INVESTIGATION AND INTERPRETATION OF LARGE MASS SPECTROMETRY IMAGING DATASETS

by

ALAN M. RACE

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
May 2016

Abstract

Mass spectrometry imaging (MSI) enables two- and three-dimensional overviews of hundreds of unlabelled molecular species including drugs, metabolites, lipids and proteins in complex samples such as intact tissue. Mining and interpreting such complex datasets manually is a tedious, laborious and error prone task. In this research, a new extensible software platform is presented, suitable for spectral preprocessing, multivariate analysis and visualisation of large MSI datasets. New solutions for data conversion and the benefits of the new software for review of data generated using a variety of MS imaging modalities using several major vendors instruments is illustrated. Principal component analysis (PCA), in combination with data binning or other reduction algorithms, has been widely used in the unsupervised processing of MSI data and as a dimensionality reduction method prior to clustering and spatial segmentation. Standard implementations of PCA require the entire dataset to be stored in memory, necessitating a compromise between the number of pixels and the number of peaks to include. In this research a new method which has no limitation on the number of pixels and allows an increased number of peaks to be retained is developed and validated against the MATLAB (The MathWorks Inc., Natick, Massachusetts) implementation of PCA (*princomp*). Application of k-means clustering to the reduced data was used for segmentation of a rat brain dataset. Correlation analysis was then used to identify ions associated with selected clusters. Hierarchical composition of data has been shown as an efficient method of capturing the information present within images in other fields. An adaptation of these ideas to MSI data is described and used to identify ions associated with specific spatial distributions present within the rat brain data, which were not elucidated using PCA or correlation analysis. The research presented in this thesis has resulted in new recommendations for presentation of MS images. The way in which imaging data are presented can have a significant impact on the perceived structure, especially when using false colour to display images. This is of great

significance in MSI as the majority of data analysis and interpretation is performed by manual inspection of significant ion images. An evaluation of the perceptual linearity of the colour schemes used within the literature revealed that over 40% of surveyed publications used colour schemes which can result in an actively misleading representation of the data. Several suggestions for superior presentation of MS imaging data to avoid unnecessary visual artefacts are described. Finally, the software and algorithms presented were used to analyse MSI data from a traumatic brain injury model. Manual exploration and use of multivariate analysis methods such as PCA did not reveal any differences between the injured hemisphere of the brain and the control hemisphere, however the hierarchical composition algorithms identified multiple ion images which appear elevated in the injured hemisphere.

ACKNOWLEDGEMENTS

I would primarily like to thank my supervisors Dr Josephine Bunch for all her invaluable guidance and her ever-helpful, humorous and insightful discussions and Dr Iain Styles for all of his support and intellectual input throughout the PhD. Thanks also to Dr Aleš Leonardis for discussions relating to the hierarchical composition algorithm development and Dr Zsuzsanna Nagy and her group for designing and conducting the traumatic brain injury experiment.

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).

Thanks also to the members of Bunch Group past and present, specifically Rory Steven, Andrew Palmer, Rian Griffiths and Joscelyn Sarsby. Many of our discussions helped to shape the work undertaken throughout the PhD. Thanks to the members of PSIBS, who not only made it a fun place to work, but also provided unique and invaluable insight, in particular Eric Pitkeathly, Alistair Bannerman and James Brown.

Finally, thank you to my family and friends, this would not have been possible without you all.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Mass spectrometry (MS) is an analytical technique which measures the mass-to-charge ratio (m/z) of gas phase ions which have been extracted from a sample. Mass spectrometers consist of an ion source, one or more mass analysers and a detector. The specific configuration used in a given mass spectrometer often tailors the instrument to certain analytical tasks. MS has a wide list of detectable analytes (such as elements [1], drugs [2, 3], metabolites [2, 3], lipids [4, 5], peptides [6] and proteins [7]) in a vast array of applications (such as surface analysis [8], biomedical research [9], clinical analysis [10], drug discovery [11] and genomics [12]).

