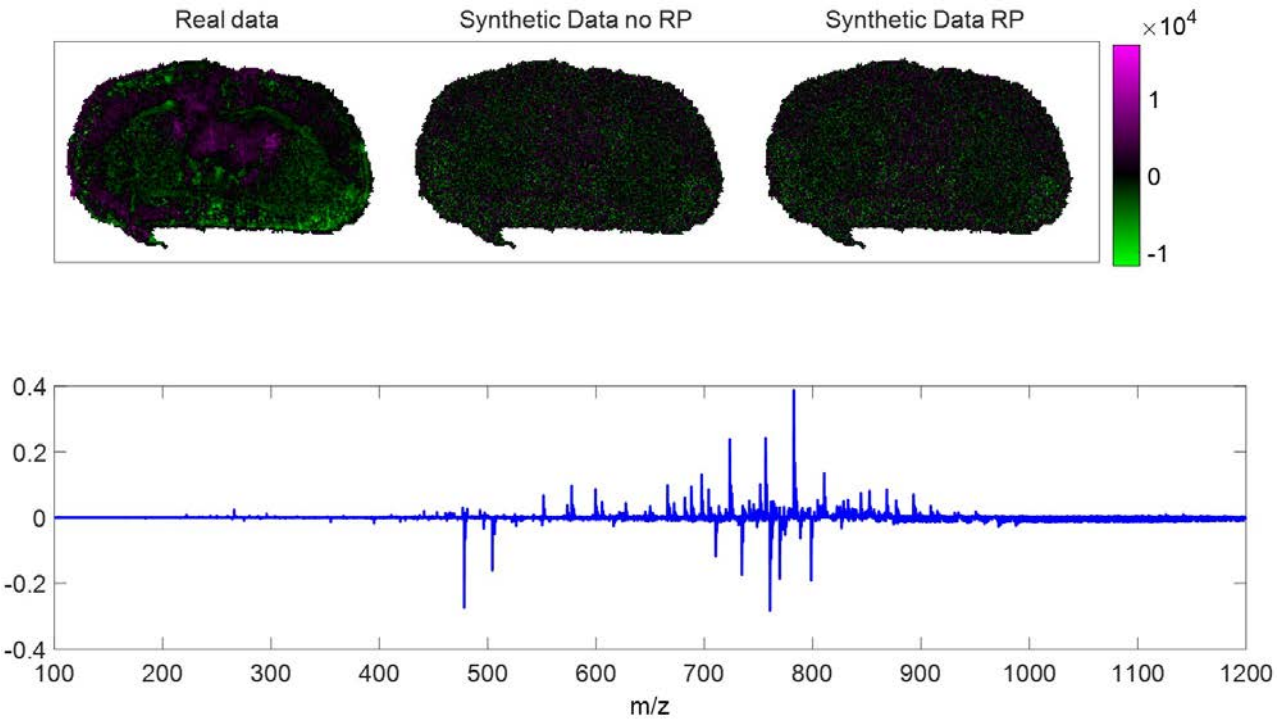

Figure 5.39: Principal component 8 from the combined real and synthetic dataset with and without random projection showing no distinct separation between any of these data.

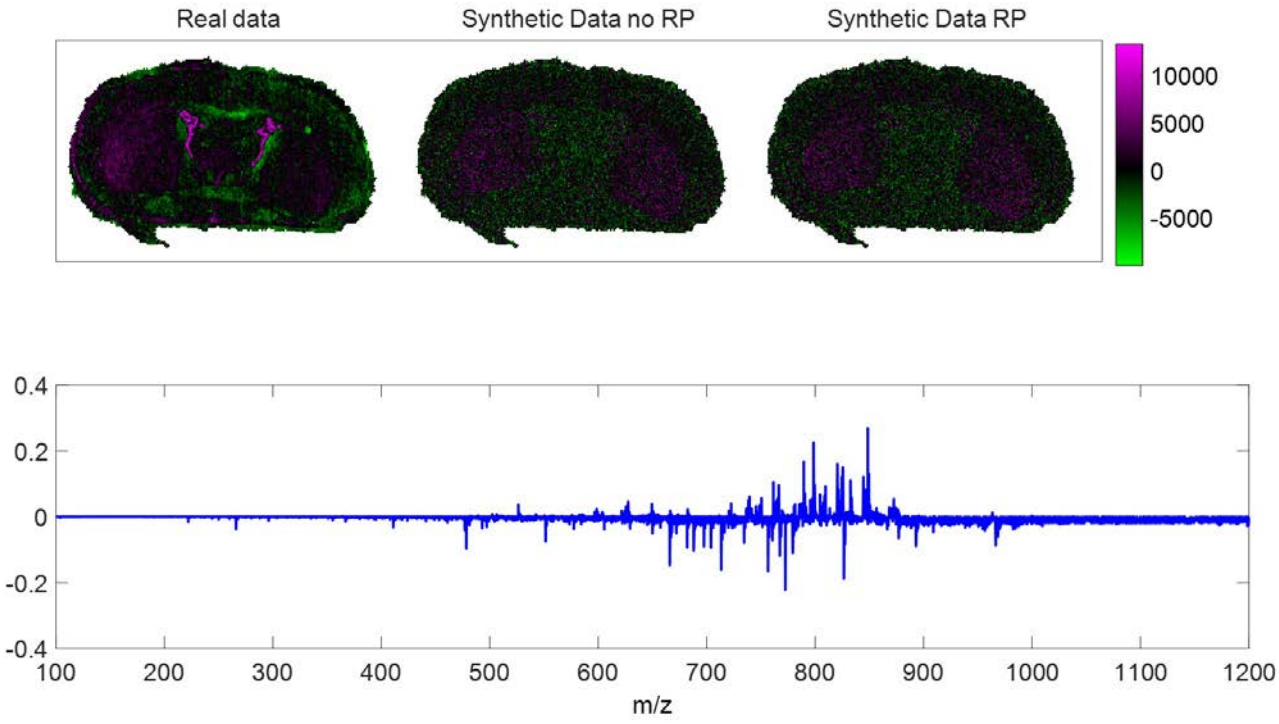

Figure 5.40: Principal component 9 from the combined real and synthetic dataset with and without random projection showing no distinct separation between any of these data.

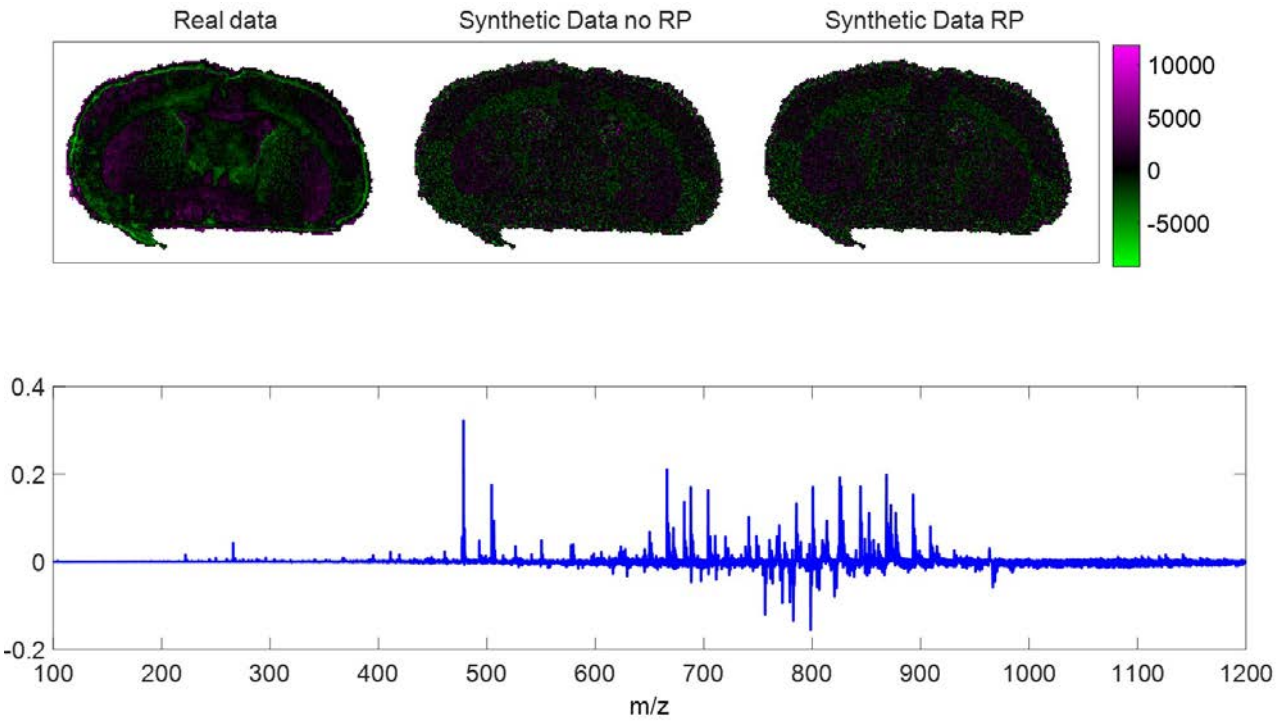

Figure 5.41: Principal component 10 from the combined real and synthetic dataset with and without random projection showing no distinct separation between any of these data.

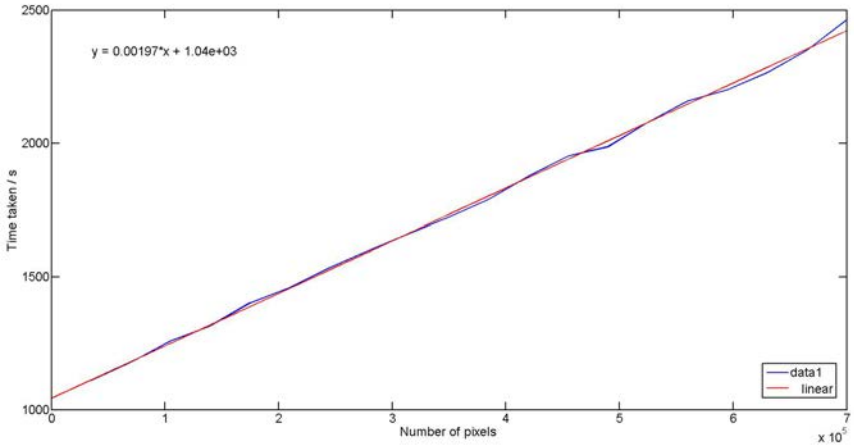

Figure 5.42: Time taken to generate synthetic data with ∼7,000 peaks without the use of random projection. This includes the offset for model generation (∼1000s).

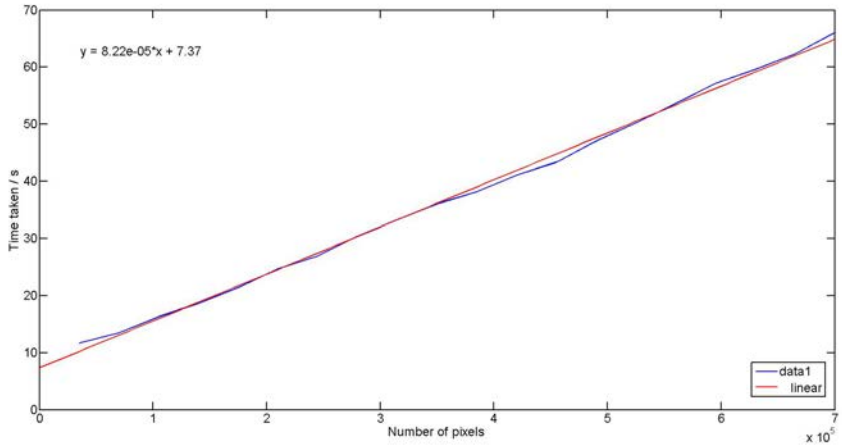

Figure 5.43: Time taken to generate synthetic data when using random projection to reduce the data to 100 dimensions. This reduces both the time taken per pixel and the model generation times significantly.

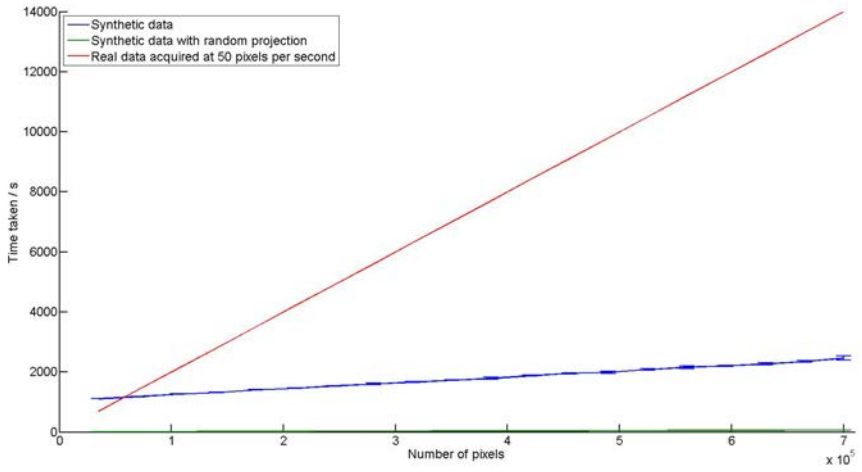

Figure 5.44: Comparison of the times taken to generate synthetic data to the current fastest experimental acquisition available.

These data were also exported into the open imzML format using the imzML converter and custom Matlab scripts, allowing them to be loaded into common MSI software packages such as SpectralAnalysis [191](Figure 5.45). This means that any processing tools available with these software can be tested on synthetic data, and by combining datasets in imzML converter [118], synthetic and real data can be directly compared with one

another (Figure 5.46). Future work could involve developing this into a software package of its own to allow users to easily generate synthetic data with predefined shapes and known spectral derivation.

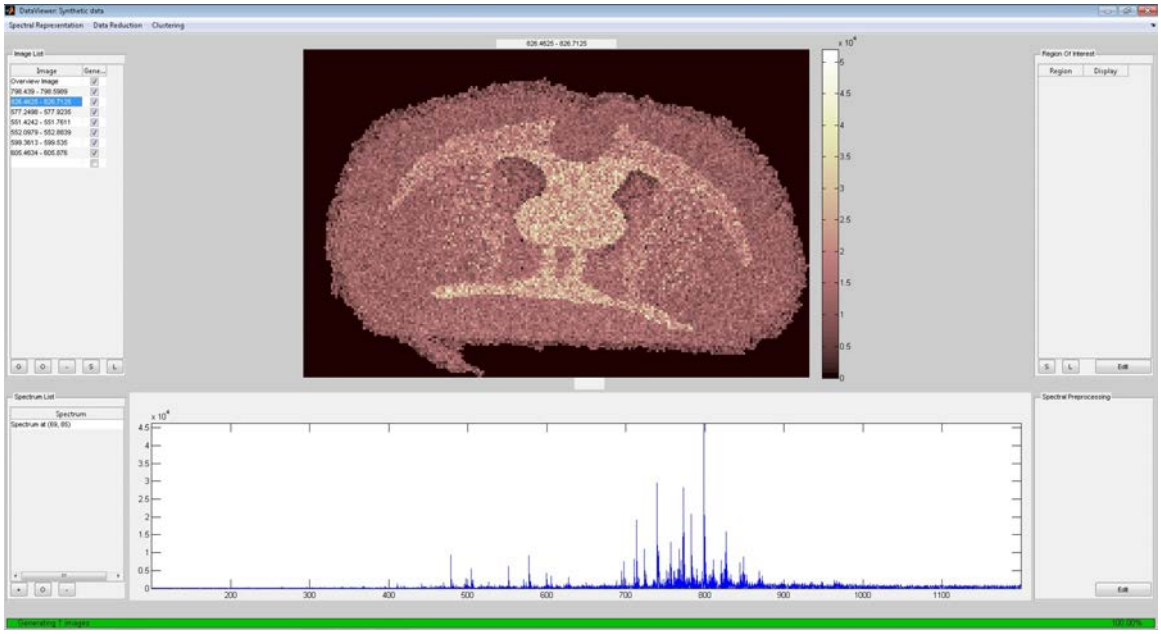

Figure 5.45: Example of synthetic dataset loaded into SpectralAnalysis software allowing any standard processing workflows to be preformed.

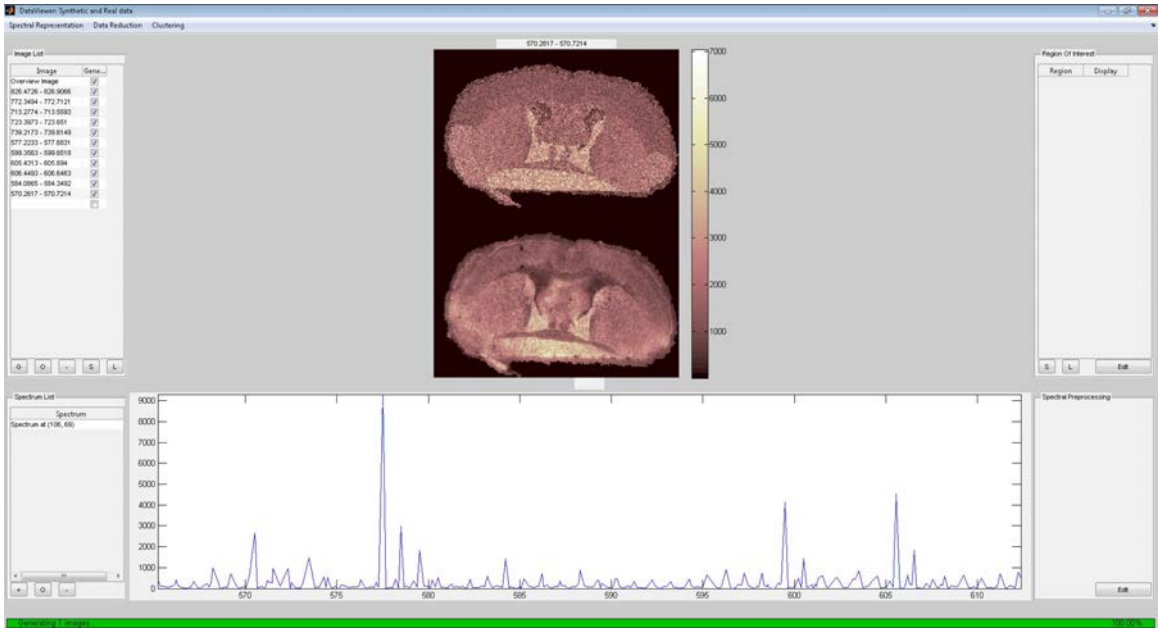

Figure 5.46: Example of a combined synthetic and real dataset loaded into SpectralAnalysis software. This allows an interactive comparison of these two datasets, as well as performance of any pre and post processing methods on these data.

5.3.2 Simulation of instrumental parameters

The generation of synthetic MSI data using statistical modelling allows users to generate datasets that contain the inherent biological variance contained within MSI data, however there remains a number of additional intstrumental variables that also need to be considered. In addition, this only allows the generation of peaklist MSI data which is not representitive of the data as it is acquired from an instrument. Understanding and simulation of instrumental variables is an area that has been investigated outside of MSI, and is beginning to emerge in MS and MSI research [177, 231, 232]. Some of the instrumental variables can be modelled based on first principles such as peak broadening [231, 232]. Other spectral features such as baselines, or artefacts of laser instability such as seen in chapter 2 are much more complex and can only be simulated heuristically. By simulating these effects on data generated via statistical modelling, different preprocessing workflows could be evaluated to determine optimal methods to perform this. These simulations have

been interfaced with the SpectralAnalysis software package [191], as alternatives to pre and post processing options (Figure 5.47). This allows users to perform these simulations on any dataset, and so a comparison of the expected spectra from a number of different instruments or configurations could be achieved.

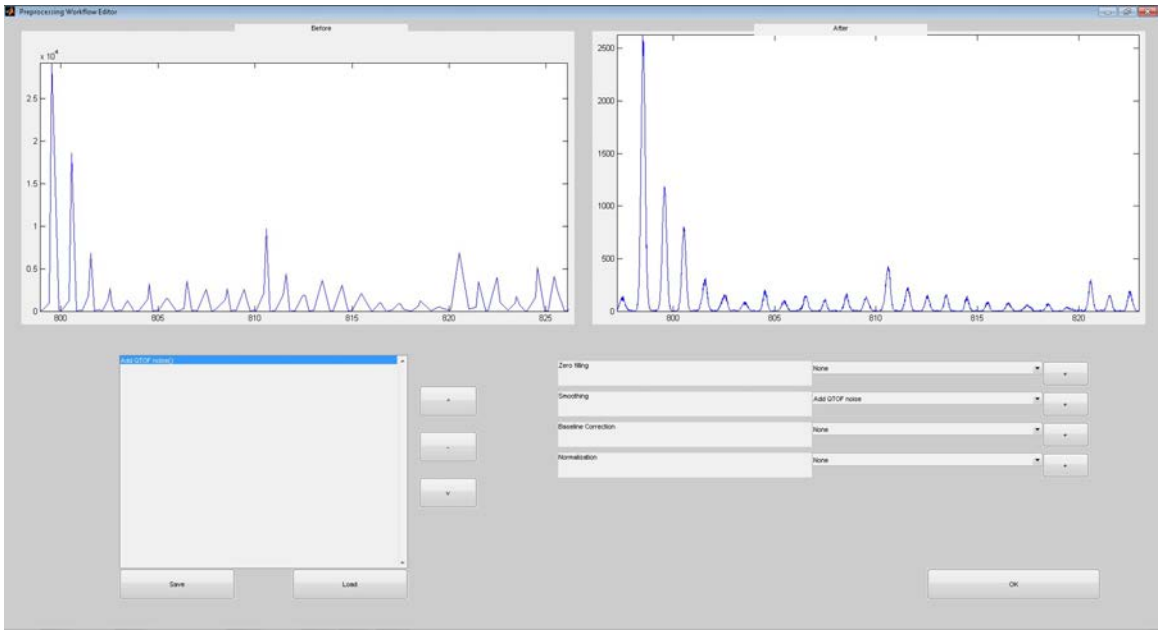

Figure 5.47: Example of Q-TOF peak broadening included into SpectralAnalysis software.

Peak broadening

Peak broadening in MS can arise from a number of sources depending on the instrument, and ionisation methods used. MALDI-TOF MS peak broadening occurs from a combination of initial velocity spread of the ions generated which will be Maxwell-Boltzmann in nature and statistically derived noise from the detector which will be governed by a Poisson distribution [231]. The velocity based distribution will contribute to around 90% of the peak broadening observed in [231]. These peak broadening simulations were performed on a synthetic spectrum from the corpus callosum region of the coronal brain dataset described previously (Figure 5.48). To simulate this effect, each ion count in the data is considered, and the m/z values are shifted based on sampling of random numbers from Maxwell-Boltzmann and Poisson distributions with the peaks centred on the orig-

inal peak of the data. Once this has been performed on all ion counts within the data, the whole spectrum is then rebinned based on either a desired bin width, or resolution (Figures 5.49 to 5.51, and Algorithm 11). To compare this simulation with real data, a mean spectum from the Rapiflex TOF dataset from chapter 6 was peak picked and converted back into a TOF spectrum with a resolution of 10,000. Simulating peaks in this way produces data that is visually comparable with real MALDI-TOF data (Figure 5.52). Peak broadening in quadrupole time-of-flight (Q-TOF) instruments occurs from inhomogeneous accelaration fields, and charge repulsion of ion packets. As a result the peak shapes in these instruments is approximately Gaussian in shape. The same basis of simulation to the TOF instruments can therefore be performed for Q-TOF instruments (Figures 5.53 to 5.55).

input : Spectral channels array S

input : Intensities array I

input : Resolution r

input : Bin width b

input : TOF or QTOF simulation

output: Peak braodened spectral channels

output: Peak braodened spectral intensities

for $i = 1 \leftarrow length(S)$ **do**
  if *TOF simulation* **then**
    create I_i Maxwell-Boltzmann distributed random numbers m;
    create I_i poisson distributed random numbers p;
    mean centre random numbers m and p and divide p by 10;
    add mean centred data to give peak distribution d;
  else
    create I_i Gaussian distributed random numbers d;
  end
  Scale the peak distribution d by the full width half maxima $2\sigma\sqrt{2\ln 2}$;
  Generate resolution scale factor as $\frac{S_i}{r}$;
  Multiply the distributed peak by the resolution scaling to give the resolved peak distribution n;
  Rebin n at a specified width b;
end

Algorithm 11: Algorithm for peak broadening simulation

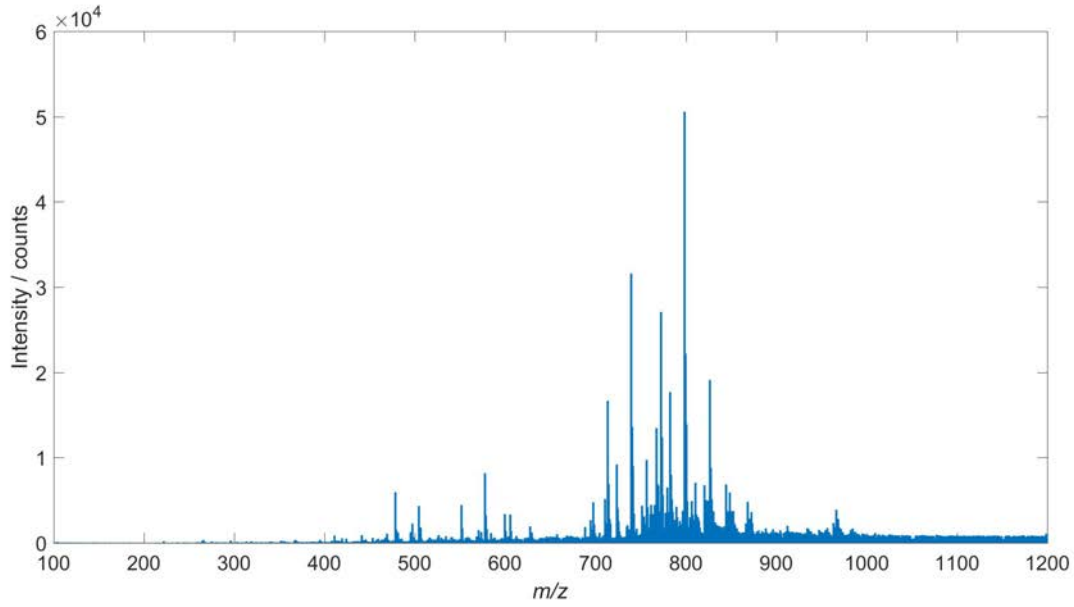

Figure 5.48: Mean spectrum from corpus callosum region of the coroanl brain data used to demonstrate the peak broadening simulations.

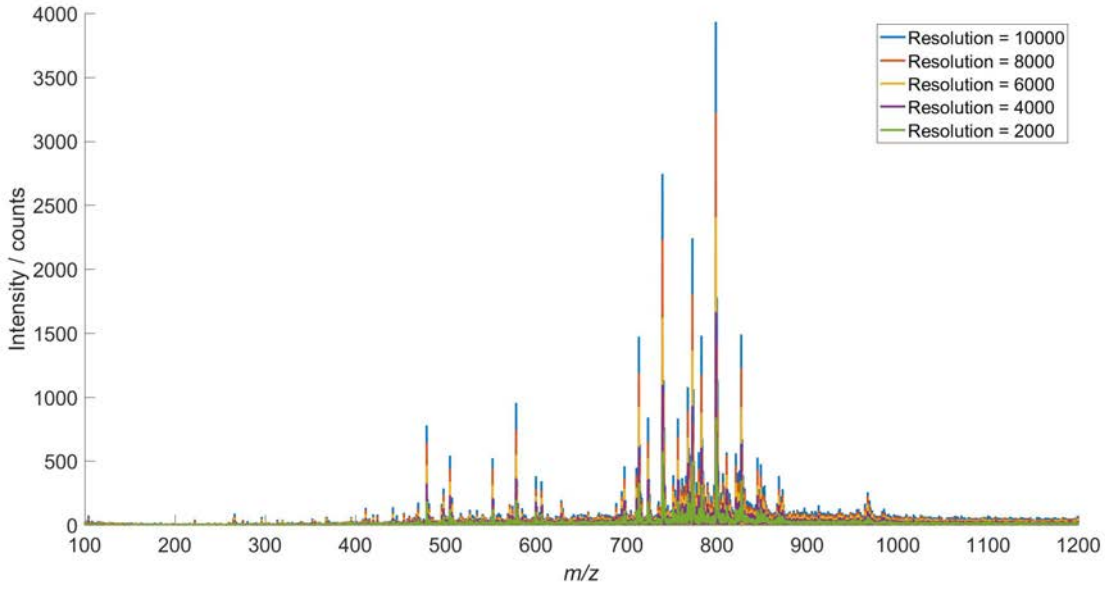

Figure 5.49: Simulation of TOF spectra at different resolutions showing the effect of resolution on peak widths and shape.

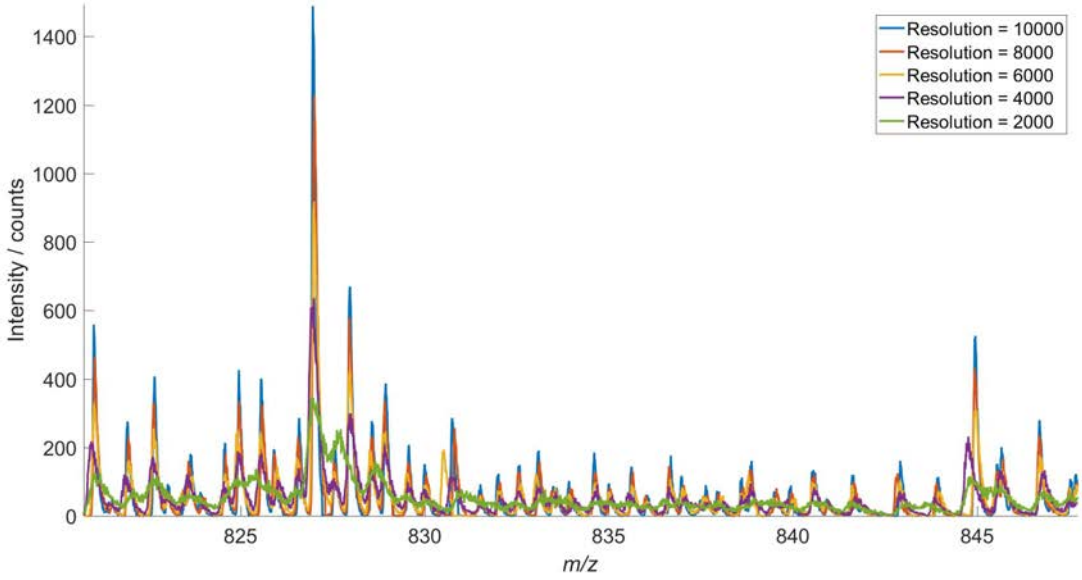

Figure 5.50: Simulation of TOF spectra at different mass resolutions: Spectral region m/z 820 to 850. As well as a peak broadening effect, a decrease in peak height is seen at lower mass resolution.

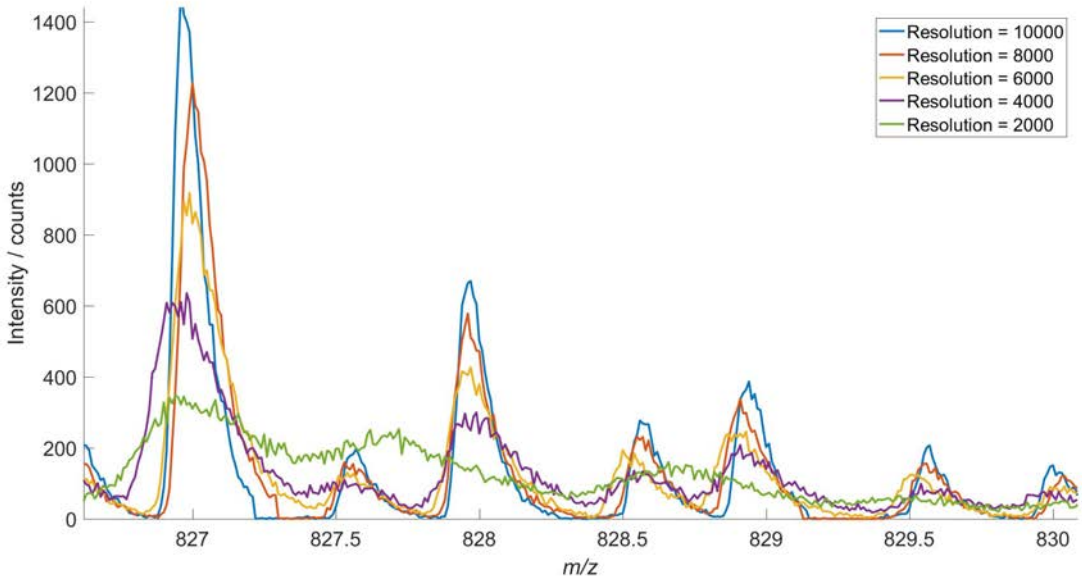

Figure 5.51: Simulation of TOF spectra at different mass resolutions: Spectral region m/z 826.5 to 830. As well as the peak height decrease, and peak broadening, at lower resolution the peaks can be seen to become less smooth.

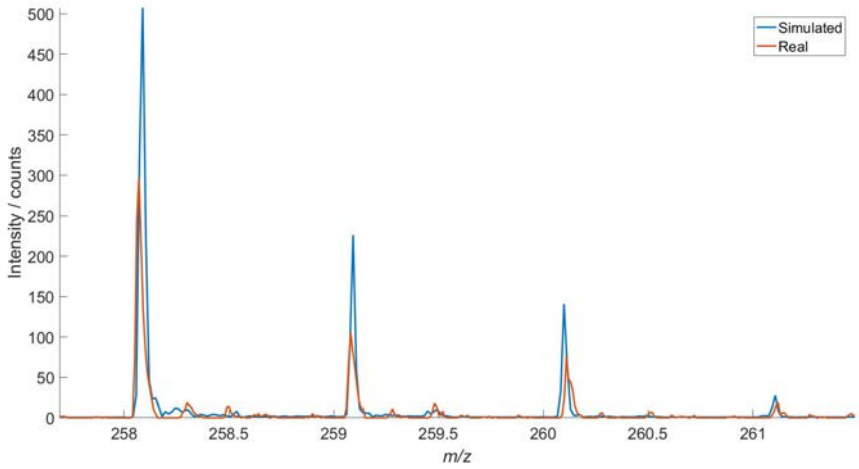

Figure 5.52: Comparison of real TOF spectrum to simulated TOF peaks. The peaks shape of the simulated data is consistent with that of the real TOF data.

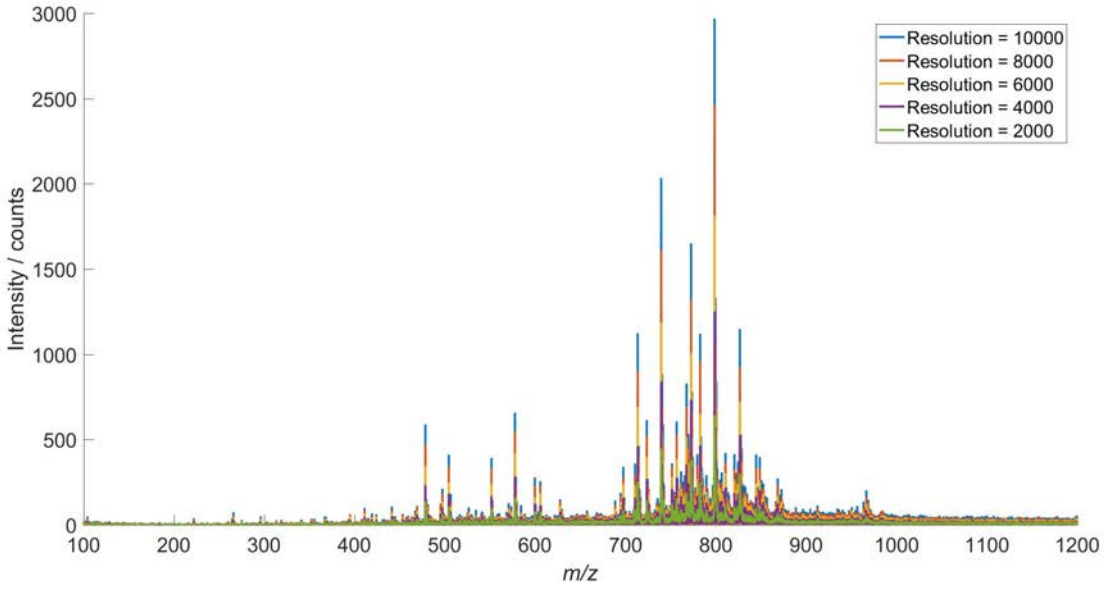

Figure 5.53: Simulation of Q-TOF spectra at different mass resolutions. As with the Q-TOF peak broadening, a decrease in peak height is seen at lower mass resolution.

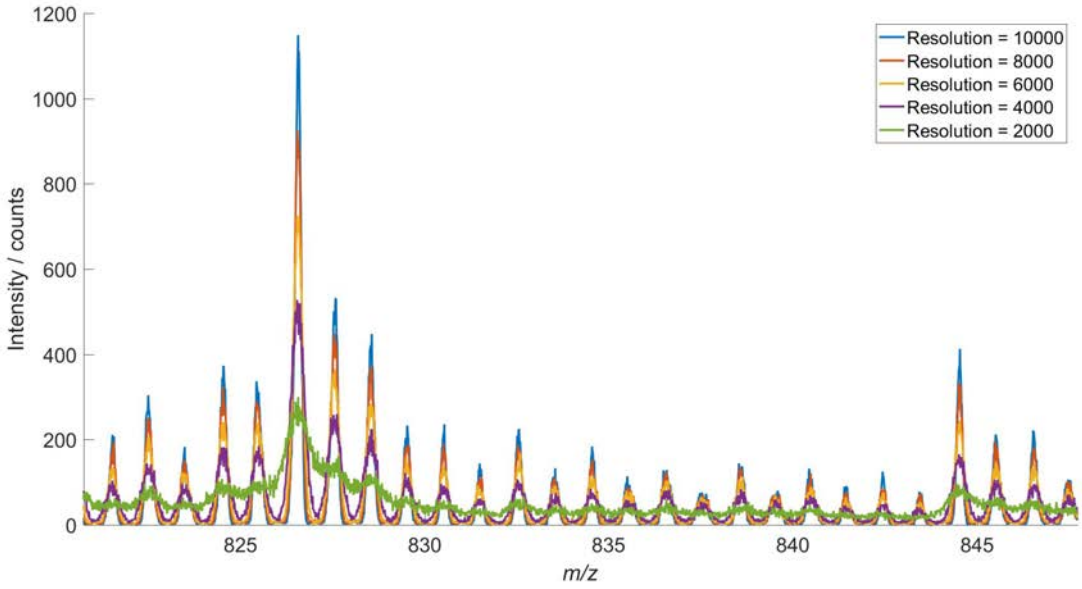

Figure 5.54: Simulation of Q-TOF spectra at different mass resolutions: Spectral region m/z 820 to 850.

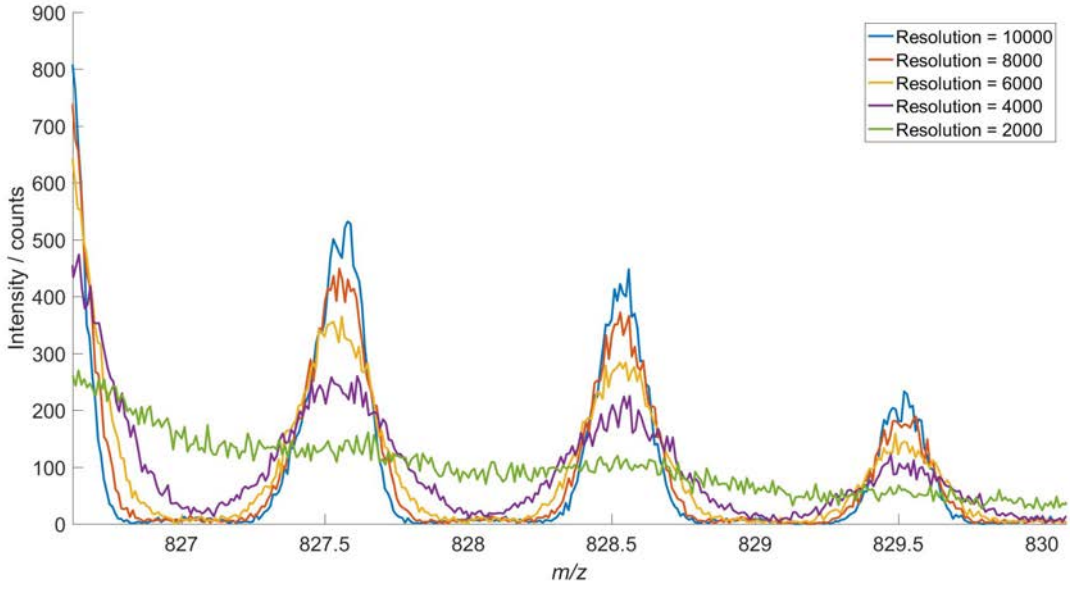

Figure 5.55: Simulation of Q-TOF spectra at different mass resolutions: Spectral region m/z 826.5 to 830.

These peak broadening simulations can be used to evaluate preprocessing such as smoothing and peak picking algorithms. To demonstrate this, the peak picked mean

spectrum from the corpus callosum brain regions was used as a reference spectrum (Figure 5.48). This peak picked spectrum was then converted back into a Q-TOF spectrum with a resolution of 10,000 (Figure 5.55), and gradient based peak picking was applied either prior to or post smoothing using a Savitzky-Golay filter with a window size of 11 and second order polynomial. To make these data directly comparible, the three spectra (original, peak picked unsmoothed and peak picked smoothed) were all rebinned to 0.01Da axes. These spectra can then be compared by computing their similarities using the metrics described in figure 1.11. This shows that smoothing prior to peak picking produces a higher similarity to original spectrum. While this result is unsurprising, this demonstrates the capability of these simulated spectra to quantitatively evaluate preprocessing workflows in MSI.

Data	Euclidean	Cosine	Correlation
Unsmoothed	90807	0.4658	0.4672
Smoothed	60532	0.1860	0.1865

Table 5.2: Similarities between the original spectrum and the simulated peak broadening, followed by peak picking with or without smoothing applied.

Peak shifting

Peak shifts in MS can be either randomly induced by fluctuations in laboratory conditions, or topographically based on uneven samples or titled sample slides [102]. To simulate random peak shifting, a random number based shift between user defined thresholds can be applied to each peak m/z (Figure 5.56, and Algorithm 12). For topographically induced changes from tilted samples, a user defined systematic increase or decrease in peak m/z can be applied across a given direction, or in a given spatial pattern (Algorithm 13). These simulated peak shifts can be used to evaluate the performace of peak alignment algorithms.

input : Spectral channels array S

input : Spectral drift range d

output: Peak shifted spectral channels P

for $i = 1 \leftarrow length(S)$ **do**

    Create random shift r within spectral drift range d;

    Add r to S_i to create shifted spectral channel P_i;

end

Algorithm 12: Algorithm for random peak shifting simulation

input : Spectral channels array S

input : Spectral drift maxima m

input : Direction of spectral drift d

input : Shift decrease or increase

input : Pixel x and y co-ordinates x and y

input : Image size z in pixels x by y

output: Peak shifted spectral channels P

for $i \leftarrow 1 length(S)$ **do**
  if $d = left\ to\ right$ **then**
    shift $s = m \times (\frac{x}{z_1})$;
  else if $d = right\ to\ left$ **then**
    shift $s = m \times (\frac{z_1 - x + 1}{z_1})$;
  else if $d = top\ to\ bottom$ **then**
    shift $s = m \times (\frac{y}{z_2})$;
  else
    shift $s = m \times (\frac{z_2 - y + 1}{z_2})$;
  end
  if *shift increase* **then**
    add shift s to S_i to create P_i;
  else
    subtract shift s from S_i to create P_i;
  end
end

Algorithm 13: Algorithm for systematic peak shifting simulation

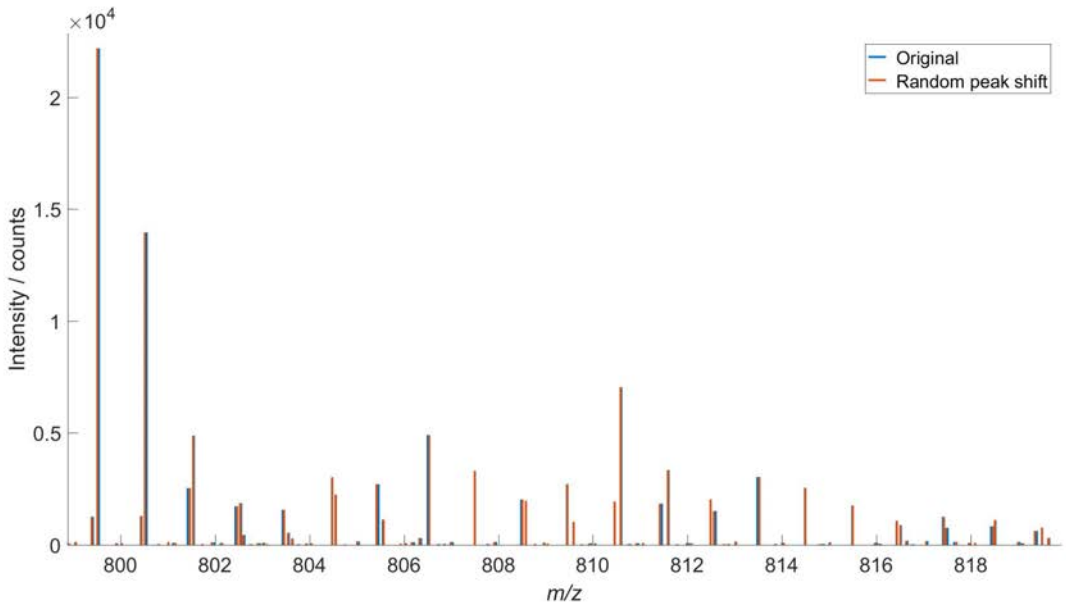

Figure 5.56: Example of random peak shifting applied to the a mass spectrum from figure 5.48.

Baselines

MS baselines occur due to unresolved high m/z species such as matrix clusters and large macromolecules. In order to truly simulate these from first principles, a full understanding of both the underlying sample composition and MS fundamentals would be required. Heuristically, baselines can be simulated as either broad MS peaks, approximately Gaussian shaped (Figure 5.57 C), or as a progressive decrease (Figure 5.57 B and Algorithm 14). These baseline simulations can be used to evaluate algorithms for baseline correction, particularly when applied in conjuction with other simulations such as peak broadening and noise.