The resulting data consists of pairs of values, m/z and a measure of ion intensity, which are plotted against one another (x- and y-axis respectively) for display purposes and form a mass spectrum. The units along the x-axis are often displayed as daltons (Da), referring to the unified atomic mass unit, however this is only appropriate when either all detected ions are singly charged, $z = 1$, or the m/z of each ion in the spectrum has been multiplied the ion's charge state. In the case where mass is measured in atomic mass units and charge is measured in elementary charge units, the Thompson (Th) unit has been proposed [13]. However this has not been widely accepted or adopted. Generally the x-axis is labelled with the dimensionless abbreviation m/z [14].

The underlying physics of the measure of ion intensity is dependent on the detector used. The intensity is often reported in various ways, such as the number of ions or

ion counts (with arbitrary units) or as a relative intensity or relative abundance (as a percentage of the largest intensity in the spectrum).

1.1 Ion Sources

The purpose of the ion source is to generate gas phase ions from the sample of interest. Two major properties are used to describe an ionisation method. The first being whether it is hard, resulting in the breaking of chemical bonds to form ions, or soft, where intact molecules are ionised. The second is whether ionisation occurs under vacuum or at atmospheric pressure, where ions are formed outside of the mass spectrometer.

An example of a hard ionisation method performed under vacuum is inductively coupled plasma mass spectrometry (ICP-MS) where an inductively coupled plasma is used to ionise the elements constituting a sample. An ambient, soft ionisation method is electrospray ionisation (ESI) where a voltage is applied to a liquid producing multiply charged ions [15]. A soft technique which can be performed either under vacuum or in ambient conditions is laser desorption/ionisation (LDI) where a laser is used to desorb and ionise molecules.

Extensions to these methods exist often to enable different classes of molecules to be detected, for example desorption electrospray ionisation (DESI) extends ESI to a surface analysis technique. In DESI, charged solvent droplets are sprayed at the sample to achieve both ejection and ionisation of intact molecules from the surface.

1.1.1 Matrix assisted laser desorption/ionisation

Matrix assisted laser desorption/ionisation (MALDI) is an extension of LDI and was the ionisation technique used to collect the majority of the data presented within this thesis. In MALDI, a matrix molecule (typically a low molecular weight organic acid) is mixed with, or deposited on top of, the sample causing co-crystallisation of the analyte and matrix molecules. A laser is then used to irradiate the matrix-analyte crystals, causing

an ablation event to desorb and ionise intact molecules. A significant benefit of MALDI when compared to other ionisation techniques is the ability to desorb and ionise a vast array of analyte classes in their singly charged state across a large mass range, ranging from drugs, through lipids and peptides, up to proteins.

The choice of matrix is often based on empirical evidence due to the complex mechanisms of ion formation. Although studies have been conducted to attempt to elucidate the ion formation process, it is still not fully understood [16, 17].

Ideal matrix compounds have strong optical absorption at the wavelength of the laser used and are often acidic, acting as a proton source to aid ionisation. The ideal matrix choice for a given experiment is also dependent on the analyte of interest, with certain matrices proving better at desorbing and ionising select molecular classes. For example, 2,5-dihydroxybenzoic acid (DHB) is often used for lipids, sinapinic acid (SA) is more commonly used for the detection of proteins and α-Cyano-4-hydroxycinnamic acid (CHCA) is frequently used for drugs and peptides.

The choice of matrix also influences the extent of metastable fragmentation [18]. Metastable fragmentation is a process where an ion with sufficient energy, after leaving the sample and entering the mass analyser(s) but before reaching the detector, fragments. Metastable fragmentation is not apparent in a linear instrument due to the parent and fragment moving with the same velocity, whereas a reflectron separates the parent and fragment ions [18].

Although MALDI is often cited as being a label free technique, in practice this comes with a caveat. While true that there is no requirement to specifically label molecules prior to analysis, the choice of both matrix and solvent system greatly impacts the capability to detect certain analyte types.