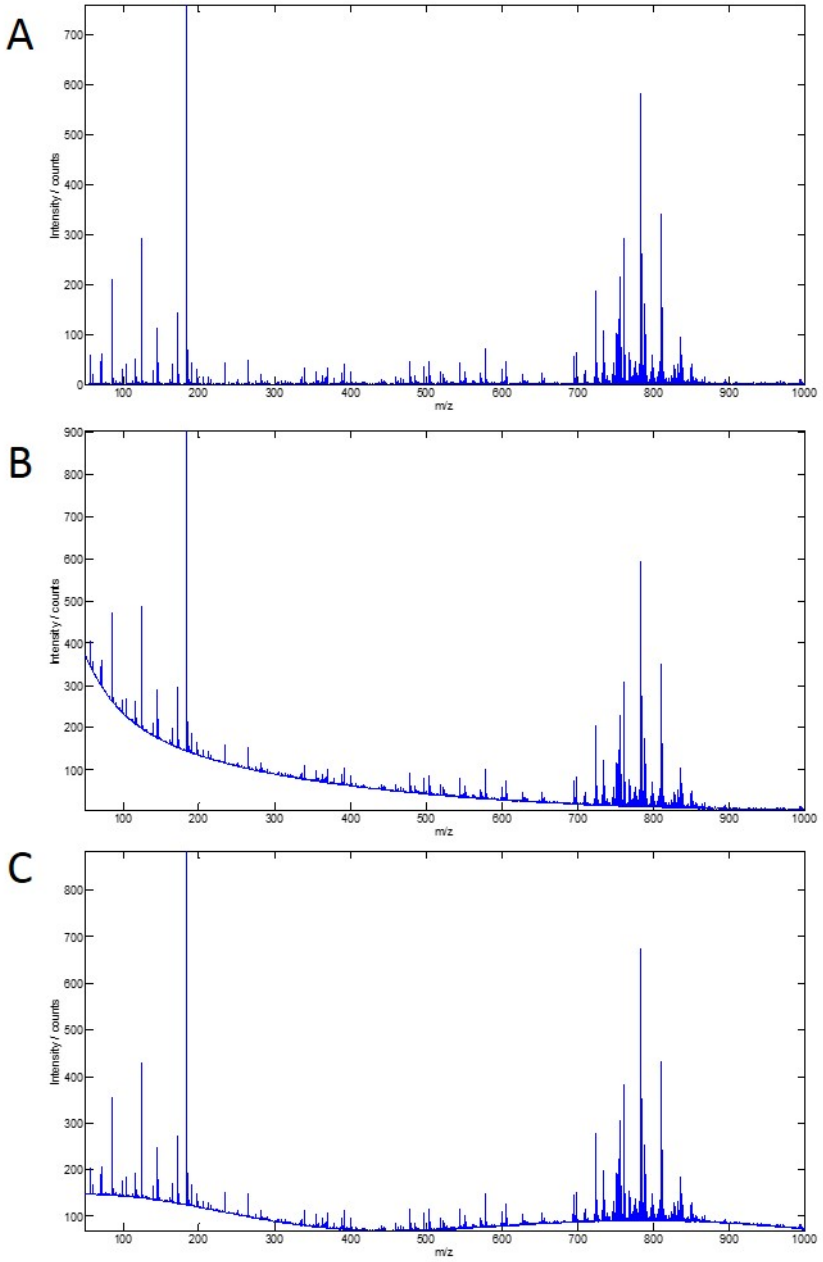

Figure 5.57: Example of baselines, progressively decreasing baseline (B), and random broad peaks (C) applied to a mass spectrum (A).

input : Spectral channel array S

input : Intensity array I

input : number of Gaussians g

input : Intensity scale of baseline b

input : Baseline type

output: Baselined intensity array B

```
Initially set up baselined intensity array as intensity array I;
Generate g random numbers within spectral channel range for Gaussian widths σ;
if Progressively decreasing baseline then
    Set Guassian means to S_1;
else
    Randomly select g Guassian means μ within spectral channel range;
end
for i = 1 ← g do
    Create Gaussian distribution G centred on μ_i with standard deviation σ_i;
    Add G to baselined intensity array B;
end
```

Algorithm 14: Algorithm for addition ofbaselines to mass spectrum

Noise

There are different types of noise in MSI, spectral white noise derived from electrical sources, or image shot noise, often referred to as salt and pepper noise in standard image processing, which may be derived from inhomogeneous matrix deposition or crystal formation. Electrical noise can be simulated heuristically by the addition of a small intensity fluctuations within a user specified threshold (Algorithm 15), and provide a means to further test smoothing algorithms. Shot noise can be simulated heuristically by randomly selecting a given number of pixels to either increase or decrease in intensity by either a

preset or randomly generated value (Algorithm 16). This was demostrated on a previously published rat brain dataset [10] (Figure 5.58), and can be used to test normalisation strategies or spatially aware denoising algorithms such as that proposed by Alexandrov *et al.* [121].

input : Intensity array I

input : Noise range n

output: Noisy intensity array N

for $i = 1 \leftarrow length(I)$ **do**
    Create random noise r as random number within range n;
    Add r to I_i to create noisy ntensity N_i;
end

Algorithm 15: Algorithm for addition of white noise to spectrum

input : Intensities matrix $\underline{I}$

input : Shot noise frequency f

input : shot noise increase magnitude m_1 and decrease m_2

input : ratio of increase to decrease r

output: Noisy intensity matrix $\underline{N}$

Randomly select f pixels to apply noise to;
Randomly partition these pixels according to the ratio of increase to decrease r;
for $i = 1 \leftarrow f$ **do**
    if *intensity increase* **then**
        Mulitply intensities of spectra $\underline{I}_i$ by m_1;
    else
        Divide intensities of spectra $\underline{I}_i$ by m_2;
    end
end

Algorithm 16: Algorithm for addition of shot noise to image

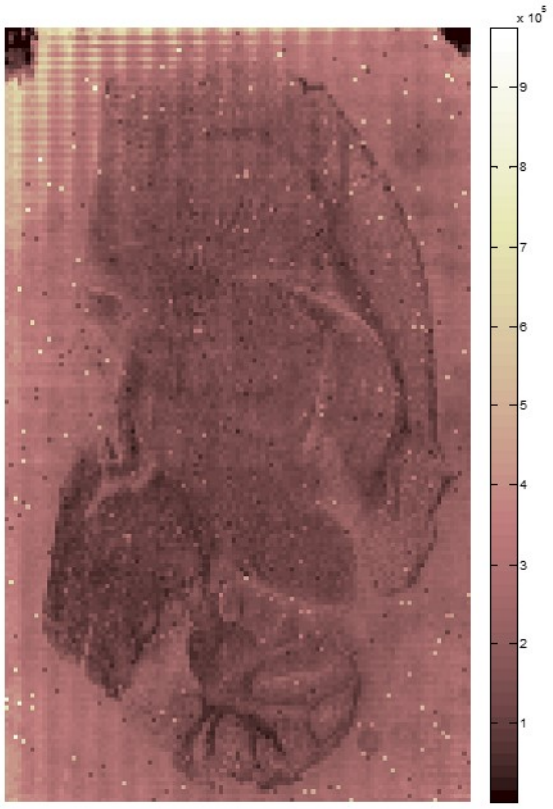

Figure 5.58: TIC from rat brain data with shot noise applied.

Intensity gradients

Intensity gradients in MSI can occur from charging effects, either on the sample if not conductive, or on quadrupoles if employed in the instrument. Typically, these will be observed as a progressive decrease in intensity throughout the course of an experiment as charge builds up (Figure 5.59, and Algorithm 17). As with the simulation of shot noise, the simulation of these intensity gradients can be used to quantitatively evaluate the performance of different normalisation strategies. A gradient is also observed from laser instability at high repetition rates seen in chapter 2, however this is not purely an intensity effect and thus a more complex model would be required to simulate this.

input : Intensity array I

input : Intensity shift percentage maxima m

input : Direction of intensity shift d

input : Shift decrease or increase

input : Pixel x and y co-ordinates x and y

input : Image size z in pixels x by y

output: Shifted intensity$P$

for $i = 1 \leftarrow length(S)$ **do**
    if $d = left\ to\ right$ **then**
        shift $s = m \times (\frac{x}{z_1})$;
    else if $d = right\ to\ left$ **then**
        shift $s = m \times (\frac{z_1 - x + 1}{z_1})$;
    else if $d = top\ to\ bottom$ **then**
        shift $s = m \times (\frac{y}{z_2})$;
    else
        shift $s = m \times (\frac{z_2 - y + 1}{z_2})$;
    end
    if *shift increase* **then**
        multiply shift s by S_i to create P_i;
    else
        divide S_i by shift s to create P_i;
    end
end

Algorithm 17: Algorithm for intensity gradient simulation

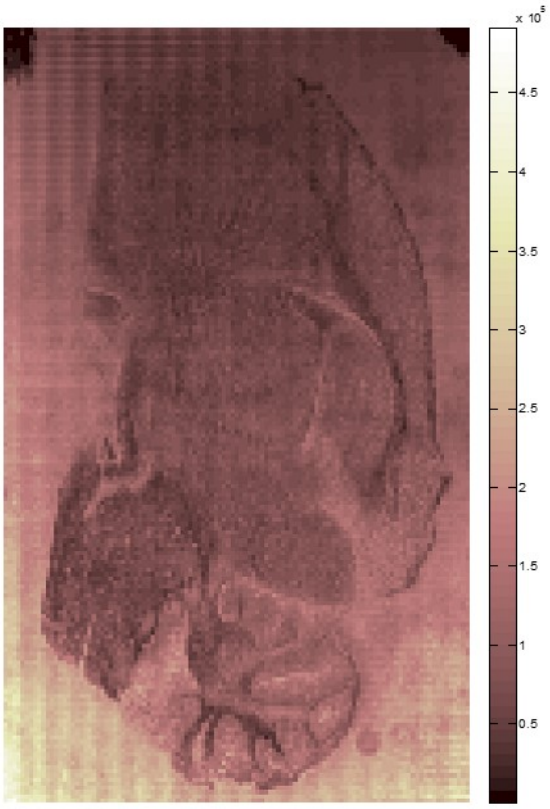

Figure 5.59: TIC from rat brain data with a twofold intensity decrease gradient applied bottom to top.

Combination of simulations

In practise, these simulations will generally be combined together in a similar way that multiple preprocessing steps will be performed sequentially. Intensity and peak shifting effects should ideally be performed first, on the peaklist data, followed by peak broadening, the addition of baselines, and finally the addition of any electrical noise into the data. By builing up these different simulations, on a synthetic dataset generated by statistical modelling, a dataset that reflects both the biological variation and complexity, along with the instrumental variability can be generated, with full knowledge of the processing that has been performed. This can allow an unbiased and quantitative evaluation of pre and post processing algorithms to be performed in a wide variety of scenarios.

5.4 Conclusions

As described in chapter 4, where possible, external evaluation methods should be used when comparing novel algorithms or parameters, using a ground truth that is representative of samples of interest. We have demonstrated that synthetic data generated by statistical modelling is a suitable means to achieve this. In addition, this approach, in combination with random projection allows large datasets both in number of pixels and mass channels, to be generated rapidly allowing evaluation and comparison of both existing and new methods as the data increases in size. Additional artefact and peak shape simulation can then be performed on these peak lists generated using this approach by modelling them on a first principals basis. Improvements could be made in synthetic spectra generation by attempting to fit a more accurate model to MSI data, such as one which does not include negatives, a feature that doesn't exist in MS. However model testing and parameter determination in high dimensions is challenging at best. Other instrumental parameters could also be simulated, and a more principaled means to generate baselines could be investigated. Ultimatelty these simulated data can then provide a means to evaluate data processing and analysis pipelines in a more systematic and quantitative manner, in order to determine the optimal methods to analyse data in MSI experiments.

CHAPTER 6

NOVEL CLUSTERING ALGORITHMS IN MSI

6.1 Introduction

As outlined in chapter 5, clustering is a powerful tool to group together similar spectra of an MSI dataset. Clustering algorithms have been shown to be capable of segmenting different anatomies within tissue [131], differentiating tumour vs. non tumour tissue types [142], and even distinguishing intra-tumour heterogeneity [152] in MSI data. There are however a number of challenges involved in clustering of MSI data. MSI data inhabits very high dimensional space, resulting in a very sparse population of data within this space. Consequently, Euclidean distance metrics converge towards ∞ as the number of dimensions increases. This is referred to as the curse of dimensionality [233]. For simple applications such as segmentation of tissue from matrix, or of highly differentiated tissues, simple clustering algorithms such as k-means and hierarchical clustering are sufficient [153]. However when a larger number of fairly similar anatomies need to be segmented, the performance of these algorithms decreases significantly. This can be overcome by the use of more sophisticated algorithms but this comes with increased computational cost, and may require additional dimensionality reduction steps [120]. The high dimensionality of MSI data also means that these datasets are extremely large, often limiting the statistical analyses that can be performed either in terms of computational complexity in both space and time. For example, agglomerative hierarchical clustering requires a full pairwise

distance matrix of the data to be calculated resulting in a time complexity of $O(n^2.d)$ and a space complexity of $n^2 + n \times d$, where n is the number of pixels, and d the dimensions or mass channels of the data. Additional to this step, the linkage method used to agglomerate the data will also further increase the overall complexity, resulting in a generalised time complexity of $O(n^2 log(n))$ [234]. This has been seen previously to be a limiting factor in clustering of MSI data where although the Ward's linkage has been previously shown to produce optimal segmentation [142], Alexandrov *et al.* were constrained to using the complete linkage by the available RAM [120]. With new developments in instrumentation and a drive towards 3D imaging, the number of pixels in MSI datasets is increasing dramatically. While this does not currently limit routine analysis, it is likely that the number of pixels will become a limiting factor for clustering MSI data in the near future. To further add to the challenge, MSI data suffers from a high degree of pixel to pixel variability, arising from a number of different factors such as sample preparation [12] and instrumental variations [150].

To achieve more widespread uptake, clustering algorithms that can efficiently and accurately cluster large and complex MSI datasets are required. This chapter uses the methods described in chapter 5 to evaluate the accuracy of clustering algorithms previously used in MSI and those used in other fields. This is done with a view towards emerging large datasets, and following this, a novel algorithm based on the combination of two-phase and graph based clustering is proposed that is both accurate and efficient.

6.2 Experimental

Synthetic data

For the comparison of clustering accuracy with respect to number of pixels, synthetic datasets were generated using the statistical modelling described in chapter 5, using three anatomical regions from the reference data (brain stem, lateral septal complex, and isocortex), and 3000 to 30,000 pixels in increments of 3000 pixels. These datasets were then

used to compare the accuracy of the clustering algorithms described in section 1.4.4. In the case of k-means, bisecting k-means, and two-phase k-means clustering, three replicates were used to minimize the effects of the random starting positions. FLAME, and DBSCAN were performed using a number of different nearest neighbours (20 to 200 in increments of 20), and the most accurate result was plotted. DBSCAN was also performed using a range of epsilon values from 0.1 to 0.9 in increments of 0.1. Graph cuts and two-phase graph cuts were performed using a number of different eigenvectors (20 to 200 in increments of 20 for graph cuts, and 5 to 25 in increments of 5 for two phase graph cuts), and the most accurate result plotted.

A large synthetic brain MSI image was generated based on the original masks from chapter 5 scaled up by a factor of 3 in x and y to give a total of 187,425 pixels. This represents the size of the data, had it been acquired at 15 μ m pixels, or if nine serial sections were acquired.

Experimental data

Transverse brain sections (12 μm) were thaw mounted onto glass slides before matrix coating (CHCA, 5 mg/mL, 80% methanol, 0.1% TFA) using an automated pneumatic sprayer (TM-sprayer, HTX imaging, Chapel Hill, NC, USA). Images were acquired using a Synapt G2Si (Waters, Manchester, UK), with a pixel size of 30 μm in both x and y, and an m/z range of 100-1200 Da.

Mouse colon data were obtained externally through collaboration with AstraZeneca. Breifly, mouse colons were collected, prepared using the Swiss Roll technique [235] and embedded in 2.5 % carboxymethyl cellulose (Sigma-Aldrich) in sterile water. Approval for this animal study was given prior to initiation by an internal University of Glasgow ethics committee and by the U.K. Home Office. The embedded colons were snap frozen in a slurry of ethanol and crushed dry ice, and then stored at -80 °C. The colons were cut to 10 μm sections in a cryostat microtome at -18 °C. The sections were cut in a specific order to take 3 sections for histology and 2 sections for MSI. Sections for MSI were

thaw mounted on conductive indium tin oxide (ITO) coated slides (Bruker Daltonics, Germany) and sections for histology onto normal microscope slides. Slides were stored at -80 °C until imaging or staining. Tissue sections thaw mounted onto ITO slide for imaging were dried in a stream of nitrogen when removed from -80 °C storage. Optical images were taken using a standard flatbed scanner (Seiko Epson, Negano, Japan) prior to sample preparation and MALDI matrix application. Sections were treated with 2,4-diphenyl-pyranlium tetrafluoroborate (DPP-TFB) to derivitise endogenous primary amines as previously described by Shariatgorji *et al.* [236]. Briefly, DPP-TFB, 9.6 mg was dissolved in 1.2 ml of 100% methanol and sonicated for 10 min and 3.5 μl of trimethylamine was added to 6 ml of 70% methanol. The DPP-TFB solution was gradually added to the 70% methanol and this final solution was sprayed onto slides using an automatic matrix sprayer (TM-Sprayer, HTX Technologies) at 0.08 mL/min, 80 °C with 30 passes. The slide was incubated in a chamber with 50% methanol for 15 min, and dried every 5 min under a nitrogen stream. Derivatisation was performed to target endogenous metabolites out with the scope of this experiment. High spatial resolution analysis was acquired using a RapiFlex MALDI TOF/TOF (Bruker Daltonics, Germany) in reflectron positive ion mode, using a pixel size of 5 μm in both x and y, over an m/z range of 200-1000 Da.

Two phase graph cuts algorithm

The two phase graph cuts algorithm used is described as follows.

input : data matrix of n pixels by d mass channels

input : desired number of subets s

input : number of eigenvectors to cluster the subsets e_s and compression set e_c

input : K-means clustering parameters

output: Vector length n of cluster assignment

```
Randomly assign the data into s equal sized subsets S_{1-s};
for i ← 1 to s do
    load into RAM data from subset S_i;
    cluster S_i into k clusters using the graph cuts algorithm 5 using the top e_s
     eigenvectors;
    if i == 1 then
        create a compression set, a compressed representation of the subset S_i from
         the cluster centroids;
    else
        add cluster centroids of the subset S_i to the compression set;
    end
end
cluster the compression set using the graph cuts algorithm 5 using the top e_c
 eigenvectors;
assign the cluster identities from the compression set to the pixels from subsets;
```

Algorithm 18: Two-phase graph cuts clustering algorithm

Data processing and analysis

Data processing was performed on an Intel Xeon quad core CPU E5-2637 v2 (3.50 GHz) with 64 GB of RAM. All data were converted from proprietary format to the mzML format using msconvert as part of ProteoWizard [197] software then into imzML using imzML-Converter [118]. This was then imported into MATLAB (version R2014a and statistics

toolbox, The Math-Works, Inc., Natick, MA, USA) using the SpectralAnalysis software package [191]. Data acquired on the QSTAR were zero-filled using the QSTARZeroFilling in SpectralAnalysis, followed by three iterations of Savitzky-Golay smoothing with a window size of 7 and second order polynomial, and the negative signals produced by the smoothing were then removed. The data acquired on the Synapt were zero filled using the interpolated rebinning from SpectralAnalysis with a bin size of 0.01 and no smoothing was applied. Total spectra were then generated from each dataset, and these were peak picked using the gradient method in SpectralAnalysis, and the peak intensities were extracted for individual pixels. k-means clustering was performed using the function kmeans from the Matlab Statistics toolbox using the parameters specified in the upcoming experiments, with three replicates and random starting clusters.

Data partitioning for the different subsets in two-phase k-means and two-phase graph cuts clustering was performed by pseudo-random assignment of each pixel into a predefined number of subsets such that the subsets are of equal size, and no spectrum is not assigned a subset. The subset sizes were 17,000 spectra for the synthetic data, 25,000 spectra for the transverse brain data, and 19,000 for the gut data. In all cases, the cosine similarity measure was used for weighted graph construction and k-means clustering.

When clustering the smaller datasets using graph cuts, k-means clustering was performed on the smallest 250 eigenvectors of the connectivity graph. For the two-phase graph cuts of the large synthetic brain data, k-means was performed on the smallest 500 eigenvectors of the connectivity graph of the subsets, and the smallest 20 eigenvectors of the graph of the compression set. For the two-phase graph cuts of the transverse brain data, k-means was performed on the smallest 600 eigenvectors of the connectivity graph of the subsets, and the smallest 10 eigenvectors of the graph of the compression set. In the two-phase graph cuts of the mouse colon data, k-means was performed on the smallest 600 eigenvectors of the connectivity graph of the subsets, and the smallest 10 eigenvectors of the graph of the compression set. For a summary of these parameters see table 6.1.

Dataset	Subset eigenvectors	Compression set eigenvectors	Subset pixels
Large synthetic brain	500	20	17000
Transverse brain	600	10	25000
Mouse colon	600	20	19000

Table 6.1: Parameters used in the two phase graph cuts clustering on the larger individual datasets

6.3 Results and Discussion

6.3.1 Comparison of clustering on synthetic MSI data

To evaluate the accuracy of these algorithms with respect to the number of pixels in the data, the clustering algorithms described in section 1.4.4 were applied to synthetic data with three anatomical regions (brain stem, lateral septal complex, and isocortex) with 3000 to 30000 pixels. These results were evaluated using the Jaccard index as this was deemed most appropriate in section 4.3.1 (Figure 6.1). These results show the commonly used algorithms in MSI do not accurately cluster these data, and in most cases perform similarly to one another. Additionally, many of the other algorithms used in other fields such as DBSCAN and affinity propogation perform significantly worse than these. The graph cuts algorithm was found to very accurately cluster these data and thus is of interest to investigate further. There are a number of other variables that could affect clustering accuracy, and these will be discussed further in the future work.

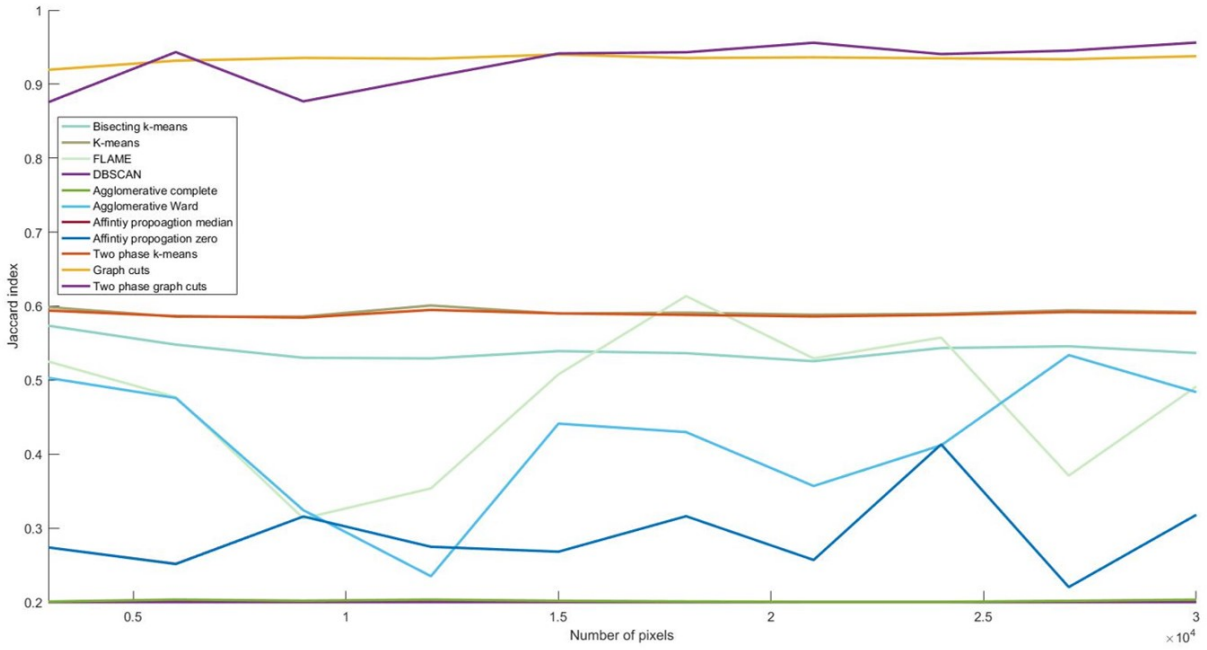

Figure 6.1: Comparison of the clustering algorithms described in section 1.4.4 performed on synthetic data created by modelling as multivariate normal distribution from chapter 5. This shows much more accurate segmentation when using the graph based clustering approaches as compared to all other algorithms.

6.3.2 Graph based clustering

As discussed, the fewer variables involved in a clustering algorithm the easier it is to use by non-experts. The graph cuts algorithm used for this work aims to minimise this by performing the clustering on the top eigenvectors computed directly from the pairwise distance matrix considered as a weighted graph (Algorithm 5). This removes the need to select the number of nearest neighbours to use, as well as the variations in Laplacian formation. In doing so, only the parameters required to perform the k-means clustering, and the number of eigenvectors to be used must be supplied by the user.

By performing this graph cuts clustering algorithm on a synthetic mass spectrometry imaging dataset, with varying numbers of eigenvectors, it can be seen there is a clear optimal range in the number of eigenvectors to use (Figure 6.2). Below this, the accuracy

of the clustering shows a sharp decrease, and above this, there is an increase in the variability of the results, along with a decrease in accuracy. There are ways to determine the optimal number of eigenvectors to use, however, as with estimating the number of clusters in a dataset, these can often be difficult to obtain [237].

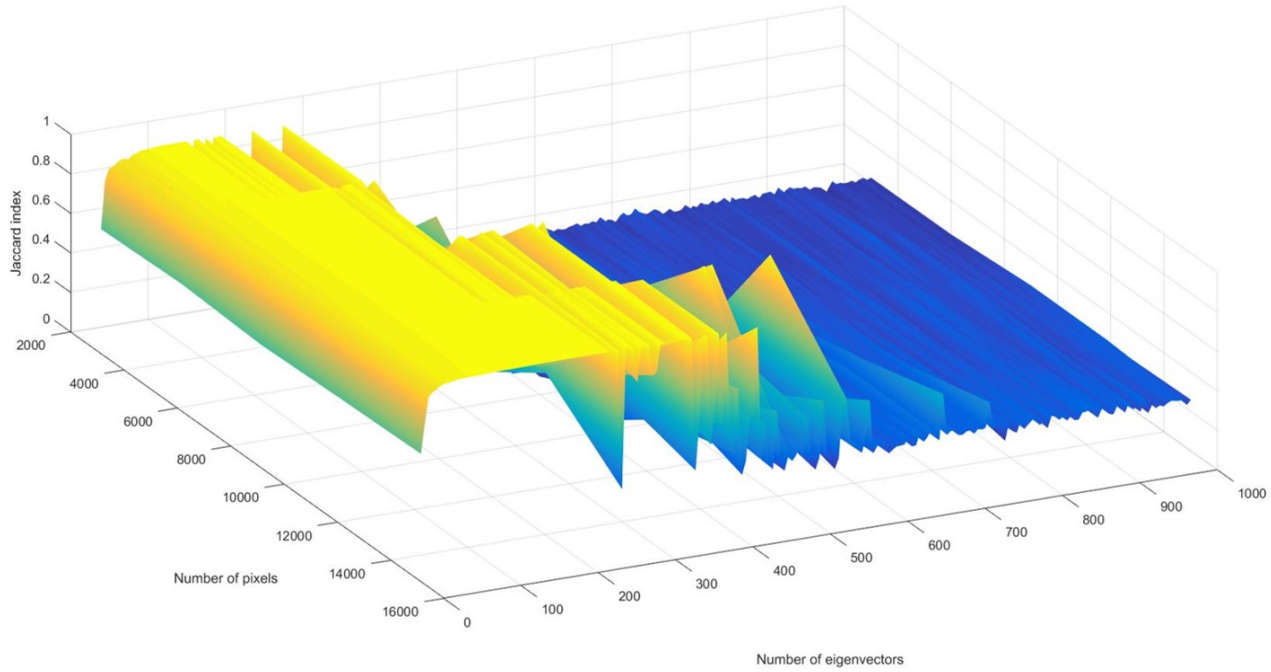

Figure 6.2: Evaluation of the accuracy of the graph cuts clustering with respect to the number of pixels and eigenvectors. There is a clear optimal region at around 100 eigenvectors, and the range of optimal eigenvectors increases as the number of pixels increases.

Graph cuts on biological data

Using graph cuts clustering on an MSI image of coronal brain shows much clearer anatomical segmentation when compared to other clustering methods such as k-means and hierarchical clustering (Figure 6.3). When clustering these data, only the graph cuts segmentation separates the isocortex (green) from olfactory areas (purple), and identifies the caudoputamen (dark blue) and brain stem (red) areas. Along with the results on synthetic MSI data, this highlights the potential usefulness of this clustering algorithm for MSI data. This improvement is also observed in clustering in an MSI image of sagittal rat

brain (Figure 6.4), where only the graph cuts algorithm is able to clearly segment caudate putamen (turquoise), cerebral cortex (orange and grey), thalamus (light green), midbrain (cream and blue) and hippocampus (purple). These initial results are not readily generalisable as it cannot be ruled out that the inherent characteristics of these datasets are more favourable to this approach and more controlled experiments are required.

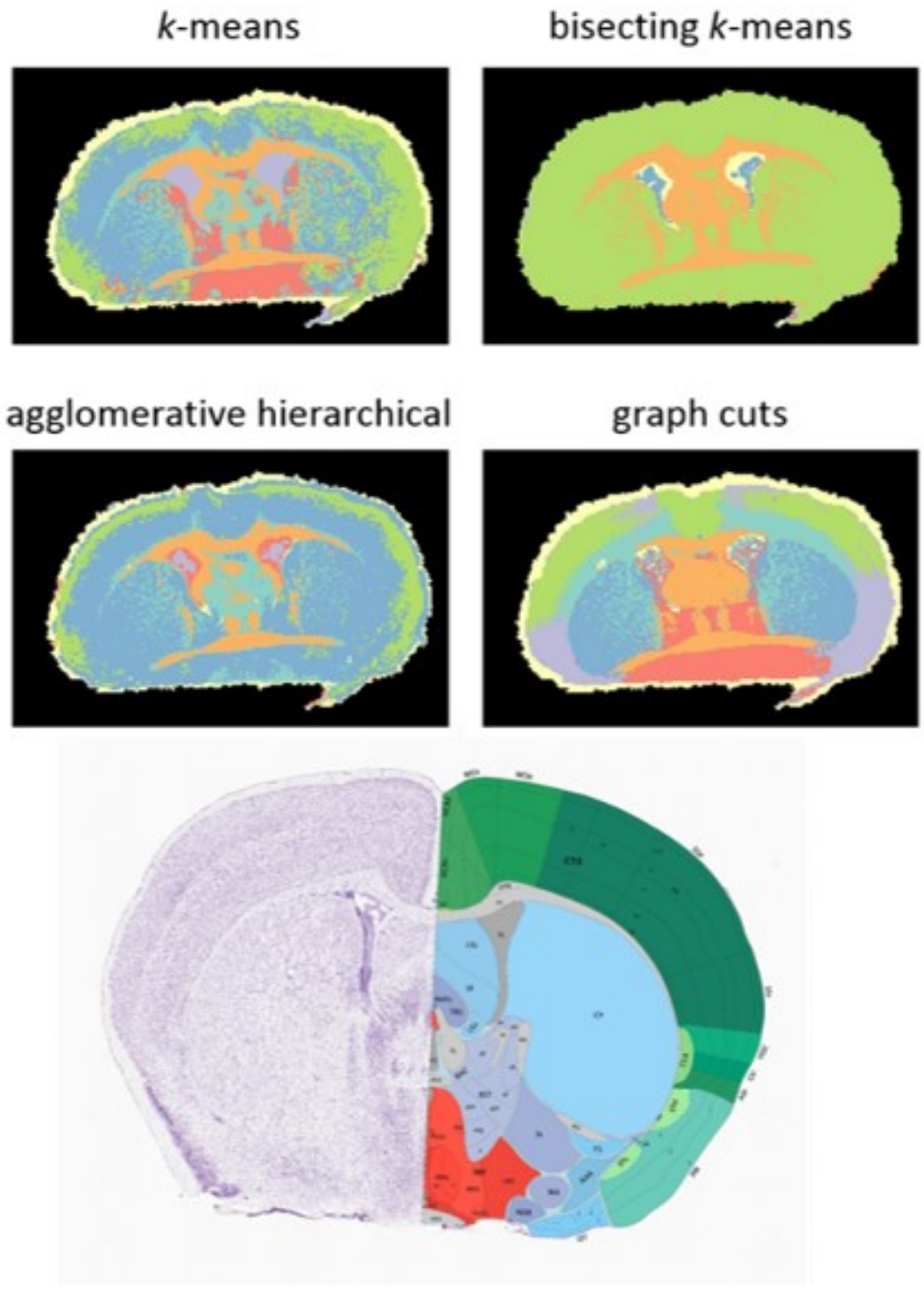

Figure 6.3: Comparison of three of the commonly used clustering algorithms in MSI, along with graph cuts on the coronal mouse brain dataset presented in chapter 5. The graph cuts algorithm more accurately segments the expected anatomies, as compared to the Allen brain atlas (bottom) [224].

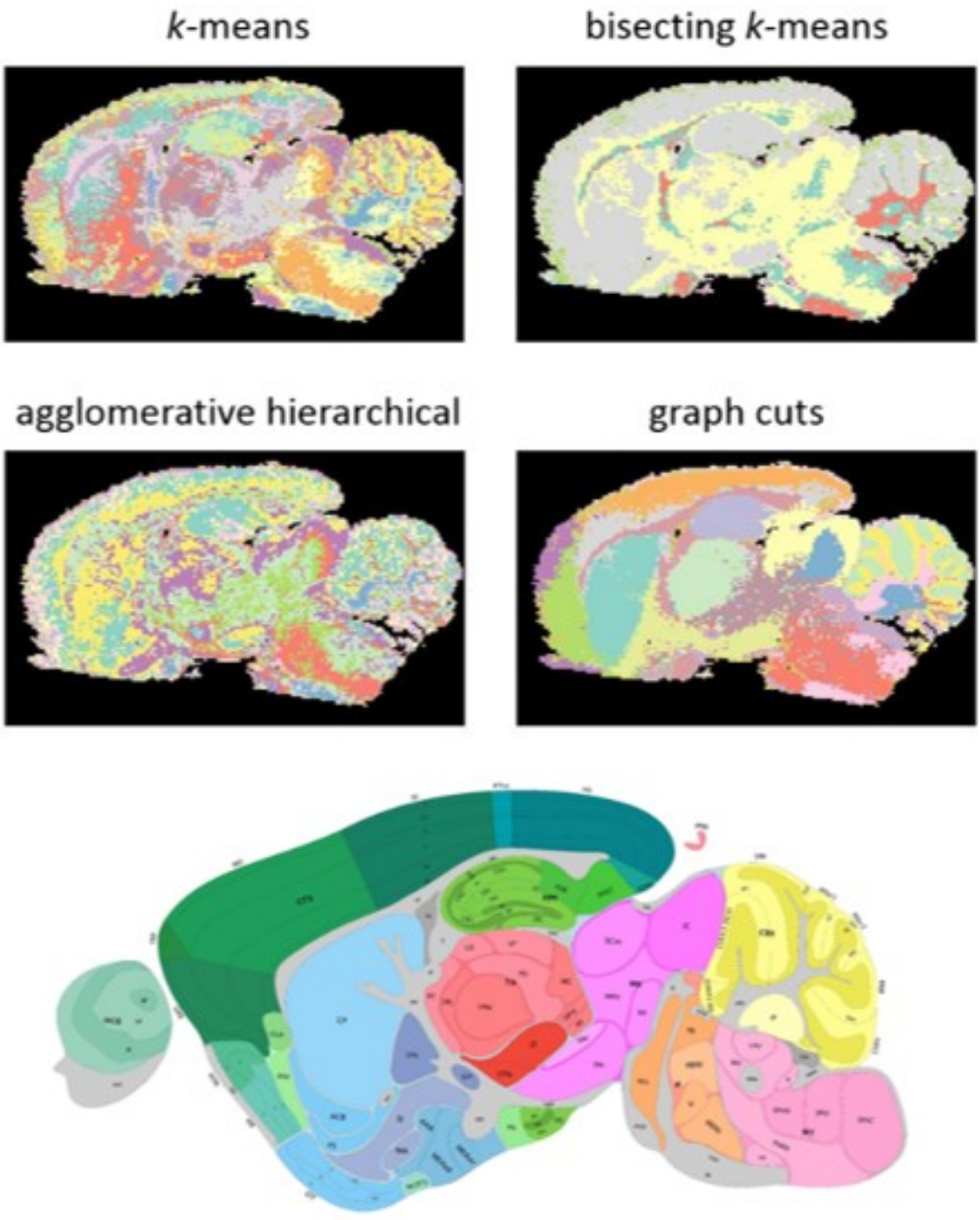

Figure 6.4: Comparison of three of the commonly used clustering algorithms, along with graph cuts on the sagittal rat brain dataset presented in chapter 5 and found at [10]. As with the results from coronal mouse brain, the graph cuts algorithm more accurately segments the expected anatomies, as compared to the Allen brain atlas (bottom) [224].

In a clinical setting, the result of the clustering must be robust to any noise or spectral differences in the data, due to the natural pixel to pixel variability within an MS image. In order to analyse a dataset with a known reduction in the spectral quality of the data, a series of previously published mouse brain datasets acquired at decreasing laser fluence were studied [56]. The spectral quality of this data decreases as the fluence falls below the threshold for ionisation (Figures 6.5). This reflects some of the extremes in the variability seen in MSI datasets where artefacts, e.g. those from inhomogeneous matrix deposition, cause localised deviations in spectra quality. In this situation, the graph cuts clustering algorithm is visibly superior to k-means clustering at segmenting the anatomical features in the tissue (Figure 6.6). This makes it more suitable for use when anatomical segmentation is the desired result of the clustering, such as in tumour differentiation. This result can be attributed to the preservation of connectivity when using the graph cuts clustering algorithm. Since the data acquired at lowest fluence will be similar to that at the next lowest, which will in turn be similar to the next lowest and so on. Therefore there is connectivity between the data acquired at the lowest fluence and the data at the highest. If the connectivity is broken, such as by only clustering a lower and higher fluence dataset, the graph cuts algorithm is able to distinguish between these experimental variances (Figure 6.7). A similar result can be achieved in simulated data derived from connected normal distributions (Figure 6.8 and 6.9). Therefore, in studies where there is likely to be an incremental change, graph cuts clustering should be used when the desired result is to cluster these incremental changes together. If the desire is to segment these incremental changes, then the k-means clustering algorithm is suitable. In all cases of clustering, the two-phase approach can be taken to improve the efficiency of the clustering.

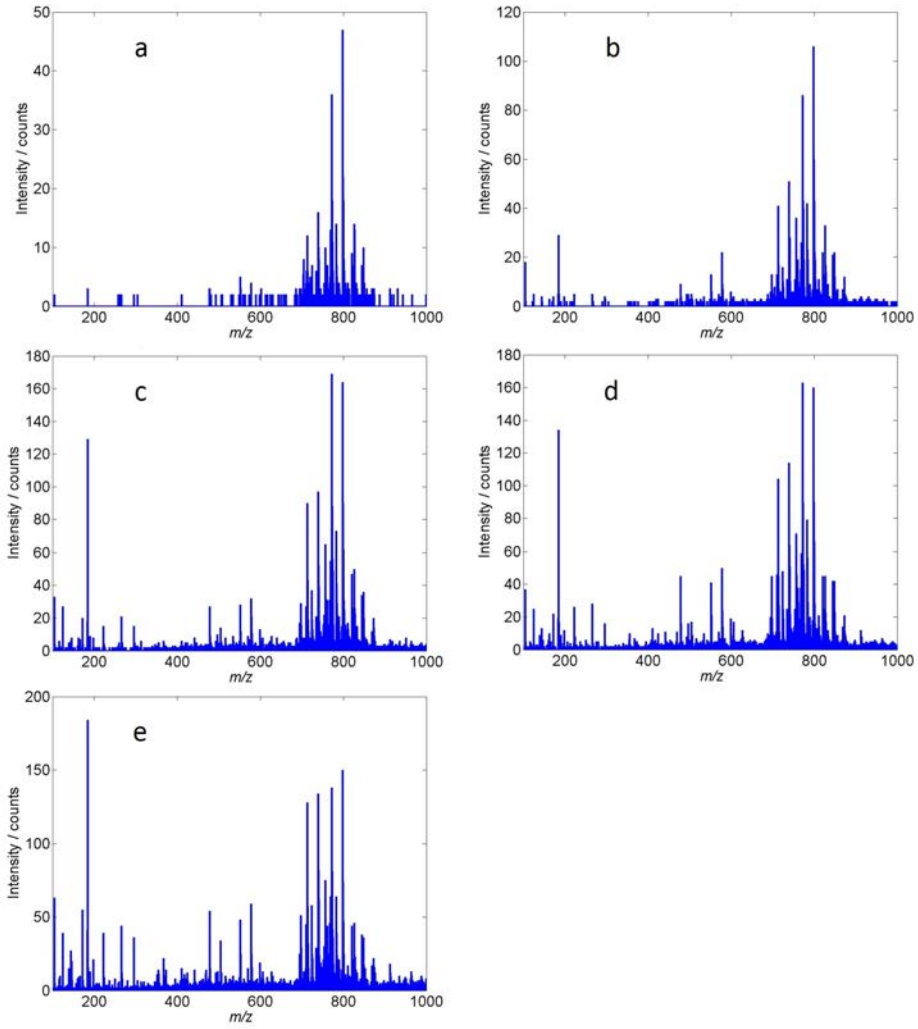

Figure 6.5: Spectra from coronal mouse brain acquired at progressively increasing fluences from $35.6 Jm^{-2}$ (a), $51.3 Jm^{-2}$ (b), $78.7 Jm^{-2}$ (c), $114.5 Jm^{-2}$ (d) up to $149.8 Jm^{-2}$ (e). The spectral quality at 35.6 and $51.3 Jm^{-2}$ is drastically poorer than that at higher fluences, and differences in the ratios between certain ions such as m/z 184 and 798 can be observed.

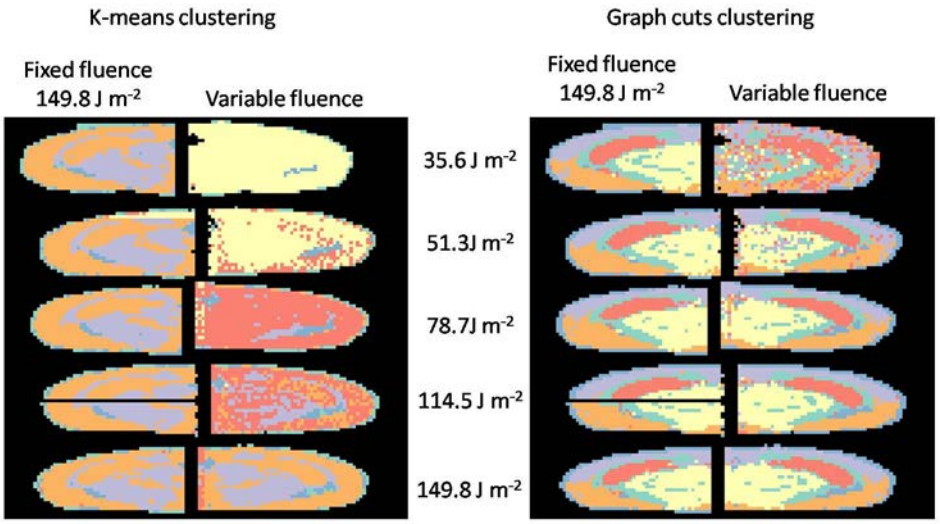

Figure 6.6: Comparison of k-means and graph cuts clustering on the data from coronal mouse brain acquired at varying fluences. The k-means clustering segments the different fluences from one another, while the graph cuts algorithm segments the anatomical features in the brain.

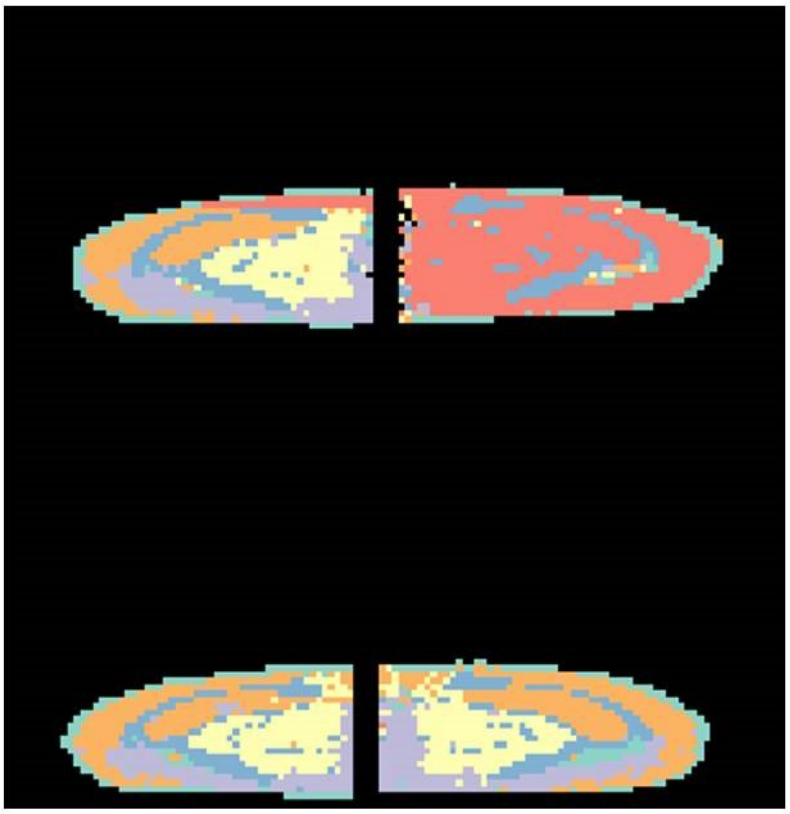

Figure 6.7: Result of graph cuts clustering on only the variable and control tissues acquired at 51.3 and 149.8 Jm^{-2}. Since there is no longer connectivity between the data within each of the sections, the different fluences are now segmented from one another.

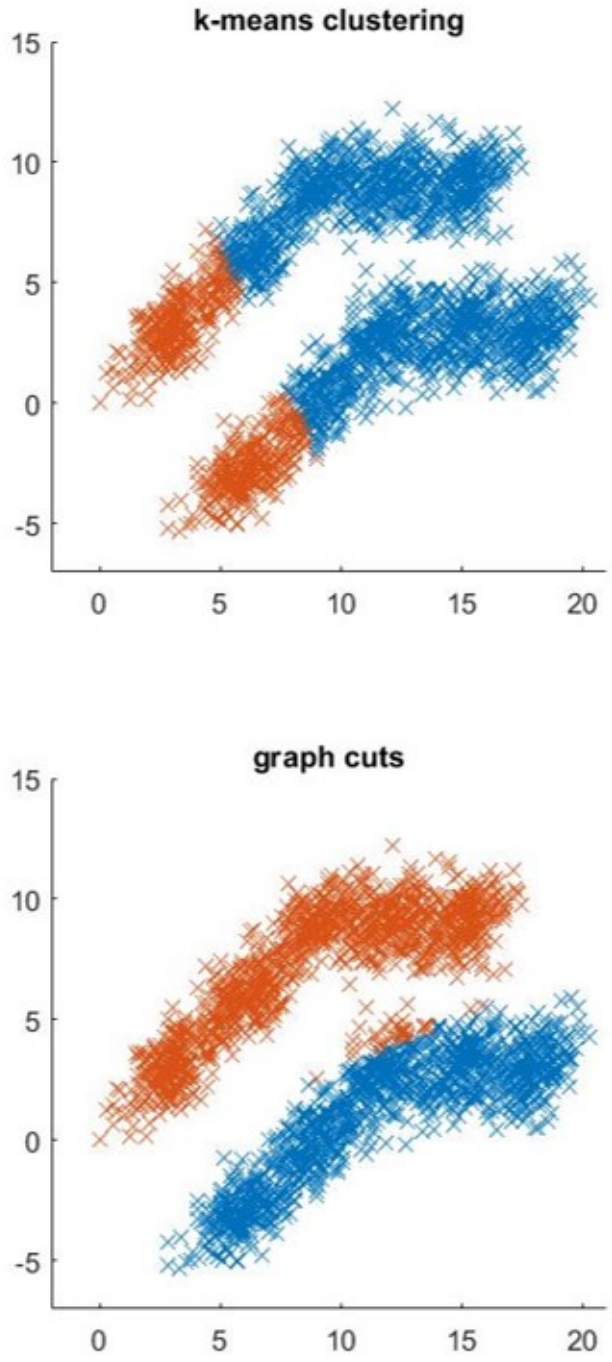

Figure 6.8: Comparison of graph cuts with k-means clustering on two clusters of connected but non Gaussian data. The graph cuts more accurately segments the banana shaped clusters from one another.