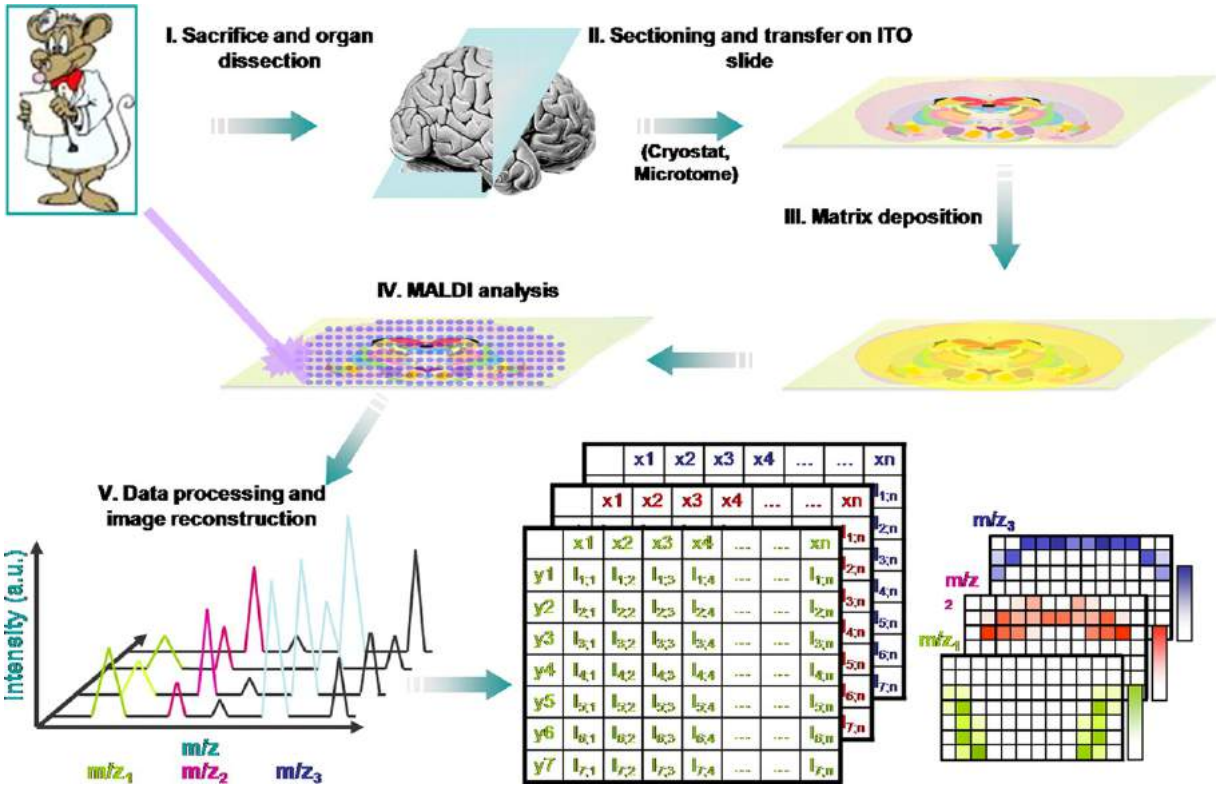


Figure 1.1: Overview of the sample preparation and analysis process for MALDI mass spectrometry imaging. Reproduced from [19]

1.2 Mass Spectrometry Imaging

Although some literature refers to this topic as "imaging mass spectrometry" (IMS), this acronym can easily be confused with ion-mobility spectrometry and so the term "mass spectrometry imaging" (MSI) will be used throughout. MSI is a technique where a sample is interrogated at spatially resolved locations resulting in a mass spectrum per location. The spatial distribution of ions can then be visualised by plotting the intensity detected at each spatial location. An overview of the typical MALDI MSI workflow is given in Figure 1.1 with steps I-IV discussed in Section 1.2.1 and step V discussed in Sections 1.6 and 1.7. An overview of the style of data produced in a MSI experiment is shown in Figure 1.2.