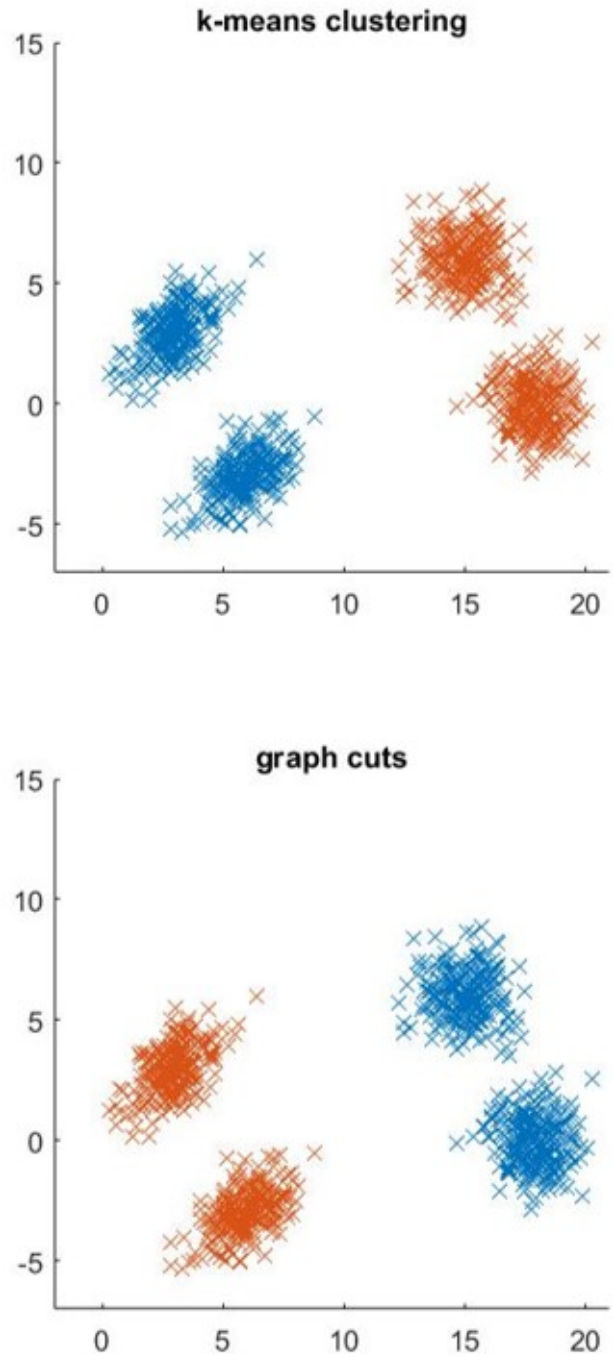

Figure 6.9: Comparison of graph cuts with k-means clustering on two clusters of connected but non Gaussian data with the central portion of the banana removed. Since there is no longer connectivity along its length, the graph cuts algorithm produces the same results as k-means clustering.

6.3.3 Two-phase clustering approach

Given a dataset with n pixels, being divided into k clusters, using a two-phase clustering algorithm such as k-means, that does not require a full pairwise distance matrix, with s subsets. Assuming the mass channels remain constant, at any one time, one subset of the data of size $\frac{n}{s}$ and the compression set of size sk needs to be loaded into RAM, giving a total of $\frac{n}{s} + sk$ memory. In order to find the minima of this, we can differentiate this with respect to the number of subsets, giving

$$\frac{d(\frac{n}{s} + sk)}{ds} = 0$$

resulting in

$$-\frac{n}{s^2} + k = 0$$

rearranging this gives

$$sk = \frac{n}{s}$$

which as established previously sk is the compression set size, and $\frac{n}{s}$ is the subset size. Therefore this algorithm is at its most memory efficient when the subsets are of equal size to the compression set. This results in $RAM_{min} \propto \sqrt{nk}$ meaning that as the number of pixels increases, the required RAM only increases by the square root of the number of pixels rather than linearly as loading the full dataset requires.

In clustering synthetic data of varying size, it can be seen that the two phase k-means clustering is around three times faster than the standard k-means algorithm (Figure 6.10). While this improvement is minimal, there is a significant potential for further optimisation of the two phase k-means clustering, through parallelisation, the use of either distributed computing networks, or graphics card processing. Critically, the results obtained from this clustering show no statistically significant loss in accuracy as compared to using k-means clustering (Figure 6.11). There is some increase in variability in these results which can be attributed to the additional random selection of subsets.

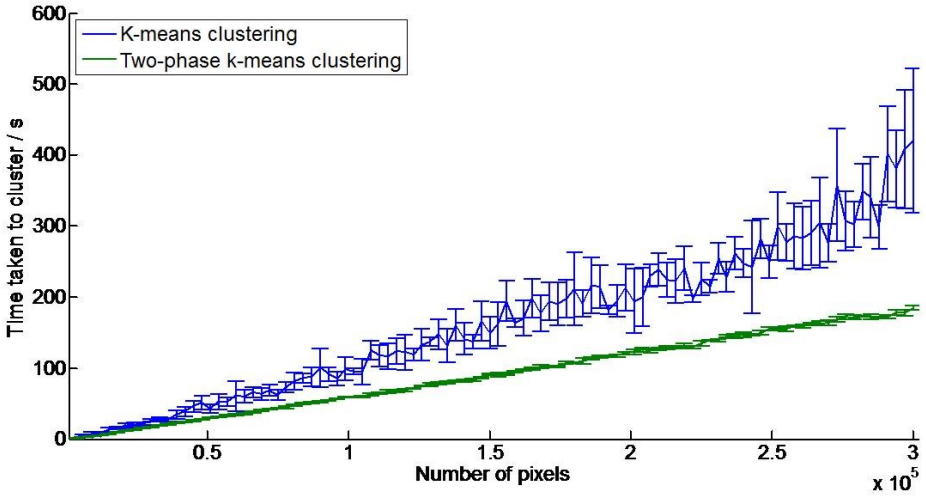

Figure 6.10: Comparison of the time taken to perform k-means or two phase k-means clustering on synthetic data with varying number of pixels. As well as a speed increase of around three times, the standard deviation of the time taken (represented by the errorbar) also decreases.

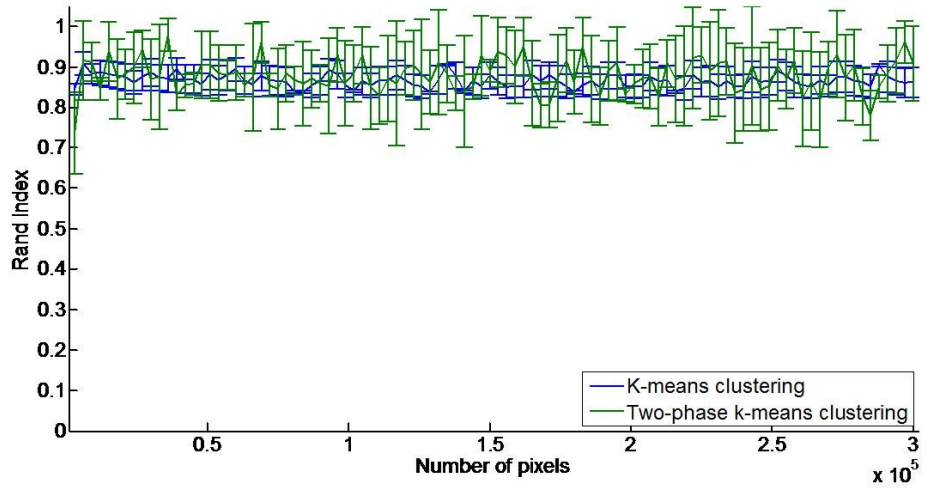

Figure 6.11: Comparison of the accuracy of k-means vs two phase k-means clustering on synthetic data with varying number of pixels. While the variability of the result increases when using the two phase k-means (larger errorbars), the mean accuracy is unchanged.

6.3.4 Two-phase graph cuts clustering

Efficiency

For the two-phase graph cuts clustering algorithm, or any algorithm that requires a full pairwise distance matrix calculation, the number of mass channels d will also affect the optimum number of subsets, and there are two possible RAM limiting steps. The first is the storage of the subset of data $\frac{nd}{s}$, the associated pairwise distance matrix $(\frac{n}{s})^2$ and the compression set sk. The other possible limitation is the compression set sk and its associated pairwise distance matrix $(sk)^2$. The most efficient possibility is when these two sets of variables are of equal size i.e.

$$\frac{nd}{s} + (\frac{n}{s})^2 + sk = sk + (sk)^2$$

rearranging this gives the formula

$$s^4k^2snd + n^2 = 0$$

quartic equations of the form $y = ax^4 + bx^3 + cx^2 + dx + e$ such as this can be solved using Ferraris solution [238], and since the cubic quadratic terms b and c are both zero, and all the other terms a, d and e are all positive, this will result in one real, positive solution for the most efficient number of subsets s [239]. If, as is the case for large MSI datasets, $n >> d$ then only the pairwise distance matrices of either subsets or the compression set need be considered, resulting in a more general RAM requirement of the algorithm to be $(\frac{n}{s})^2$. Based on the greatest efficiency when subsets and compression set are of equal size, this gives $RAM_{min} \propto nk$, as compared to $RAM = n^2$ required for standard pairwise distance matrices. Since $k << n$ the two phase graph cuts clustering scales linearly with n, as compared to n^2 without using the two phase approach. This is increasingly important as n increases since it becomes too large even for high performance PCs above around 75,000 pixels (Figure 6.13).

While using smaller subsets will generally improve efficiency, the primary basis of the two-phase clustering is that the subsets of data should be representative of the original data. Therefore using larger subsets is advisable where possible. While this may seem problematic, the larger the dataset, the larger the subsets will be, and therefore be more likely to be representative of the original data, allowing for preservation of efficiency. Typically, the largest possible subsets that can be processed within time and RAM constraints should be used within the time or RAM constraints available to the user. Additionally, the complexity of the pairwise distance matrix calculation required to perform the graph cuts clustering scales by $O(n^2.d)$. As a result, the time taken to perform graph cuts clustering increases rapidly as the number of pixels increases, even below the threshold memory available (Figure 6.12). In comparison, the time taken to perform the two-phase graph cuts clustering, when performed at peak efficiency scales linearly with n and is significantly lower than that of the standard graph cuts (Figure 6.12).

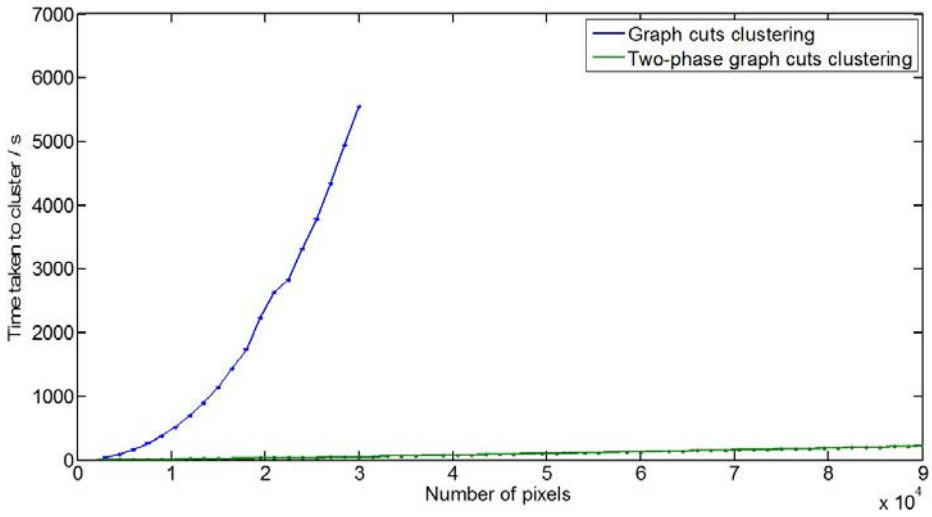

Figure 6.12: Comparison of the time taken to perform graph cuts vs two phase graph cuts clustering, with varying pixel numbers from 3,000 to 300,000. Above $\sim 30,000$ pixels, the time and memory constraints of pairwise distance calculation for graph cuts clustering becomes unfeasible.

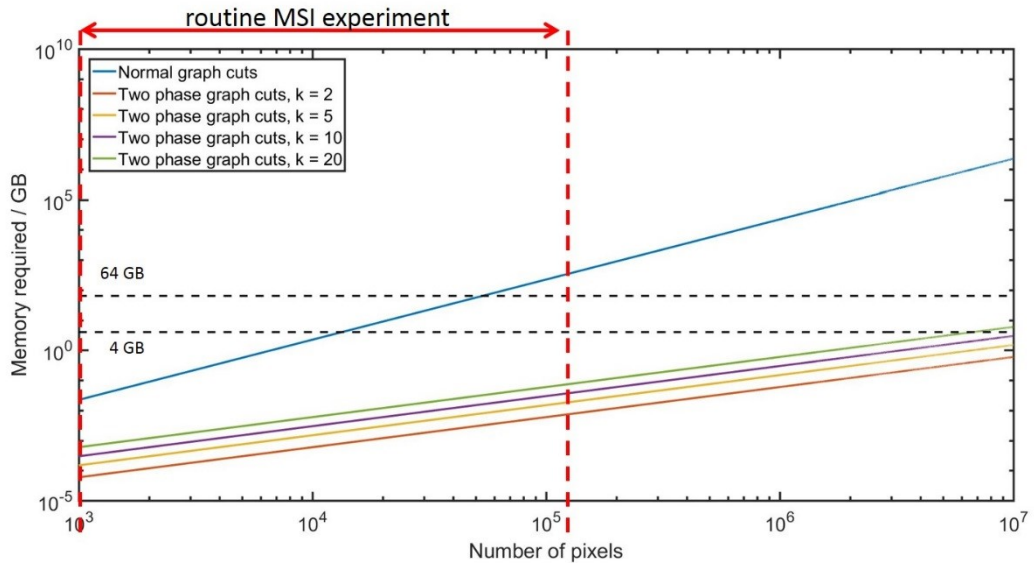

Figure 6.13: Comparison of the amount of RAM required to cluster data of varying pixel sizes using either the standard or two phase graph cuts methods. The black horizontal line represent the ranges for typical computing specifications.

Application of two phase graph cuts

Two-phase graph cuts clustering was then applied to large MSI datasets, both synthetic and derived from biological samples. To test the accuracy of the two-phase graph cuts algorithm on large and complex datasets, a large synthetic dataset comprising of 7 regions totalling 187,425 pixels and 8,193 mass channels was generated using the statistical modelling described in chapter 5. This would require 11.4 GB to load into RAM, within the capabilities of some higher end PCs but not all standard PCs or laptops. Additionally, in order to calculate and store a full pairwise distance matrix for this dataset would require over 260 GB of RAM, well beyond even high performance PCs. While large, this dataset is still well below the size of datasets often acquired on newer generation instruments or large 3D MSI datasets [95, 154]. This dataset was then clustered using k-means, two phase k-means and two phase graph cuts clustering algorithms. The two phase k-means clustering produced almost identical results to the standard k-means algorithm (Figure 6.14), but the clustering took 1.5 GB RAM for two phase k-means vs.11.5 GB for standard

k-means. Additionally, the two phase graph cuts clustering produces much more accurate results than both the k-means and two phase k-means, as measured by the Rand index (0.98 compared to 0.90), while still requiring less RAM than the k-means clustering algorithm ($<$ 3 GB).

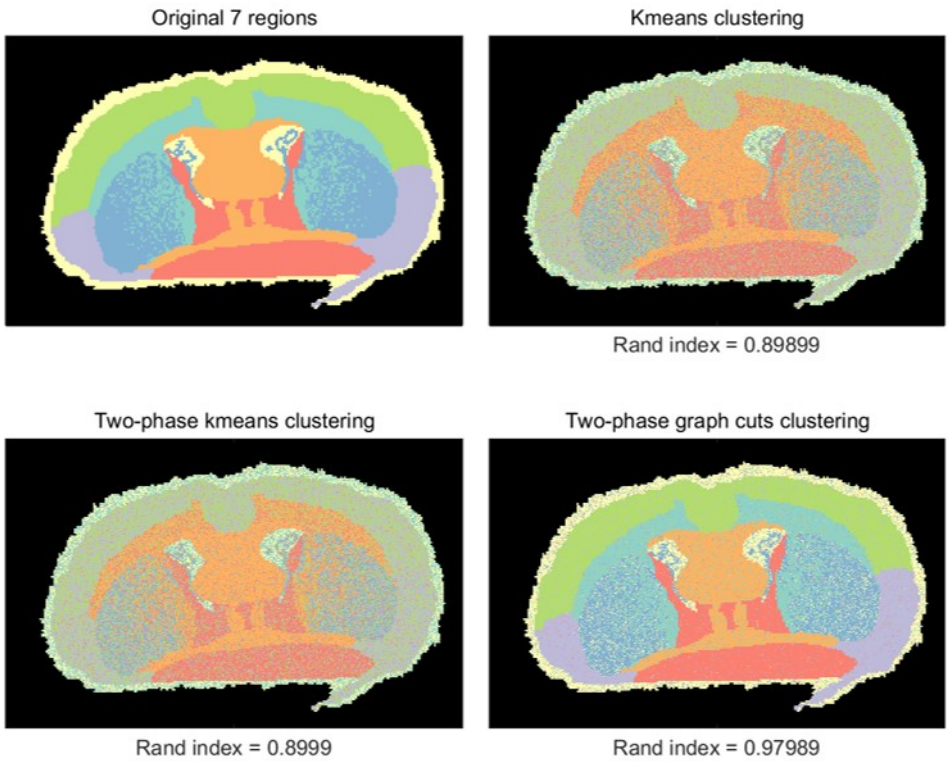

Figure 6.14: Clustering results using three different algorithms on a large synthetic dataset created using the masks shown in the top left. The two phase graph cuts clustering produces much more accurate results than the other two, and two phase and standard k-means produce almost identical results.

Two phase k-means and graph cuts clustering was then applied to large MSI datasets from biological samples. A transverse mouse brain image acquired with a pixel size of 30 μm comprising of 101,390 pixels. While not the largest MSI dataset, this represents both a large number of pixels ($>$ 100,000), rich and complex lipid spectra ($>$ 7,000 peaks), and a large number of image features ($>$ 10 anatomical regions) representing the whole

spectrum of complexity and size. As with the smaller image from the coronal mouse brain image, the two phase graph cuts clustering produces a clearer anatomical segmentation than two phase k-means clustering with respect to the expected anatomies (Figure 6.16).

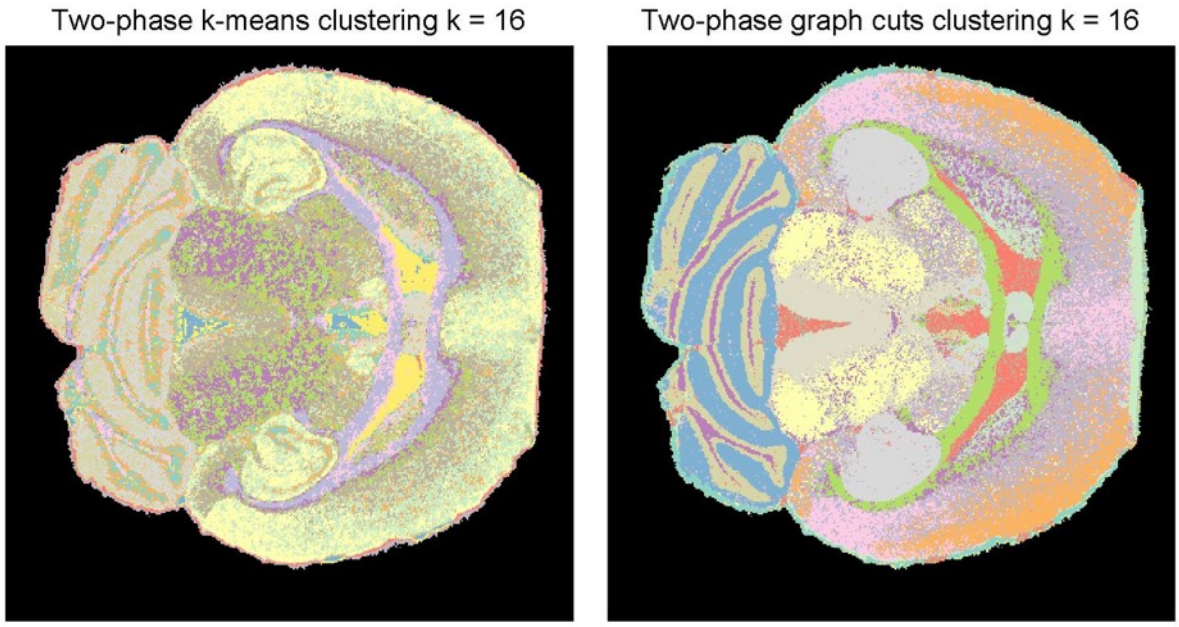

Figure 6.15: Comparison of two phase k-means and graph cuts clustering on an MSI image of transverse mouse brain. The two phase graph cuts method produces much clearer anatomical segmentation of the expected features within the brain as seen in figure 6.16.

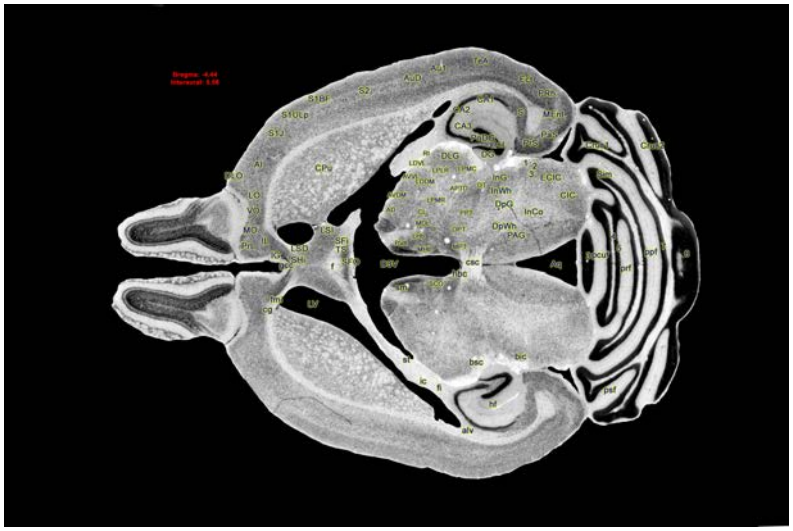

Figure 6.16: Image of an annotated transverse mouse brain image showing the expected anatomies for MSI clustering *http*: *//www.mbl.org/mblmain/atlas.html* (accessed 22/05/2017).

A larger dataset from an MSI image of gut tissue acquired with a pixel size of 5 μm was also segmented by the two phase graph cuts, and two phase k-means clustering (Figure 6.17). This dataset contained 400,625 pixels, and 6,886 spectral channels, too large to load into RAM on a standard PC (>20GB), and requiring >1TB memory to calculate a full pairwise distance matrix. In addition to MSI analysis, histopathological assessment was performed using an H&E stained serial tissue section (Figure 6.18). Four distinct anatomical layers are readily apparent; the mucosa, the sub-mucosa, the muscularis propria (externa) and the serosa. The mucosa (red) represents the innermost layer of the colon and can be sub-divided further into the epithelium, a supportive lamina propria and an outer muscularis mucosae. The mucosal epithelial layer is formed from tightly packed glands (or crypts) that open onto the surface epithelium. The neck of the glands are lined by absorptive epithelial cells, goblet cells and enteroendocrine cells, whereas stem cells and transit amplifying cells are located towards the base of the glands. The sub-mucosa (green) lies directly beneath the mucosa. The muscularis propria (grey) surrounds the sub-mucosa and consists of the inner circular and outer longitudinal smooth muscle layers.