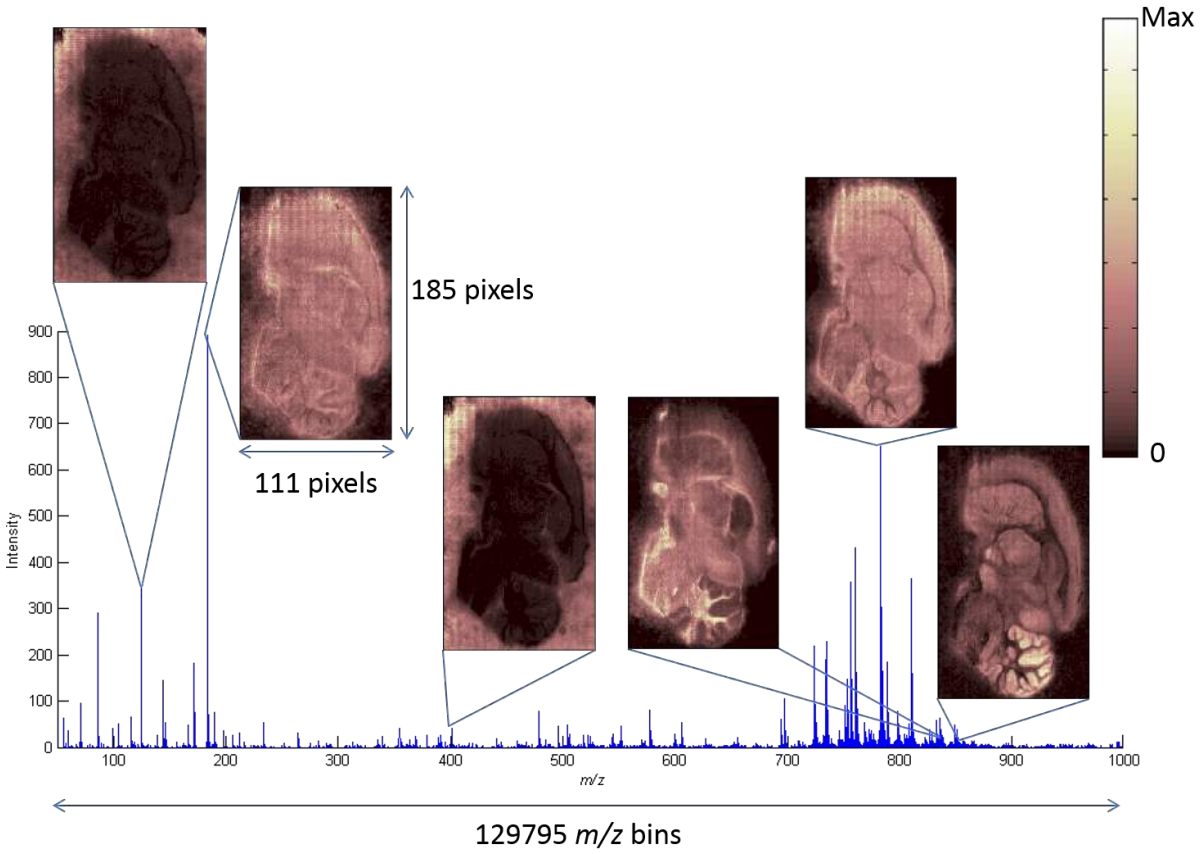


Figure 1.2: An overview of MSI data, where a spectrum is acquired at every discrete spatial location and an ion image is generated by plotting the intensity of a given peak at each spatial location to which a false colour scheme is often applied.

1.2.1 Sample Preparation

Sample preparation for most MSI techniques involves mounting a thin tissue section (10-20 μm) onto a substrate (which could be glass, ITO coated glass or stainless steel depending on the ionisation technique and instrument used). In the case of biological samples these could be either fresh frozen or fixed prior to sectioning [4, 20]. For techniques such as DESI, liquid extraction surface analysis (LESA) and secondary ion mass spectrometry (SIMS) imaging this is all the preparation required.

In the case of MALDI MSI, as when performing standard MALDI MS, a matrix must be applied to the sample. The ideal matrix application method will deposit a given volume of matrix homogeneously over the sample. The cheapest and easiest method is to use an artistic airbrush to coat the sample [21]. However, the homogeneity of the deposited layer is heavily dependent on the user. To remove the human error from this, a number of dedicated robotic sprayers have been developed [22] as well as modified inkjet printers [23]. When high spatial resolution is critical, sublimation has been shown to result in the smallest crystal size [24].

1.2.2 Spatial Resolution

Spatial resolution is heavily dependent on the ionisation method used. In the case of MALDI the limit is dependent the focusing capabilities of the laser and the matrix crystal size, in SIMS the focusing of the ion beam, in DESI the tip-to-surface distance, nebulisation gas pressure, solvent flow rate and solvent composition all affect the footprint of the solvent spray [25] and in LESA the microjunction size. Every reduction in pixel size comes at a cost of sensitivity and so there is often a trade off to be made between pixel size and sensitivity, with sensitivity being the major limiting factor.

In MALDI MSI, the size of the crystals formed plays a large role in the limit of spatial resolution that can be achieved. Crystals must be smaller than the pixel size or an inaccurate representation of the spatial distribution will be acquired. The matrix choice

and application method employed both affect the size of crystals formed [26].

It is often possible to increase the spatial resolution in MALDI MSI by employing 'oversampling' whereby pixels are overlapped with one another and at each pixel the sample is exhausted in the analysis process before moving onto the next pixel [27]. This means that signal is only detected from remains of the sample in the non-overlapping region. As the minimum movement of a stage is often smaller than the probe size, oversampling can provide significant spatial resolution improvements. Oversampling has been employed in other techniques such as DESI, however as the sample is not wholly consumed at each pixel location, signal is detected from a wider region than the specified pixel size.

1.3 Mass analysers

The purpose of a mass analyser is to separate ions in the gas-phase according to their mass-to-charge ratio. The choice of mass analyser usually involves consideration of speed, sensitivity, mass range, and mass resolving power with no one mass analyser optimal for all situations.

1.3.1 Mass Resolution and Resolving Power

When discussing the performance of mass spectrometers it is important to include information about the mass resolving power. The precise definition is one of the most confusing subjects of terminology used in mass spectrometry [28]. The IUPAC definition has evolved over time. Initially two separate methods of determining resolving power were presented, the first measured from a single peak consisting of single-charged ions of mass m, the second defined in terms of overlap between two peaks [29].

Later, 'resolution' was introduced and adopted the definition previously assigned to resolving power [30]. Mass resolving power was then defined as the ability to distinguish between ions differing in the quotient m/z by a small increment, characterised by giving the peak width expressed as a function of mass for fifty and five percent of the maximum

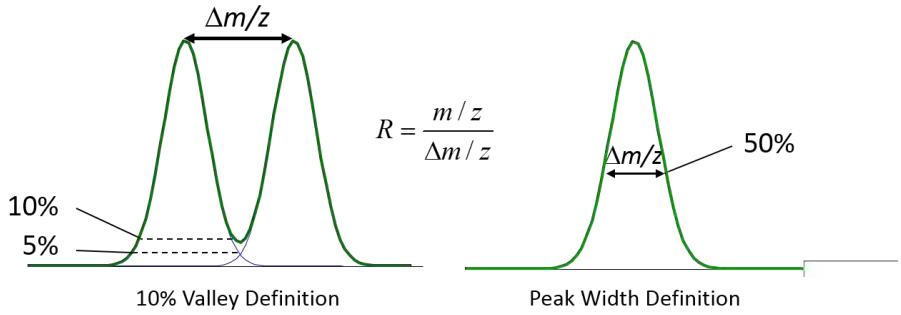


Figure 1.3: Two methods for determining mass resolution of a mass spectrometer, the 10 percent valley definition and the peak width definition.

peak height [30].