The outer most layer, the serosa (blue), consists of a thin layer of connective tissue lined by a single layer of mesothelial cells forming the visceral peritoneum. It is important to note that due to the orientation of the colon within the Swiss roll, minor region differences in plane of the tissue are evident within the section. In addition, in some areas, the serosa and muscularis propria in particular are variably intact.

Comparison of the area of the H&E section analysed and the two phase k-means clustering clearly demonstrates that the cluster provides little or no discrimination between the various anatomical regions of the colon (Figure 6.17). In contrast, the two phase graph cuts can discriminate tissue from non-tissue and appears to start to identify specific regions (Figure 6.18). The differentiation between mucosa (red in H&E stain) and underlying sub-mucosa(green in H&E) / muscularis externa (grey in H&E stain) is particularly clear. Although the limits of resolution do not allow individual cell identification, the appearance of the clustering within the mucosa is reminiscent of the histological appearance of the mucosa and may partially recapitulate the glandular structure of the epithelium. The slight differences observed in the two phase clustering between the sections of colon within the Swiss roll may be a consequence of the variability in the section of plane as previous described.

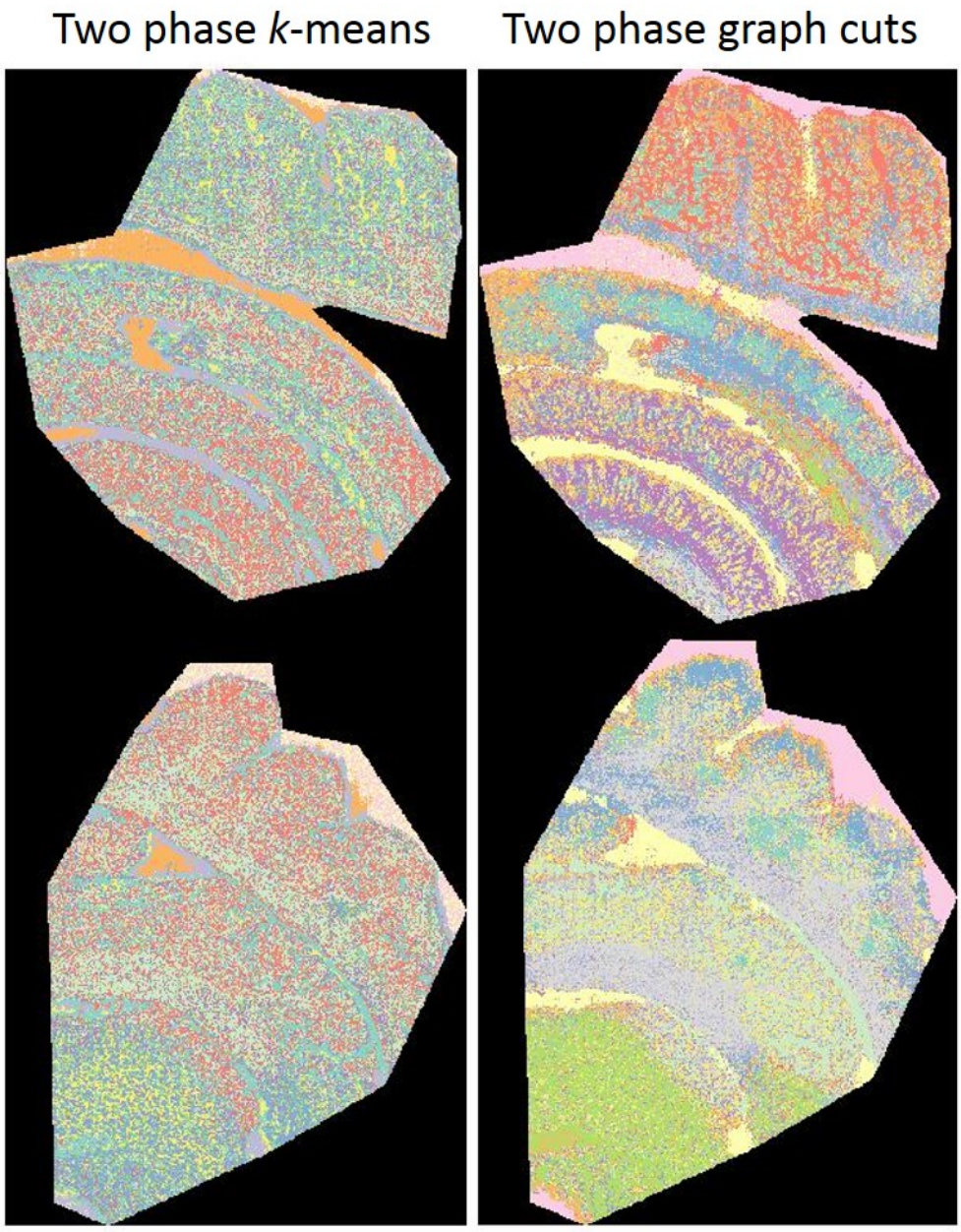

Figure 6.17: Clustering results using two phase k-means or graph cuts on a large image from mouse colon. The two phase graph cuts produces a much better segmentation as compared to labelled H&E stained serial sections 6.18.

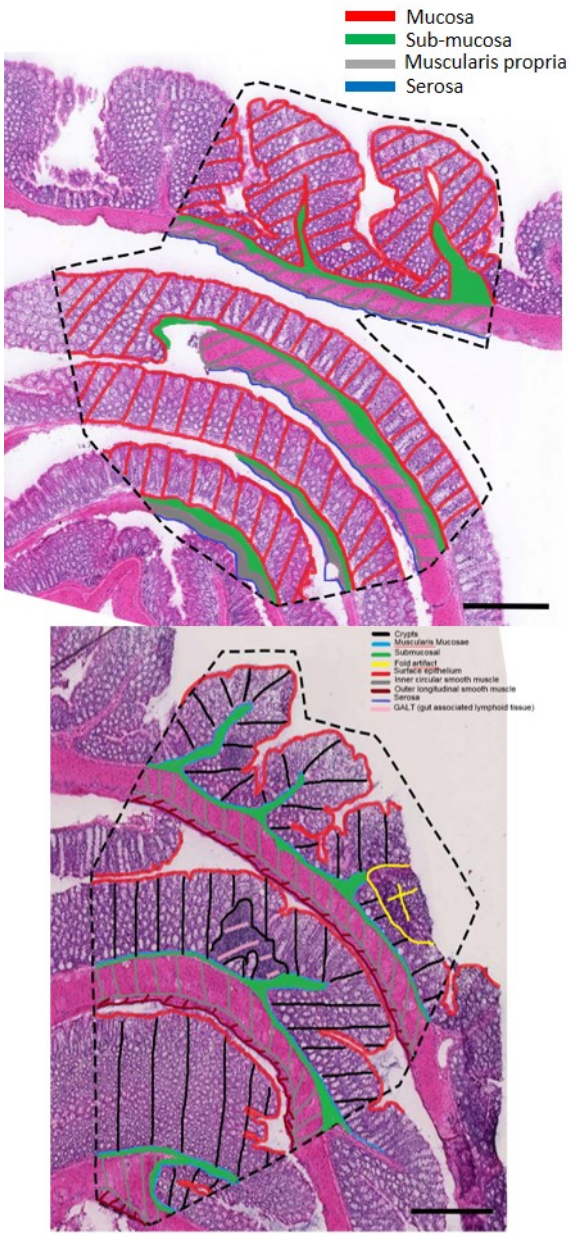

Figure 6.18: Labelled H&E stained serial tissue sections to the images from 6.17.

6.4 Conclusions

There are a wide variety of different clustering algorithms used within MSI and even more used in other fields. Here we have used the synthetic data from chapter 5 to evaluate the performance of a number of these algorithms in MSI. The graph cuts algorithm produces

the most accurate clustering results, but comes with high computational costs to do so. In cases where the full pairwise distance matrix cannot be stored in memory, or the data itself is too large to load into RAM, the two phase clustering approach can be used to reduce this cost and speed up the clustering process. This comes with only a very minimal change in the segmentation result. This two phase clustering can also be applied to any future clustering algorithm developed to increase efficiency. With new developments in instrumentation, as well as a growing need to combine and compare mutiple datasets together, data size and complexity are expected to dramatically increase in MSI in the near future. The algorithms developed will allow fast and accurate segmentation of these new generation datasets. The implementation of the two-phase graph cuts algorithm could be further improved by parallelising the preprocessing and clustering of each subset of data, and if the data were small enough, graphics processing unit (GPU) processing could be used to further speed this up. This could potentially allow for psuedo real time clustering of data as they are acquired. This could allow users to view a segmented image immediately after data are acquired from the instrument, allowing for further detailed investigation of specific areas of interest.

Using this methodology, a more complete evaluation of different variables on clustering accuracy could be performed. There are however a huge number of different, potentially interdependent variables that could be evaluated and to compare all would be extremely challenging and time consuming. For example, in the synthetic data examples shown here, each anatomical region in the data has equal numbers of pixels. Different numbers of pixels in each region would be more reflective of real MSI data. Additionally, this could be used to test the effectiveness of different clustering algorithms towards small spatial features contained within a large image. The possible combinations of numbers of pixels, number of regions, and size of each region is almost limitless and thus it would be more advisable to select these to reflect the feature sizes experienced in real MSI data. Another area that could be investigated by these means is the use of dimensionality reduction steps prior to clustering. Currently in MSI this is acheived either by PCA, t-SNE, or

random projection [131, 133, 240]. The effect of reducing the data to different numbers of dimensions was investigated by Palmer *et al.* [133] for random projection but a complete evaluation of different methods, has yet to be performed. While a full detailed evaluation is an interesting future consideration, the focus of these studies was to identify potential algorithms used in other fields that may be applicable to MSI data, and in this case, the graph cuts algorithm was chosen for further investigation based on the high accuracy. A full and detailed investigation into these variables could include analysis of the effect of different preprocessing or dimensionality methods prior to clustering, investigation into the effect of differential cluster sizes on accuracy with different algorithms, comparison of algorithm accuracies with number of clusters in the data, or combinations of these variables. These investigations could form the basis for future work in this area.

CHAPTER 7

CONCLUSIONS AND FUTURE DIRECTIONS

This thesis presents novel multivariate analysis into fundamental studies and emerging fields within MSI, and provides various means to quantifiy the performace of various algorithms for MSI datasets, and to understand the statistic distribution of MSI data. Finally, an efficient and accurate clustering algorithm, two phase graph cuts is presented, and its performance demonstrated on both synthetic and real MSI data.

In chapter 2, PCA is used in raster mode imaging to invesitgate the effect of laser parameters repetiton rate and pulse energy alonside the stage raster speed. While the full influence of repetition rate, pulse energy, and stage speed in raster mode MSI of biological samples remains undetermined, some of the effect of these variables are now better understood. The predominant factor that affects the all ion intensities is by far the energy per pulse, however in most commercial systems, the means to measure this is not present. The laser energy is also seen to be extremely variable during the initial laser firing, and a shutter based system investigated to remove this variance.The laser setup described provides a means for online measurement of laser energy during MSI experiments, as well as a shutter to maintain steady state operation to minimise energy fluctuations. Following on from this, there is a m/z dependant change in ions intensities with varying repetition rates, and this variability is not based on either lipid class or adduct formation. Furthermore, the experimental aspects of this chapter, including the use of a shutter to reduce laser variability, serial coronal mouse brain sections with each

half imaged separately, and multivariate data analysis, offers good practice guidelines to improve the quality and quantity of information gained from future fundamentals studies for MALDI tissue imaging, and can also apply to many other fundamentals studies.

MALDI-MSI is being increasingly utilised in both clinical and preclinical research, but a full understanding of all of the different variables involved is still not fully understood. As a result, research into MALDI fundamentals remain a vital area of work to understand the data acquired from MSI experiments, and the work described here will further the communities understanding in this field. There are a number of exciting areas to investigate the use of multivariate analysis in MALDI fundamentals. This investigation could be extended into the use of additonal matrices, and how they affect data from other tissues, or targeting different classes of molecules. Additionally, methods such as PCA could be used to determine the effects of the physicochemical properties of matrices and analytes on ion yields, and classification methods could be used to determine the potential usefulness of new matrices based on their properties.

In chapter 3 novel data processing is described to allow more comprehensive and detailed analysis of LESA FAIMS data. The conversion of the 2D sweep data into the imzML format allows users to determine optimal FAIMS transmission conditions for a range of different ions. This could also be used to mine data from fundamentals studies in FAIMS which, like MALDI remains unknown. In order for these softwares to become more widely used in the FAIMS community, it would be useful to include the addition of FAIMS parameters into the imzML format, or to create software to analyse the FAIMS data from the mzML format. The analysis of imaging data with an additional FAIMS or other ion mobility dimension remains a challenge. This could be investigaed using alternative methods such as 2D peak picking methods. In addition to this, if deconvolution is to be routinely used in LESA-MSI, significant speed improvements are required in the deconvolution methods, as well as well characterised benchmarks for performance.

From the work presented in chapter 4, external evaluation of clustering algorithms using the Jaccard index is advised where possible, and the synthetic data presented in

chapter 5 provide a means to quickly and reliable generate data for this purpose. Where clustering is performed in an exploratory manner, external evaluation is not possible. However when using algorithms that assume normality, testing against this normality using chi squared quantile tests are a robust means to determine the appropriate choice of distance metric to calculate spectral similarity. This could also potentially be used to estimate the number of clusters in the data. By clustering the data with varying values of k, the appropriate number of clusters could potentially be determined based on where large deviations from normal distributions are observed. However, this then requires some threshold of normality approximation to be supplied which would require further investigation to optimise.

Another area of future work in synthetic data would involve the generation of a detailed, well curated dataset to act as the benchmark for the generation of these synthetic data. This could also have accompanying anatomy specific LC and MS/MS structure elucidation to relate peaks detected back into actual detected molecules. Alongside a set of descriptive metadata, this would allow researchers to generate very detailed synthetic data with knowledge of spatial and spectral information. Combining this with a easy to use interface to generate data, and direct export to imzML format would allow mass spectrometrists to evaluate algorithms developed for MSI data, on all of the existing software platforms.

There still remains the challenge of how to appropriately model data that does not fit the assumption of normality. More complex model generation would allow a better representation of MSI data to be created. Additionally, reducing the speckling of the images, such as by placing spectrally similar pixels closer to each other would create synthetic data that are visibly more similar to real MSI data.

By combining the synthetic data and the instrumental simulations described in chapter 5, benchmark MS spectra could be generated to provide a means to evaluate different preprocessing pipelines. This is an area that currently lacks a defined optimal workflow, and there is a large degree of discrepancies within the field. Development of preprocessing

guidelines would benefit the community by allowing consistent and optimal processing of data. By generating a series of spectra with known instrumental variables, and applying different preprocessing pipelines with some scoring metric, evolutionary algorithms could be used to learn the optimal preprocessing pipelines for these data.

The work presented in Chapter 6 briefly investigates the accuracy of a number of different algorithms with respect to the number of pixels present within the data alongside measures of efficiency. Using the synthetic data from chapter 5, there are many other possible variables that can be investigated. For example, these synthetic data could be used to investigate how the numbers of peaks, clusters, and relative cluster sizes affect the accuracy of various clustering algorithms. For example, datasets with some very small clusters may have these missed by some algorithms but not others. This would be of particular benefit in MSI research, where biological features can often be very small, and so easily missed.

The two phase graph cuts algorithm described in chapter 6 can also be applied to a number of different studies, such as investigations into tumour growth or heterogeneity, anatomical segmentation for quantification studies, or feature selection for more detailed analysis such as high mass or spatial resolution analysis, or MS/MS experiments.

There is still a need to be able to include new data into a segmentation of existing data. One means to acheive this could be to segment a dataset using either clusterinig or dimensionality reduction (by reducing to three dimensions), and using this to train classifiers such as neural networks and deep learning. This would also allow even larger datasets than demonstrated in this thesis to be segmented. These segmentations could alse be used to identify regions of interest for further analysis such as by high mass or spatial resolution MSI, MS/MS imaging of specific regions, or MALDI or DESI guided LESA sampling, followed by LC/MS/MS.

LIST OF REFERENCES

[1] Richard M Caprioli, Terry B Farmer, and Jocelyn Gile. Molecular imaging of biological samples: localization of peptides and proteins using maldi-tof ms. *Analytical chemistry*, 69(23):4751–4760, 1997.

[2] Brendan Prideaux and Markus Stoeckli. Mass spectrometry imaging for drug distribution studies. *Journal of proteomics*, 75(16):4999–5013, 2012.

[3] Daisuke Miura, Yoshinori Fujimura, and Hiroyuki Wariishi. In situ metabolomic mass spectrometry imaging: recent advances and difficulties. *Journal of proteomics*, 75(16):5052–5060, 2012.

[4] Yuki SUGIURA, Shuichi SHIMMA, and Mitsutoshi SETOU. Thin sectioning improves the peak intensity and signal-to-noise ratio in direct tissue mass spectrometry. *Journal of the Mass Spectrometry Society of Japan*, 54(2):45–48, 2006.

[5] Rita Casadonte and Richard M Caprioli. Proteomic analysis of formalin-fixed paraffin-embedded tissue by maldi imaging mass spectrometry. *Nature protocols*, 6(11):1695–1709, 2011.

[6] Erin H Seeley, Stacey R Oppenheimer, Deming Mi, Pierre Chaurand, and Richard M Caprioli. Enhancement of protein sensitivity for maldi imaging mass spectrometry after chemical treatment of tissue sections. *Journal of the American Society for Mass Spectrometry*, 19(8):1069–1077, 2008.

[7] Erika R Amstalden Van Hove, Donald F Smith, Lara Fornai, Kristine Glunde, and Ron MA Heeren. An alternative paper based tissue washing method for mass spectrometry imaging: localized washing and fragile tissue analysis. *Journal of the American Society for Mass Spectrometry*, 22(10):1885–1890, 2011.

[8] M Reid Groseclose, Malin Andersson, William M Hardesty, and Richard M Caprioli. Identification of proteins directly from tissue: in situ tryptic digestions coupled with imaging mass spectrometry. *Journal of Mass Spectrometry*, 42(2):254–262, 2007.

[9] Jonathan Stauber, Luke MacAleese, Julien Franck, Emmanuelle Claude, Marten Snel, Basak Kükrer Kaletas, Ingrid MVD Wiel, Maxence Wisztorski, Isabelle Fournier, and Ron MA Heeren. On-tissue protein identification and imaging by maldi-ion mobility mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 21(3):338–347, 2010.

[10] Claire L Carter, Cameron W McLeod, and Josephine Bunch. Imaging of phospholipids in formalin fixed rat brain sections by matrix assisted laser desorption/ionization mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 22(11):1991–1998, 2011.

[11] Yuki Sugiura, Mitsutoshi Setou, and Daisuke Horigome. Matrix choice. In *Imaging Mass Spectrometry*, pages 55–69. Springer, 2010.

[12] Richard JA Goodwin. Sample preparation for mass spectrometry imaging: small mistakes can lead to big consequences. *Journal of proteomics*, 75(16):4893–4911, 2012.

[13] Edmond Hoffmann. *Mass spectrometry*. Wiley Online Library, 1996.

[14] Raymond Castaing and G Slodzian. Microanalyse par émission ionique secondaire, j. *J Microsc*, 1(395):138, 1960.

[15] PC Zalm. Secondary ion mass spectrometry. *Vacuum*, 45(6-7):753–772, 1994.

[16] Nicholas Winograd. The magic of cluster sims. *Analytical Chemistry*, 77(7):142–A, 2005.

[17] Praveen Kumar, Michael Pfeffer, Benjamin Willsch, Oliver Eibl, Lluís Yedra, Santhana Eswara, Jean-Nicolas Audinot, and Tom Wirtz. Direct imaging of dopant distributions across the si-metallization interfaces in solar cells: Correlative nanoanalytics by electron microscopy and nanosims. *Solar Energy Materials and Solar Cells*, 160:398–409, 2017.

[18] Aurélie Fauveau, Benoit Martel, Jordi Veirman, Sébastien Dubois, Anne Kaminski-Cachopo, and Frédérique Ducroquet. Comparison of characterization techniques for measurements of doping concentrations in compensated n-type silicon. *Energy Procedia*, 92:691–696, 2016.

[19] Jean-Luc Vorng, Anna M Kotowska, Melissa K Passarelli, Andrew West, Peter S Marshall, Rasmus Havelund, Martin P Seah, Colin T Dollery, Paulina D Rakowska, and Ian S Gilmore. Semi-empirical rules to determine drug sensitivity and ionization efficiency in sims using a model tissue sample. *Analytical Chemistry*, 2016.

[20] Quentin P Vanbellingen, Anthony Castellanos, Monica Rodriguez-Silva, Iru Paudel, Jeremy W Chambers, and Francisco A Fernandez-Lima. Analysis of chemotherapeutic drug delivery at the single cell level using 3d-msi-tof-sims. *Journal of The American Society for Mass Spectrometry*, 27(12):2033–2040, 2016.

[21] Tanja K Claus, Benjamin Richter, Vincent Hahn, Alexander Welle, Sven Kayser, Martin Wegener, Martin Bastmeyer, Guillaume Delaittre, and Christopher Barner-Kowollik. Simultaneous dual encoding of three-dimensional structures by light-induced modular ligation. *Angewandte Chemie International Edition*, 55(11):3817–3822, 2016.

[22] JF Mahoney, J Perel, SA Ruatta, PA Martino, S Husain, Kelsey Cook, and TD Lee. Massive cluster impact mass spectrometry: A new desorption method for the analysis of large biomolecules. *Rapid communications in mass spectrometry*, 5(10):441–445, 1991.

[23] John F Mahoney, D Shannon Cornett, Terry D Lee, and Douglas F Barofsky. Formation of multiply charged ions from large molecules using massive-cluster impact. *Rapid Communications in Mass Spectrometry*, 8(5):403–406, 1994.

[24] Alain Brunelle, David Touboul, and Olivier Laprévote. Biological tissue imaging with time-of-flight secondary ion mass spectrometry and cluster ion sources. *Journal of Mass Spectrometry*, 40(8):985–999, 2005.

[25] Jean-Luc Guerquin-Kern, Ting-Di Wu, Carmen Quintana, and Alain Croisy. Progress in analytical imaging of the cell by dynamic secondary ion mass spectrometry (sims microscopy). *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1724(3):228–238, 2005.

[26] Peter Kingshott, Sally McArthur, Helmut Thissen, David G Castner, and Hans J Griesser. Ultrasensitive probing of the protein resistance of peg surfaces by secondary ion mass spectrometry. *Biomaterials*, 23(24):4775–4785, 2002.

[27] Morgan C Putnam, Michael A Filler, Brendan M Kayes, Michael D Kelzenberg, Yunbin Guan, Nathan S Lewis, John M Eiler, and Harry A Atwater. Secondary ion mass spectrometry of vapor- liquid- solid grown, au-catalyzed, si wires. *Nano letters*, 8(10):3109–3113, 2008.

[28] NC Fenner and NR Daly. Laser used for mass analysis. *Review of Scientific Instruments*, 37(8):1068–1070, 1966.

[29] Michael Karas and Franz Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical chemistry*, 60(20):2299–2301, 1988.

[30] Renato Zenobi. Laser-assisted mass spectrometry. *CHIMIA International Journal for Chemistry*, 51(10):801–803, 1997.

[31] Franz Hillenkamp. Method and apparatus for maldi analysis, August 29 2000. US Patent 6,111,251.

[32] Richard JA Goodwin, Stephen R Pennington, and Andrew R Pitt. Protein and peptides in pictures: imaging with maldi mass spectrometry. *Proteomics*, 8(18):3785–3800, 2008.

[33] Andre Zavalin, Erik M Todd, Patrick D Rawhouser, Junhai Yang, Jeremy L Norris, and Richard M Caprioli. Direct imaging of single cells and tissue at sub-cellular spatial resolution using transmission geometry maldi ms. *Journal of Mass Spectrometry*, 47(11):1473–1481, 2012.