Most recently, resolving power has been defined as the measure of the ability to provide a specified value of mass resolution. Resolution has been defined as the observed m/z divided by the smallest difference $\Delta(m/z)$ for two ions that can be separated, as given in Equation 1.1, with the requirement that the procedure by which $\Delta(m/z)$ was defined and measured, and the m/z value at which the measurement was made must be reported [31].

$$R = \frac{(m/z)}{\Delta(m/z)} \tag{1.1}$$

Two methods for determining the resolution are provided, the 10 percent valley definition and the peak width definition shown diagrammatically in Figure 1.3. The 10 percent valley definition is the value of Equation 1.1 for two peaks of equal (or near equal) height at m/z and $m/z + \Delta(m/z)$ that are separated by a valley which at its minima is 10% of the lower peak height. The peak width definition is the value of Equation 1.1, where $\Delta(m/z)$ is the width of a single peak of a singly charged ion at a specified fraction (0.5, 5 or 50%) of the peak height. It is not always possible to find a pair of peaks matching the necessary criteria for the 10 percent valley definition and so all resolving power and resolution values calculated and quoted within this thesis use the peak width definition at 50%.

1.3.2 Quadrupole

Quadrupole mass analysers consist of four (or six, or eight, in the case of the hexapole and octupoles respectively) perfectly parallel cylindrical, or preferably hyperbolic, rods to which potentials are applied [28]. The oscillating electric field separates ions based on the stability of their trajectories through the mass analyser.

These can be operated in two modes, radio frequency (RF) only and as a mass filter. When operating in RF only mode, ions with a mass higher than a limit defined by the value of the RF voltage have a stable trajectory and thus the quadrupole acts as an ion guide. Higher masses suffer from lower transmission due to poorer focusing, the efficiency of which is inversely proportional to m/z. When operating as a mass filter a direct current (DC) voltage is applied with the RF voltage which causes ions which do not have a certain m/z (determined by the voltages applied) to have an unstable trajectory and so will not reach the detector.

Quadrupoles have superior power to focus ions towards the axis of the ion guide due to the shape of the potential well and thus have better transmission efficiency than the other multipoles. However, octupoles have a wider mass range for simultaneous transmission of ions than quadrupoles.

Multiple quadrupoles can be included in tandem, with quadrupole mass spectrometers stylised with Q and RF only quadrupoles with q. The instrument used to collect the majority of the data presented within this thesis, the QSTAR XL (AB SCIEX), is a QqTOF instrument. This consists of q0, an RF only mode quadrupole which focuses and transfers ions into the high vacuum region, Q1, which operates in RF only mode to transmit ions unless tandem MS is being performed where a resolving DC voltage is applied for mass resolving mode, q2 which is housed inside the collision cell and operates in RF mode at all times and then a time-of-flight tube (discussed in Section 1.3.3) [32].

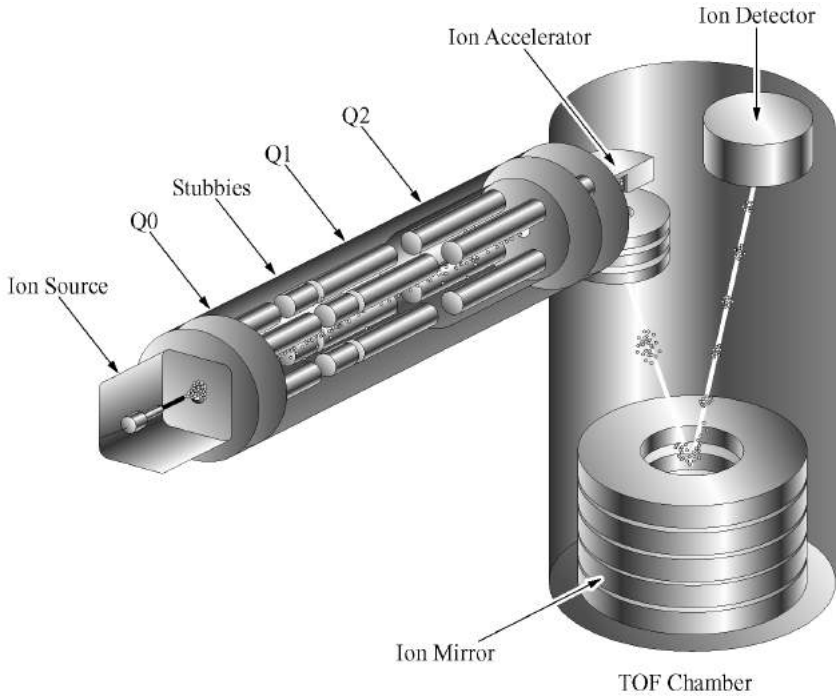


Figure 1.4: QSTAR XL ion path chamber. Reproduced from [32]