[34] Farida Benabdellah, Alexandre Seyer, Loïc Quinton, David Touboul, Alain Brunelle, and Olivier Laprévote. Mass spectrometry imaging of rat brain sections: nanomolar sensitivity with maldi versus nanometer resolution by tof–sims. *Analytical and bioanalytical chemistry*, 396(1):151–162, 2010.

[35] Tatiana C Rohner, Dieter Staab, and Markus Stoeckli. Maldi mass spectrometric imaging of biological tissue sections. *Mechanisms of ageing and development*, 126(1):177–185, 2005.

[36] Sarah R Barger, B Chris Hoefler, Andrés Cubillos-Ruiz, William K Russell, David H Russell, and Paul D Straight. Imaging secondary metabolism of streptomyces sp. mg1 during cellular lysis and colony degradation of competing bacillus subtilis. *Antonie Van Leeuwenhoek*, 102(3):435–445, 2012.

[37] Drew Sturtevant, Young-Jin Lee, and Kent D Chapman. Matrix assisted laser desorption/ionization-mass spectrometry imaging (maldi-msi) for direct visualization of plant metabolites in situ. *Current opinion in biotechnology*, 37:53–60, 2016.

[38] R Mirnezami, K Spagou, PA Vorkas, MR Lewis, J Kinross, E Want, H Shion, RD Goldin, A Darzi, Z Takats, et al. Chemical mapping of the colorectal cancer microenvironment via maldi imaging mass spectrometry (maldi-msi) reveals novel cancer-associated field effects. *Molecular oncology*, 8(1):39–49, 2014.

[39] David Bonnel, Rémi Longuespee, Julien Franck, Morad Roudbaraki, Pierre Gosset, Robert Day, Michel Salzet, and Isabelle Fournier. Multivariate analyses for biomarkers hunting and validation through on-tissue bottom-up or in-source decay in maldi-msi: application to prostate cancer. *Analytical and bioanalytical chemistry*, 401(1):149–165, 2011.

[40] György Marko-Varga, Thomas E Fehniger, Melinda Rezeli, Balázs Döme, Thomas Laurell, and Ákos Végvári. Drug localization in different lung cancer phenotypes by maldi mass spectrometry imaging. *Journal of proteomics*, 74(7):982–992, 2011.

[41] Brendan Prideaux, Véronique Dartois, Dieter Staab, Danielle M Weiner, Anne Goh, Laura E Via, Clifton E Barry III, and Markus Stoeckli. High-sensitivity maldi-mrm-ms imaging of moxifloxacin distribution in tuberculosis-infected rabbit lungs and granulomatous lesions. *Analytical chemistry*, 83(6):2112–2118, 2011.

[42] Dennis Trede, Stefan Schiffler, Michael Becker, Stefan Wirtz, Klaus Steinhorst, Jan Strehlow, Michaela Aichler, Jan Hendrik Kobarg, Janina Oetjen, Andrey Dyatlov, et al. Exploring three-dimensional matrix-assisted laser desorption/ionization imaging mass spectrometry data: three-dimensional spatial segmentation of mouse kidney. *Analytical chemistry*, 84(14):6079–6087, 2012.

[43] Markus Stoeckli, Dieter Staab, Alain Schweitzer, James Gardiner, and Dieter Seebach. Imaging of a β-peptide distribution in whole-body mice sections by maldi mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 18(11):1921–1924, 2007.

[44] Robert C Murphy, Joseph A Hankin, and Robert M Barkley. Imaging of lipid species by maldi mass spectrometry. *Journal of lipid research*, 50(Supplement):S317–S322, 2009.

[45] Manuela Peukert, Andrea Matros, Giuseppe Lattanzio, Stephanie Kaspar, Javier Abadía, and Hans-Peter Mock. Spatially resolved analysis of small molecules by matrix-assisted laser desorption/ionization mass spectrometric imaging (maldi-msi). *New Phytologist*, 193(3):806–815, 2012.

[46] David Calligaris, Remi Longuespee, Delphine Debois, Daiki Asakawa, Andrei Turtoi, Vincent Castronovo, Agnes Noel, Virginie Bertrand, Marie-Claire De Pauw-Gillet, and Edwin De Pauw. Selected protein monitoring in histological sections by targeted maldi-fticr in-source decay imaging. *Analytical chemistry*, 85(4):2117–2126, 2013.

[47] Malin Andersson, M Reid Groseclose, Ariel Y Deutch, and Richard M Caprioli. Imaging mass spectrometry of proteins and peptides: 3d volume reconstruction. *Nature Methods*, 5(1):101–108, 2008.

[48] F Hillenkamp and H Ehring. Laser desorption mass spectrometry. part i: Basic mechanisms and techniques. In *Mass Spectrometry in the Biological Sciences: A Tutorial*, pages 165–179. Springer, 1992.

[49] Klaus Dreisewerd. The desorption process in maldi. *Chemical reviews*, 103(2):395–426, 2003.

[50] Lisa M Preston Schaffter, Gary R Kinsel, and DH Russell. Effects of heavy-atom substituents on matrices used for matrix-assisted laser desorptionionization mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 5(9):800–806, 1994.

[51] H Ehring and BUR Sundqvist. Excited state relaxation processes of maldi-matrices studied by luminescence spectroscopy. *Applied surface science*, 96:577–580, 1996.

[52] Michael Karas, Doris Bachmann, U el Bahr, and Franz Hillenkamp. Matrix-assisted ultraviolet laser desorption of non-volatile compounds. *International journal of mass spectrometry and ion processes*, 78:53–68, 1987.

[53] Edda Lehmann, Richard Knochenmuss, and Renato Zenobi. Ionization mechanisms in matrix-assisted laser desorption/ionization mass spectrometry: contribution of pre-formed ions. *Rapid communications in mass spectrometry*, 11(14):1483–1492, 1997.

[54] Richard Knochenmuss. Ion formation mechanisms in uv-maldi. *Analyst*, 131(9):966–986, 2006.

[55] Klaus Dreisewerd, Martin Schürenberg, Michael Karas, and Franz Hillenkamp. Influence of the laser intensity and spot size on the desorption of molecules and ions in matrix-assisted laser desorption/ionization with a uniform beam profile. *International Journal of Mass Spectrometry and Ion Processes*, 141(2):127–148, 1995.

[56] Rory T Steven, Alan M Race, and Josephine Bunch. Probing the relationship between detected ion intensity, laser fluence, and beam profile in thin film and tissue in maldi msi. *Journal of The American Society for Mass Spectrometry*, pages 1–10, 2016.

[57] Klaus Dreisewerd, Martin Schürenberg, Michael Karas, and Franz Hillenkamp. Matrix-assisted laser desorption/ionization with nitrogen lasers of different pulse widths. *International journal of mass spectrometry and ion processes*, 154(3):171–178, 1996.

[58] Paul J Trim, Marie-Claude Djidja, Sally J Atkinson, Keith Oakes, Laura M Cole, David MG Anderson, Philippa J Hart, Simona Francese, and Malcolm R Clench. Introduction of a 20 khz nd: Yvo4 laser into a hybrid quadrupole time-of-flight mass spectrometer for maldi-ms imaging. *Analytical and bioanalytical chemistry*, 397(8):3409–3419, 2010.

[59] Jeffrey M Spraggins and Richard M Caprioli. High-speed maldi-tof imaging mass spectrometry: rapid ion image acquisition and considerations for next generation instrumentation. *Journal of the American Society for Mass Spectrometry*, 22(6):1022–1031, 2011.

[60] Marcel Wiegelmann, Jens Soltwisch, Thorsten W Jaskolla, and Klaus Dreisewerd. Matching the laser wavelength to the absorption properties of matrices increases the ion yield in uv-maldi mass spectrometry. *Analytical and bioanalytical chemistry*, 405(22):6925–6932, 2013.

[61] Max Born and Emil Wolf. Principles of optics ch. 9, 1999.

[62] G Westmacott, W Ens, F Hillenkamp, K Dreisewerd, and M Schürenberg. The influence of laser fluence on ion yield in matrix-assisted laser desorption ionization mass spectrometry. *International Journal of Mass Spectrometry*, 221(1):67–81, 2002.

[63] Rory Thomas Steven. *Investigations in MALDI-MSI using a high repetition rate laser*. PhD thesis, University of Birmingham, 2014.

[64] Klaus Dreisewerd. Recent methodological advances in maldi mass spectrometry. *Analytical and bioanalytical chemistry*, 406(9-10):2261–2278, 2014.

[65] Bernhard Spengler and Martin Hubert. Scanning microprobe matrix-assisted laser desorption ionization (smaldi) mass spectrometry: instrumentation for sub-

micrometer resolved ldi and maldi surface analysis. *Journal of the American Society for Mass Spectrometry*, 13(6):735–748, 2002.

[66] Sabine Guenther, Martin Koestler, Oliver Schulz, and Bernhard Spengler. Laser spot size and laser power dependence of ion formation in high resolution maldi imaging. *International Journal of Mass Spectrometry*, 294(1):7–15, 2010.

[67] Zoltan Takats, Justin M Wiseman, Bogdan Gologan, and R Graham Cooks. Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. *Science*, 306(5695):471–473, 2004.

[68] Andre Venter, Paul E Sojka, and R Graham Cooks. Droplet dynamics and ionization mechanisms in desorption electrospray ionization mass spectrometry. *Analytical Chemistry*, 78(24):8549–8555, 2006.

[69] Zoltan Takats, Justin M Wiseman, and R Graham Cooks. Ambient mass spectrometry using desorption electrospray ionization (desi): instrumentation, mechanisms and applications in forensics, chemistry, and biology. *Journal of Mass Spectrometry*, 40(10):1261–1275, 2005.

[70] Crystal Huynh and Jan Halámek. Trends in fingerprint analysis. *TrAC Trends in Analytical Chemistry*, 2016.

[71] Vilmos Kertesz and Gary J Van Berkel. Improved imaging resolution in desorption electrospray ionization mass spectrometry. *Rapid Communications in Mass Spectrometry*, 22(17):2639–2644, 2008.

[72] Yong-Seung Shin, Barbara Drolet, Richard Mayer, Kurt Dolence, and Franco Basile. Desorption electrospray ionization-mass spectrometry of proteins. *Analytical chemistry*, 79(9):3514–3518, 2007.

[73] Justin M Wiseman, Demian R Ifa, Qingyu Song, and R Graham Cooks. Tissue imaging at atmospheric pressure using desorption electrospray ionization (desi) mass spectrometry. *Angewandte Chemie International Edition*, 45(43):7188–7192, 2006.

[74] Justin M Wiseman, Demian R Ifa, Yongxin Zhu, Candice B Kissinger, Nicholas E Manicke, Peter T Kissinger, and R Graham Cooks. Desorption electrospray ionization mass spectrometry: Imaging drugs and metabolites in tissues. *Proceedings of the National Academy of Sciences*, 105(47):18120–18125, 2008.

[75] FM Green, TL Salter, IS Gilmore, P Stokes, and G O'Connor. The effect of electrospray solvent composition on desorption electrospray ionisation (desi) efficiency and spatial resolution. *Analyst*, 135(4):731–737, 2010.

[76] Anna Bodzon-Kulakowska, Anna Drabik, Joanna Ner, Jolanta Helena Kotlinska, and Piotr Suder. Desorption electrospray ionisation (desi) for beginners–how to adjust settings for tissue imaging. *Rapid Communications in Mass Spectrometry*, 28(1):1–9, 2014.

[77] FM Green, P Stokes, C Hopley, MP Seah, IS Gilmore, and G OConnor. Developing repeatable measurements for reliable analysis of molecules at surfaces using desorption electrospray ionization. *Analytical chemistry*, 81(6):2286–2293, 2009.

[78] Gary J Van Berkel, Vilmos Kertesz, Kenneth A Koeplinger, Marissa Vavrek, and Ah-Ng Tony Kong. Liquid microjunction surface sampling probe electrospray mass spectrometry for detection of drugs and metabolites in thin tissue sections. *Journal of mass spectrometry*, 43(4):500–508, 2008.

[79] Vilmos Kertesz and Gary J Van Berkel. Fully automated liquid extraction-based surface sampling and ionization using a chip-based robotic nanoelectrospray platform. *Journal of mass spectrometry*, 45(3):252–260, 2010.

[80] Gary J Van Berkel and Vilmos Kertesz. Continuous-flow liquid microjunction surface sampling probe connected on-line with high-performance liquid chromatography/mass spectrometry for spatially resolved analysis of small molecules and proteins. *Rapid Communications in Mass Spectrometry*, 27(12):1329–1334, 2013.

[81] Patrick J Roach, Julia Laskin, and Alexander Laskin. Nanospray desorption electrospray ionization: an ambient method for liquid-extraction surface sampling in mass spectrometry. *Analyst*, 135(9):2233–2236, 2010.

[82] Kristina Clemons, Chinyere Nnaji, and Guido F Verbeck. Overcoming selectivity and sensitivity issues of direct inject electrospray mass spectrometry via dapnensi-ms. *Journal of The American Society for Mass Spectrometry*, 25(5):705–711, 2014.

[83] Joscelyn Sarsby, Nicholas J Martin, Patricia F Lalor, Josephine Bunch, and Helen J Cooper. Top-down and bottom-up identification of proteins by liquid extraction surface analysis mass spectrometry of healthy and diseased human liver tissue. *Journal of The American Society for Mass Spectrometry*, 25(11):1953–1961, 2014.

[84] Daniel Eikel, Marissa Vavrek, Sheri Smith, Carol Bason, Suzie Yeh, Walter A Korfmacher, and Jack D Henion. Liquid extraction surface analysis mass spectrometry (lesa-ms) as a novel profiling tool for drug distribution and metabolism analysis: the terfenadine example. *Rapid Communications in Mass Spectrometry*, 25(23):3587–3596, 2011.

[85] Matthew J Walworth, Mariam S ElNaggar, Joseph J Stankovich, Chuck Witkowski, Jeremy L Norris, and Gary J Van Berkel. Direct sampling and analysis from solid-phase extraction cards using an automated liquid extraction surface analysis nano-electrospray mass spectrometry system. *Rapid Communications in Mass Spectrometry*, 25(17):2389–2396, 2011.

[86] Maxence Wisztorski, Benoit Fatou, Julien Franck, Annie Desmons, Isabelle Farré, Eric Leblanc, Isabelle Fournier, and Michel Salzet. Microproteomics by liquid extraction surface analysis: Application to ffpe tissue to study the fimbria region of tubo-ovarian cancer. *PROTEOMICS-Clinical Applications*, 7(3-4):234–240, 2013.

[87] Elizabeth C Randall, Josephine Bunch, and Helen J Cooper. Direct analysis of intact proteins from escherichia coli colonies by liquid extraction surface analysis mass spectrometry. *Analytical chemistry*, 86(21):10504–10510, 2014.

[88] Reinaldo Almeida, Zane Berzina, Eva C Arnspang, Jan Baumgart, Johannes Vogt, Robert Nitsch, and Christer S Ejsing. Quantitative spatial analysis of the mouse brain lipidome by pressurized liquid extraction surface analysis. *Analytical chemistry*, 87(3):1749–1756, 2015.

[89] Nicholas J Martin, Josephine Bunch, and Helen J Cooper. Dried blood spot proteomics: surface extraction of endogenous proteins coupled with automated sample preparation and mass spectrometry analysis. *Journal of The American Society for Mass Spectrometry*, 24(8):1242–1249, 2013.

[90] John G Swales, James W Tucker, Michael J Spreadborough, Suzanne L Iverson, Malcolm R Clench, Peter JH Webborn, and Richard JA Goodwin. Mapping drug distribution in brain tissue using liquid extraction surface analysis mass spectrometry imaging. *Analytical chemistry*, 87(19):10146–10152, 2015.

[91] Rian L Griffiths, Alex Dexter, Andrew J Creese, and Helen J Cooper. Liquid extraction surface analysis field asymmetric waveform ion mobility spectrometry mass spectrometry for the analysis of dried blood spots. *Analyst*, 140(20):6879–6885, 2015.

[92] Clara L Feider, Natalia Elizondo, and Livia S Eberlin. Ambient ionization and faims mass spectrometry for enhanced imaging of multiply charged molecular ions in biological tissues. *Analytical Chemistry*, 88(23):11533–11541, 2016.

[93] Joscelyn Sarsby, Rian L Griffiths, Alan M Race, Josephine Bunch, Elizabeth C Randall, Andrew J Creese, and Helen J Cooper. Liquid extraction surface analysis mass spectrometry coupled with field asymmetric waveform ion mobility spectrometry for analysis of intact proteins from biological substrates. *Analytical chemistry*, 87(13):6794–6800, 2015.

[94] David A Barnett, Randy W Purves, Barbara Ells, and Roger Guevremont. Separation of o-, m-and p-phthalic acids by high-field asymmetric waveform ion mobility spectrometry (faims) using mixed carrier gases. *Journal of mass spectrometry*, 35(8):976–980, 2000.

[95] Nina Ogrinc Potočnik, Tiffany Porta, Michael Becker, Ron Heeren, and Shane R Ellis. Use of advantageous, volatile matrices enabled by next-generation high-speed matrix-assisted laser desorption/ionization time-of-flight imaging employing a scanning laser beam. *Rapid Communications in Mass Spectrometry*, 29(23):2195–2203, 2015.

[96] Andre Zavalin, Junhai Yang, Kevin Hayden, Marvin Vestal, and Richard M Caprioli. Tissue protein imaging at 1 μm laser spot diameter for high spatial resolution and high imaging speed using transmission geometry maldi tof ms. *Analytical and bioanalytical chemistry*, 407(8):2337–2342, 2015.

[97] Donglu Zhang and Sekhar Surapaneni. *ADME-enabling technologies in drug design and development*. John Wiley & Sons, 2012.

[98] Andreas Römpp, Jean-Pierre Both, Alain Brunelle, Ron MA Heeren, Olivier Laprévote, Brendan Prideaux, Alexandre Seyer, Bernhard Spengler, Markus Stoeckli, and Donald F Smith. Mass spectrometry imaging of biological tissue: an approach for multicenter studies. *Analytical and bioanalytical chemistry*, 407(8):2329–2335, 2015.

[99] Facundo Fernandez, Charles McEwen, Jack A Syage, Zhang Ouyang, Gary Hieftje, Andre Venter, K Hiraoka, Graham Cooks, Akos Vertes, Jen-Taie Shiea, et al. *Ambient ionization mass spectrometry*. Royal Society of Chemistry, 2014.

[100] Hanan Awad, Mona M Khamis, and Anas El-Aneed. Mass spectrometry, review of the basics: ionization. *Applied Spectroscopy Reviews*, 50(2):158–175, 2015.

[101] BA Mamyrin, VI Karataev, DV Shmikk, and VA Zagulin. The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer with high resolution. *Soviet Journal of Experimental and Theoretical Physics*, 37:45, 1973.

[102] Liam A McDonnell, Todd H Mize, Stefan L Luxembourg, Sander Koster, Gert B Eijkel, Elisabeth Verpoorte, Nico F de Rooij, and Ron MA Heeren. Using matrix peaks to map topography: Increased mass resolution and enhanced sensitivity in chemical imaging. *Analytical chemistry*, 75(17):4373–4381, 2003.

[103] Igor V Chernushevich, Alexander V Loboda, and Bruce A Thomson. An introduction to quadrupole–time-of-flight mass spectrometry. *Journal of Mass Spectrometry*, 36(8):849–865, 2001.

[104] M Guilhaus, D Selby, and V Mlynski. Orthogonal acceleration time-of-flight mass spectrometry. *Mass spectrometry reviews*, 19(2):65–107, 2000.

[105] Peter H Dawson. *Quadrupole mass spectrometry and its applications*. Elsevier, 2013.

[106] Raymond E Kaiser, R Graham Cooks, George C Stafford, John EP Syka, and Philip H Hemberger. Operation of a quadrupole ion trap mass spectrometer to achieve high mass/charge ratios. *International journal of mass spectrometry and ion processes*, 106:79–115, 1991.

[107] Fred W McLafferty. Tandem mass spectrometry. *Science*, 214(4518):280–287, 1981.

[108] MT Rodgers, Kent M Ervin, and Peter B Armentrout. Statistical modeling of collision-induced dissociation thresholds. *The Journal of chemical physics*, 106(11):4499–4508, 1997.

[109] Robert L Fitzgerald, Carol L O'Neal, Bradley J Hart, Alphonse Poklis, and David A Herold. Comparison of an ion-trap and a quadrupole mass spectrometer using diazepam as a model compound. *Journal of analytical toxicology*, 21(6):445–450, 1997.

[110] Alan G Marshall, Christopher L Hendrickson, and George S Jackson. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass spectrometry reviews*, 17(1):1–35, 1998.

[111] Evgene N Nikolaev, Gleb N Vladimirov, Roland Jertz, and Gökhan Baykut. From supercomputer modeling to highest mass resolution in ft-icr. *Mass Spectrometry*, 2(Spec Iss), 2013.

[112] Alan G Marshall and Christopher L Hendrickson. Fourier transform ion cyclotron resonance detection: principles and experimental configurations. *International Journal of Mass Spectrometry*, 215(1):59–75, 2002.

[113] Jeffrey M Spraggins, David G Rizzo, Jessica L Moore, Kristie L Rose, Neal D Hammer, Eric P Skaar, and Richard M Caprioli. Maldi fticr ims of intact proteins: Using mass accuracy to link protein images with proteomics data. *Journal of The American Society for Mass Spectrometry*, 26(6):974–985, 2015.

[114] Qizhi Hu, Robert J Noll, Hongyan Li, Alexander Makarov, Mark Hardman, and R Graham Cooks. The orbitrap: a new mass spectrometer. *Journal of mass spectrometry*, 40(4):430–443, 2005.

[115] Paulo J Amorim Madeira, Carlos M Borges, and Pedro A Alves. *High resolution mass spectrometry using FTICR and orbitrap instruments*. INTECH Open Access Publisher, 2012.

[116] Fulvio Magni, Maciej Lalowski, Veronica Mainini, Martina Marchetti-Deschmann, Clizia Chinello, Andrea Urbani, and Marc Baumann. Proteomics imaging and the kidney. *J Nephrol*, 26(3):430–436, 2013.

[117] NE Manicke, AL Dill, DR Ifa, and RG Cooks. High-resolution tissue imaging on an orbitrap mass spectrometer by desorption electrospray ionization mass spectrometry. *Journal of mass spectrometry*, 45(2):223–226, 2010.

[118] Alan M. Race, Iain B. Styles, and Josephine Bunch. Inclusive sharing of mass spectrometry imaging data requires a converter for all. *Journal of Proteomics*, 75(16):5111 – 5112, 2012.

[119] Alan M Race. *Investigation and interpretation of large mass spectrometry imaging datasets*. PhD thesis, University of Birmingham, 2016.

[120] Theodore Alexandrov and Jan Hendrik Kobarg. Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. *Bioinformatics*, 27(13):i230–i238, 2011.

[121] Theodore Alexandrov, Michael Becker, Soeren-Oliver Deininger, Gunther Ernst, Liane Wehder, Markus Grasmair, Ferdinand von Eggeling, Herbert Thiele, and Peter Maass. Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering. *Journal of proteome research*, 9(12):6535–6546, 2010.

[122] Judith M Fonville, Claire L Carter, Luis Pizarro, Rory T Steven, Andrew D Palmer, Rian L Griffiths, Patricia F Lalor, John C Lindon, Jeremy K Nicholson, Elaine Holmes, et al. Hyperspectral visualization of mass spectrometry imaging data. *Analytical chemistry*, 85(3):1415–1423, 2013.

[123] Steven Rudich and Avi Wigderson. *Computational complexity theory*. American Mathematical Soc., 2004.

[124] Keigo Kimura, Yuzuru Tanaka, and Mineichi Kudo. A fast hierarchical alternating least squares algorithm for orthogonal nonnegative matrix factorization. In *Asian Conference on Machine Learning*, pages 129–141, 2015.

[125] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[126] Victor Y Pan and Zhao Q Chen. The complexity of the matrix eigenproblem. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 507–516. ACM, 1999.

[127] Virginia Vassilevska Williams. Multiplying matrices in o (n2. 373) time. *preprint*, 2014.

[128] Limin Fu and Enzo Medico. Flame, a novel fuzzy clustering method for the analysis of dna microarray data. *BMC bioinformatics*, 8(1):1, 2007.

[129] Alan M. Race, Rory T. Steven, Andrew D. Palmer, Iain B. Styles, and Josephine Bunch. Memory efficient principal component analysis for the dimensionality reduction of large mass spectrometry imaging data sets. *Analytical Chemistry*, 85(6):3071–3078, 2013.

[130] Yuki Sugiura, Yoshiyuki Konishi, Nobuhiro Zaima, Shigeki Kajihara, Hiroki Nakanishi, Ryo Taguchi, and Mitsutoshi Setou. Visualization of the cell-selective distribution of pufa-containing phosphatidylcholines in mouse brain by imaging mass spectrometry. *Journal of lipid research*, 50(9):1776–1788, 2009.

[131] Gregor McCombie, Dieter Staab, Markus Stoeckli, and Richard Knochenmuss. Spatial and spectral correlations in maldi mass spectrometry images by clustering and multivariate analysis. *Analytical chemistry*, 77(19):6118–6124, 2005.

[132] Xing-Chuang Xiong, FANG Xiang, Zheng Ouyang, You Jiang, Ze-Jian Huang, and Yu-Kui Zhang. Feature extraction approach for mass spectrometry imaging data using non-negative matrix factorization. *Chinese Journal of Analytical Chemistry*, 40(5):663–669, 2012.

[133] Andrew D Palmer, Josephine Bunch, and Iain B Styles. Randomized approximation methods for the efficient compression and analysis of hyperspectral data. *Analytical chemistry*, 85(10):5078–5086, 2013.

[134] Michael Hanselmann, Marc Kirchner, Bernhard Y Renard, Erika R Amstalden, Kristine Glunde, Ron MA Heeren, and Fred A Hamprecht. Concise representation of mass spectrometry images by probabilistic latent semantic analysis. *Analytical chemistry*, 80(24):9649–9658, 2008.

[135] Spencer A Thomas, Alan M Race, Rory T Steven, Ian S Gilmore, and Josephine Bunch. Dimensionality reduction of mass spectrometry imaging data using autoencoders. In *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*, pages 1–7. IEEE, 2016.

[136] Alan M Race and Josephine Bunch. Optimisation of colour schemes to accurately display mass spectrometry imaging data based on human colour perception. *Analytical and bioanalytical chemistry*, 407(8):2047–2054, 2015.

[137] Donald Michie, David J Spiegelhalter, and Charles C Taylor. *Machine learning, neural and statistical classification*. Citeseer, 1994.

[138] M Reid Groseclose, Pierre P Massion, Pierre Chaurand, and Richard M Caprioli. High-throughput proteomic analysis of formalin-fixed paraffin-embedded tissue microarrays using maldi imaging mass spectrometry. *Proteomics*, 8(18):3715–3724, 2008.

[139] Lisa H Cazares, Dean Troyer, Savvas Mendrinos, Raymond A Lance, Julius O Nyalwidhe, Hind A Beydoun, Mary Ann Clements, Richard R Drake, and O John Semmes. Imaging mass spectrometry of a specific fragment of mitogen-activated protein kinase/extracellular signal-regulated kinase kinase kinase 2 discriminates cancer from uninvolved prostate tissue. *Clinical Cancer Research*, 15(17):5541–5551, 2009.

[140] Theodore Alexandrov. Maldi imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC bioinformatics*, 13(Suppl 16):S11, 2012.

[141] Kirill A Veselkov, Reza Mirnezami, Nicole Strittmatter, Robert D Goldin, James Kinross, Abigail VM Speller, Tigran Abramov, Emrys A Jones, Ara Darzi, Elaine Holmes, et al. Chemo-informatic strategy for imaging mass spectrometry-based hyperspectral profiling of lipid signatures in colorectal cancer. *Proceedings of the National Academy of Sciences*, 111(3):1216–1221, 2014.

[142] Soren-Oliver Deininger, Matthias P Ebert, Arne Futterer, Marc Gerhard, and Christoph Rocken. Maldi imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *Journal of proteome research*, 7(12):5230–5236, 2008.

[143] Stefan M Willems, Alexandra van Remoortere, René van Zeijl, André M Deelder, Liam A McDonnell, and Pancras CW Hogendoorn. Imaging mass spectrometry of myxoid sarcomas identifies proteins and lipids specific to tumour type and grade, and reveals biochemical intratumour heterogeneity. *The Journal of pathology*, 222(4):400–409, 2010.

[144] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[145] Derya Birant and Alp Kut. St-dbscan: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1):208–221, 2007.

[146] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[147] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[148] Avrim Blum and Shuchi Chawla. Learning from labeled and unlabeled data using graph mincuts. 2001.

[149] DT Pham, SS Dimov, and CD Nguyen. A two-phase k-means algorithm for large datasets. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 218(10):1269–1273, 2004.

[150] Rory T Steven, Alex Dexter, and Josephine Bunch. Investigating maldi msi parameters (part 2)–on the use of a mechanically shuttered trigger system for improved laser energy stability. *Methods*, 2016.

[151] Sören-Oliver Deininger, Dale S Cornett, Rainer Paape, Michael Becker, Charles Pineau, Sandra Rauser, Axel Walch, and Eryk Wolski. Normalization in maldi-tof imaging datasets of proteins: practical considerations. *Analytical and bioanalytical chemistry*, 401(1):167–181, 2011.

[152] Emrys A Jones, Alexandra van Remoortere, René JM van Zeijl, Pancras CW Hogendoorn, Judith VMG Bovée, André M Deelder, and Liam A McDonnell. Multiple statistical analysis techniques corroborate intratumor heterogeneity in imaging mass spectrometry datasets of myxofibrosarcoma. *PLoS One*, 6(9):e24913, 2011.

[153] Walid M Abdelmoula, Ricardo J Carreira, Reinald Shyti, Benjamin Balluff, Rene JM van Zeijl, Else A Tolner, Boudewijn FP Lelieveldt, Arn MJM van den Maagdenberg, Liam A McDonnell, and Jouke Dijkstra. Automatic registration of mass spectrometry imaging data sets to the allen brain atlas. *Analytical chemistry*, 86(8):3947–3954, 2014.

[154] Janina Oetjen, Kirill Veselkov, Jeramie Watrous, James S McKenzie, Michael Becker, Lena Hauberg-Lotte, Jan Hendrik Kobarg, Nicole Strittmatter, Anna K Mróz, Franziska Hoffmann, et al. Benchmark datasets for 3d maldi-and desi-imaging mass spectrometry. *GigaScience*, 4(1):1, 2015.

[155] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[156] ER Muir, IJ Ndiour, NA LeGoasduff, Richard A Moffitt, Ying Liu, M Cameron Sullards, Alfred H Merrill, Yanfeng Chen, and May D Wang. Multivariate analysis of imaging mass spectrometry data. In *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*, pages 472–479. IEEE, 2007.

[157] Andrew D Palmer, Josephine Bunch, and Iain B Styles. The use of random projections for the analysis of mass spectrometry imaging data. *Journal of The American Society for Mass Spectrometry*, 26(2):315–322, 2015.

[158] Theodore Alexandrov, Michael Becker, Orlando Guntinas-Lichius, Günther Ernst, and Ferdinand von Eggeling. Maldi-imaging segmentation is a powerful tool for spatial functional proteomic analysis of human larynx carcinoma. *Journal of cancer research and clinical oncology*, 139(1):85–95, 2013.

[159] Douglass R Cutting, David R Karger, Jan O Pedersen, and John W Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In

Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, pages 318–329. ACM, 1992.

[160] Paul S Bradley and Usama M Fayyad. Refining initial points for k-means clustering. In *ICML*, volume 98, pages 91–99. Citeseer, 1998.

[161] Walid M Abdelmoula, Karolina Skraskova, Benjamin Balluff, Ricardo J Carreira, Else A Tolner, Boudewijn PF Lelieveldt, Laurens van der Maaten, Hans Morreau, Arn MJM van den Maagdenberg, Ron MA Heeren, et al. Automatic generic registration of mass spectrometry imaging data to histology using nonlinear stochastic embedding. *Analytical chemistry*, 86(18):9204–9211, 2014.

[162] Dan Pelleg and Andrew Moore. Accelerating exact k-means algorithms with geometric reasoning. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 277–281. ACM, 1999.

[163] Dennis Trede, Stefan Schiffler, Michael Becker, Stefan Wirtz, Jan Strehlow, Jan Hendrik Kobarg, Klaus Steinhorst, Michaela Aichler, Janina Oetjen, Andrey Dyatlov, et al. O5. scils lab: software for analysis and interpretation of large maldi-ims datasets. *OurCon 2012*, page 50.

[164] Michael Steinbach, George Karypis, Vipin Kumar, et al. A comparison of document clustering techniques. In *KDD workshop on text mining*, volume 400, pages 525–526. Boston, 2000.

[165] Janina Oetjen, Michaela Aichler, Dennis Trede, Jan Strehlow, Judith Berger, Stefan Heldmann, Michael Becker, Michael Gottschalk, Jan Hendrik Kobarg, Stefan Wirtz, et al. Mri-compatible pipeline for three-dimensional maldi imaging mass spectrometry using paxgene fixation. *Journal of proteomics*, 90:52–60, 2013.

[166] Lukas Krasny, Franziska Hoffmann, Günther Ernst, Dennis Trede, Theodore Alexandrov, Vladimir Havlicek, Orlando Guntinas-Lichius, Ferdinand von Eggeling, and Anna C Crecelius. Spatial segmentation of maldi ft-icr msi data: a powerful tool to explore the head and neck tumor in situ lipidome. *Journal of The American Society for Mass Spectrometry*, 26(1):36–43, 2015.

[167] Gabriele De Sio, Andrew James Smith, Manuel Galli, Mattia Garancini, Clizia Chinello, Francesca Bono, Fabio Pagni, and Fulvio Magni. A maldi-mass spectrometry imaging method applicable to different formalin-fixed paraffin-embedded human tissues. *Molecular bioSystems*, 11(6):1507–1514, 2015.

[168] JK Boelke, M Gerhard, FM Schleif, J Decker, M Kuhn, T Elssner, W Pusch, and M Kostrzewa. Clinprotools 2.0 user documentation, 2005. *Available in the ClinProtClinProTools*, 2.

[169] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[170] Robin Sibson. Slink: an optimally efficient algorithm for the single-link cluster method. *The computer journal*, 16(1):30–34, 1973.

[171] Daniel Defays. An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):364–366, 1977.

[172] William HE Day and Herbert Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24, 1984.

[173] Sören-Oliver Deininger, Michael Becker, and Detlev Suckau. Tutorial: multivariate statistical treatment of imaging data for clinical biomarker discovery. *Mass Spectrometry Imaging: Principles and Protocols*, pages 385–403, 2010.

[174] Rima Ait-Belkacem, Caroline Berenguer, Claude Villard, L'Houcine Ouafik, Dominique Figarella-Branger, Olivier Chinot, and Daniel Lafitte. Maldi imaging and in-source decay for top-down characterization of glioblastoma. *Proteomics*, 14(10):1290–1301, 2014.

[175] Christos Faloutsos and King-Ip Lin. *FastMap: A fast algorithm for indexing, datamining and visualization of traditional and multimedia datasets*, volume 24. ACM, 1995.

[176] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[177] Andrew D Palmer. *Information processing for mass spectrometry imaging*. PhD thesis, University of Birmingham, 2014.

[178] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[179] Song Wang and Jeffrey Mark Siskind. Image segmentation with ratio cut. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(6):675–690, 2003.

[180] Yudong Chen, Shiau Hong Lim, and Huan Xu. Weighted graph clustering with non-uniform uncertainties. In *ICML*, pages 1566–1574, 2014.

[181] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining*, pages 911–916. IEEE, 2010.

[182] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[183] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[184] Joseph C Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104, 1974.

[185] Raf Van de Plas, Fabian Ojedaa, Maarten Dewile, Ludo Van, Bart De Moora Den Bosche, and Etienne Waelkensbcd. Unsupervised spatial tissue exploration via imaging mass spectrometry: Preprocessing and clustering. *Bioinformatics*, 2006.

[186] Theodore Alexandrov, Ilya Chernyavsky, Michael Becker, Ferdinand von Eggeling, and Sergey Nikolenko. Analysis and interpretation of imaging mass spectrometry data by clustering mass-to-charge images according to their spatial similarity. *Analytical chemistry*, 85(23):11189–11195, 2013.

[187] Sanaiya Sarkari, Chanchala D Kaddi, Rachel V Bennett, Facundo M Fernández, and May D Wang. Comparison of clustering pipelines for the analysis of mass spectrometry imaging data. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4771–4774. IEEE, 2014.

[188] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[189] Rebecca W Garden and Jonathan V Sweedler. Heterogeneity within maldi samples as revealed by mass spectrometric imaging. *Analytical chemistry*, 72(1):30–36, 2000.

[190] Rory T Steven, Andrew D Palmer, and Josephine Bunch. Fluorometric beam profiling of uv maldi lasers. *Journal of The American Society for Mass Spectrometry*, 24(7):1146–1152, 2013.

[191] Alan M Race, Andrew D Palmer, Alex Dexter, Rory T Steven, Iain B Styles, and Josephine Bunch. Spectralanalysis: software for the masses. *Analytical Chemistry*, 88(19):9451–9458, 2016.

[192] Jochen Th Westheide, J Sabine Becker, Ralf Jäger, Hans-Joachim Dietze, and José AC Broekaert. Analysis of ceramic layers for solid oxide fuel cells by laser ablation inductively coupled plasma mass spectroscopy. *Journal of Analytical Atomic Spectrometry*, 11(9):661–666, 1996.

[193] Shelley N Jackson, Michael Ugarov, Jeremy D Post, Thomas Egan, Denis Langlais, J Albert Schultz, and Amina S Woods. A study of phospholipids by ion mobility tofms. *Journal of the American Society for Mass Spectrometry*, 19(11):1655–1662, 2008.

[194] Peter Marshall, Valerie Toteu-Djomte, Philippe Bareille, Hayley Perry, Gillian Brown, Mark Baumert, and Keith Biggadike. Correlation of skin blanching and percutaneous absorption for glucocorticoid receptor agonists by matrix-assisted laser desorption ionization mass spectrometry imaging and liquid extraction surface analysis with nanoelectrospray ionization mass spectrometry. *Analytical chemistry*, 82(18):7787–7794, 2010.

[195] Martin RL Paine, Philip J Barker, Shane A Maclauglin, Todd W Mitchell, and Stephen J Blanksby. Direct detection of additives and degradation products from polymers by liquid extraction surface analysis employing chip-based nanospray mass spectrometry. *Rapid Communications in Mass Spectrometry*, 26(4):412–418, 2012.

[196] Magdalena Montowska, Morgan R Alexander, Gregory A Tucker, and David A Barrett. Rapid detection of peptide markers for authentication purposes in raw and cooked meat using ambient liquid extraction surface analysis mass spectrometry. *Analytical chemistry*, 86(20):10257–10265, 2014.

[197] Darren Kessner, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. Proteowizard: open source software for rapid proteomics tools development. *Bioinformatics*, 24(21):2534–2536, 2008.

[198] David M Horn, Roman A Zubarev, and Fred W McLafferty. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry*, 11(4):320–332, 2000.

[199] Kyle D Bemis, April Harry, Livia S Eberlin, Christina Ferreira, Stephanie M van de Ven, Parag Mallick, Mark Stolowitz, and Olga Vitek. Cardinal: an r package for statistical analysis of mass spectrometry-based imaging experiments. *Bioinformatics*, page btv146, 2015.

[200] HanBin Oh, Kathrin Breuker, Siu Kwan Sze, Ying Ge, Barry K Carpenter, and Fred W McLafferty. Secondary and tertiary structures of gaseous protein ions characterized by electron capture dissociation mass spectrometry and photofragment spectroscopy. *Proceedings of the National Academy of Sciences*, 99(25):15863–15868, 2002.

[201] Alexandre A Shvartsburg, Fumin Li, Keqi Tang, and Richard D Smith. Characterizing the structures and folding of free proteins using 2-d gas-phase separations: observation of multiple unfolded conformers. *Analytical chemistry*, 78(10):3304–3315, 2006.

[202] Georgios Papadopoulos, Annette Svendsen, Oleg V Boyarkin, and Thomas R Rizzo. Conformational distribution of bradykinin [bk+ 2 h] 2+ revealed by cold ion spectroscopy coupled with faims. *Journal of the American Society for Mass Spectrometry*, 23(7):1173–1181, 2012.

[203] Calvin Rorrer and Iii Leonard. *Parameters affecting performance of planar high-field asymmetric waveform ion mobility spectrometry (FAIMS)*. University of Florida, 2010.

[204] Zhongqi Zhang and Alan G Marshall. A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *Journal of the American Society for Mass Spectrometry*, 9(3):225–233, 1998.

[205] Arthur L Patterson. A direct method for the determination of the components of interatomic distances in crystals. *Zeitschrift für Kristallographie-Crystalline Materials*, 90(1-6):517–542, 1935.

[206] Michael W Senko, Steven C Beu, and Fred W McLafferty. Automated assignment of charge states from resolved isotopic peaks for multiply charged ions. *Journal of the American Society for Mass Spectrometry*, 6(1):52–56, 1995.

[207] Michael T Heideman, Don H Johnson, and C Sidney Burrus. Gauss and the history of the fast fourier transform. *Archive for history of exact sciences*, 34(3):265–277, 1985.

[208] Li Chen and Yee Leng Yap. Automated charge state determination of complex isotope-resolved mass spectra by peak-target fourier transform. *Journal of the American Society for Mass Spectrometry*, 19(1):46–54, 2008.

[209] Michael W Senko, Steven C Beu, and Fred W McLafferty. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, 6(4):229–233, 1995.

[210] Anoop M Mayampurath, Navdeep Jaitly, Samuel O Purvine, Matthew E Monroe, Kenneth J Auberry, Joshua N Adkins, and Richard D Smith. Deconmsn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. *Bioinformatics*, 24(7):1021–1023, 2008.

[211] Paulo C Carvalho, Tao Xu, Xuemei Han, Daniel Cociorva, Valmir C Barbosa, and John R Yates. Yada: a tool for taking the most out of high-resolution spectra. *Bioinformatics*, 25(20):2734–2736, 2009.

[212] Shawn Donnelly, Josette E Marrero, Trevor Cornell, Kevin Fowler, and John Allison. Analysis of pigmented inkjet printer inks and printed documents by laser desorption/mass spectrometry. *Journal of forensic sciences*, 55(1):129–135, 2010.

[213] Thomas K Burdenski Jr. Evaluating univariate, bivariate, and multivariate normality using graphical procedures. *Multiple Linear Regression Viewpoints*, 2000.

[214] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

[215] Maurice G Kendall. *A Course in the Geometry of n Dimensions*. Courier Corporation, 2004.

[216] Jianchang Mao and Anil K Jain. A self-organizing network for hyperellipsoidal clustering (hec). *Ieee transactions on neural networks*, 7(1):16–29, 1996.

[217] Greg Hamerly and Charles Elkan. Learning the k in a¿ means. *Advances in neural information processing systems*, 16:281, 2004.

[218] Douglas Steinley. K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59(1):1–34, 2006.

[219] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[220] Hubert W Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402, 1967.

[221] Donald A Darling. The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics*, 28(4):823–838, 1957.

[222] Jelle J Goeman, Sara A Van De Geer, and Hans C Van Houwelingen. Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):477–493, 2006.

[223] MJRi Healy. Multivariate normal plotting. *Applied Statistics*, pages 157–161, 1968.

[224] Ed S Lein, Michael J Hawrylycz, Nancy Ao, Mikael Ayres, Amy Bensinger, Amy Bernard, Andrew F Boe, Mark S Boguski, Kevin S Brockway, Emi J Byrnes, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–176, 2007.

[225] Andrew Palmer, Ekaterina Ovchinnikova, Mikael Thuné, Régis Lavigne, Blandine Guével, Andrey Dyatlov, Olga Vitek, Charles Pineau, Mats Borén, and Theodore Alexandrov. Using collective expert judgements to evaluate quality measures of mass spectrometry images. *Bioinformatics*, 31(12):i375–i384, 2015.

[226] Michael Ljungberg, Sven-Erik Strand, and Michael A King. *Monte Carlo calculations in nuclear medicine: Applications in diagnostic imaging*. CRC Press, 2012.

[227] Habib Zaidi. Relevance of accurate monte carlo modeling in nuclear medical imaging. *Medical physics*, 26(4):574–608, 1999.

[228] Jan Hendrik Kobarg. *Signal and image processing methods for imaging mass spectrometry data*. PhD thesis, Bremen, Universität Bremen, Diss., 2014, 2014.

[229] Annette J Dobson. *Introduction to statistical modelling*. Springer, 2013.

[230] James Jianmeng Xu. *Statistical modelling and inference for multivariate and longitudinal discrete response data*. PhD thesis, University of British Columbia, 1996.

[231] Kevin R Coombes, John M Koomen, Keith A Baggerly, Jeffrey S Morris, and Ryuji Kobayashi. Understanding the characteristics of mass spectrometry data through the use of simulation. *Cancer Informatics*, 1, 2005.

[232] Andreas Ipsen. Derivation from first principles of the statistical distribution of the mass peak intensities of ms data. *Analytical chemistry*, 87(3):1726–1734, 2015.

[233] Eamonn Keogh and Abdullah Mueen. Curse of dimensionality. In *Encyclopedia of Machine Learning*, pages 257–258. Springer, 2011.

[234] Oded Maimon and Lior Rokach. Introduction to knowledge discovery and data mining. In *Data mining and knowledge discovery handbook*, pages 1–15. Springer, 2009.

[235] Carol M Park, Philip E Reid, David C Walker, and Brian R MacPherson. A simple, practical swiss rollmethod of preparing tissues for paraffin or methacrylate embedding. *Journal of microscopy*, 145(1):115–120, 1987.

[236] Mohammadreza Shariatgorji, Anna Nilsson, Richard JA Goodwin, Patrik Källback, Nicoletta Schintu, Xiaoqun Zhang, Alan R Crossman, Erwan Bezard, Per Svenningsson, and Per E Andren. Direct targeted quantitative molecular imaging of neurotransmitters in brain tissue sections. *Neuron*, 84(4):697–707, 2014.

[237] Rong Jin, Chris Ding, and Feng Kang. A probabilistic approach for optimizing spectral clustering. In *NIPS*, pages 571–578, 2005.

[238] Stefan Neumark. *Solution of cubic and quartic equations*. Elsevier, 2014.

[239] Girolamo Cardano and Massimo Tamborini. *Artis magnae, sive, De regulis algebraicis liber unus*. Franco Angeli, 1545.

[240] Walid M Abdelmoula, Benjamin Balluff, Sonja Englert, Jouke Dijkstra, Marcel JT Reinders, Axel Walch, Liam A McDonnell, and Boudewijn PF Lelieveldt. Data-driven identification of prognostic tumor subpopulations using spatially mapped t-sne of mass spectrometry imaging data. *Proceedings of the National Academy of Sciences*, page 201510227, 2016.

Research presented in this thesis has been included in the following publications

STEVEN, R. T., DEXTER, A. & BUNCH, J. 2016b. Investigating MALDI MSI parameters (Part 2)On the use of a mechanically shuttered trigger system for improved laser energy stability. Methods.

STEVEN, R. T., DEXTER, A. & BUNCH, J. 2016a. Investigating MALDI MSI parameters (Part 1)A systematic survey of the effects of repetition rates up to 20kHz in continuous raster mode. Methods.

DEXTER, A., RACE, A., STYLES, I. & BUNCH, J. 2016. Testing for multivariate normality in mass spectrometry imaging data: A robust statistical approach for clustering evaluation and the generation of synthetic mass spectrometry imaging datasets. Analytical chemistry.

DEXTER, A., RACE, A. M., STEVEN, R. T., BARNES, J. R., HULME, H., GOODWIN, R. J., STYLES, I. B. & BUNCH, J. 2017. Two-phase and graph based clustering methods for accurate and efficient segmentation of large mass spectrometry images. Analytical Chemistry.