1.3.3 Time-of-flight (TOF)

In time-of-flight (TOF) mass analysers, ions are accelerated by an electric field and subsequently separated by their velocities as they traverse a fixed length flight tube. The relationship between time taken to reach the detector and m/z is given in Equation 1.3.

$$t^2 = \frac{m}{z}\left(\frac{L^2}{2eV_s}\right) \tag{1.2}$$

$$m/z = t^2\left(\frac{2eV_s}{L^2}\right) \tag{1.3}$$

where t is time taken, L is the length of the flight tube, e is the elementary charge and V_s is the voltage applied.

The inclusion of a reflectron, an electrostatic reflector, increases the flight path and therefore the resolution through increased separation of ions. However the more beneficial feature is the correction of the spread of kinetic energies that ions of the same m/z have when leaving the source, further improving the resolution [28]. The disadvantage of

including a reflectron is the reduction in sensitivity while also introducing a mass range limitation.

1.4 Data Formats

Software for visualising and processing MS and MSI data was originally limited to the software supplied by the corresponding instrument vendor. As each MS instrument vendor also has their own independent and incompatible data format, it is not possible to use a competing vendor's software to process or analyse data. This has implications when trying to compare data acquired on different instruments and also when sharing data with either collaborators or the community, by submitting to data repositories or as supplementary information alongside published articles. Vendor's formats are largely proprietary, eliminating them from consideration as a standard for sharing data as it would be difficult, if not impossible, to develop tools such as converters, data viewers and parsers without either full details of the format or support from the vendors. Open formats, on the other hand, enable anyone to develop software capable of reading the data. This therefore increases the amount of software available for processing data, removing the need to ensure that all collaborators have the same, often commercial and prohibitively expensive, software to view the data. Further, with the move towards standards on the reportable information, such as minimum information about a proteomics experiment (MIAPE) [33], it is important that shareable data formats cater for such details.

The simplest open format solution is to use human readable text (ASCII), however in situations where large arrays of number are stored, such as in mass spectrometry, it becomes impractical due to the inflated data size, in some cases an 18x increase in size when compared with the vendor's format [34]. The early open formats, JCAMP-DX (originally designed for infrared (IR) spectrometry but later adapted for MS [35]), netCDF and ANDI-MS suffered from low uptake due to being variably implemented, difficult to validate as well as relying on complex and rigid dictionaries which are not only

difficult to keep up to date but also prevent inclusion of vital features such as tandem mass spectrometry (MS/MS) scans [34, 36].

The next era of open formats (mzXML [34] and mzData [37, 38]) were developed simultaneously and independently, but had many similarities in their implementation. Both formats are open, vendor-neutral, capable of storing metadata, and store all information using XML. As XML is intended to be human readable, all of the data must be stored as text, reintroducing the inflated data size issues associated with ASCII. The solution to this was to store the binary data in Base64, while keeping the metadata in human readable XML. This results in a 33% increase in data size compared with binary storage, but a reduction of 325% when compared to using human readable plain text. To facilitate uptake, basic software tools were released as open source alongside the description of the mzXML format, including converters for native MS formats, parsers and data viewers [34].

Although mzXML and mzData are very similar in the handling of the data, they differ on their design philosophy of flexibility. mzData was designed to be extensible and flexible due to the inclusion of the external controlled vocabulary, whereas mzXML has a strict schema with information being described in enumerated attributes [39]. The strict schema simplifies software implementations and allows existing XML validators to be used to validate the format, whereas mzData requires custom validators due to the controlled vocabulary. However, this simplified implementation comes at a cost, as any alteration to the schema requires an update of any supporting software to be able to handle the new version of the format whereas mzData allows alterations of the controlled vocabulary without requiring additional software updates.

The existence of two similar formats that attempt to achieve the same goal caused unnecessary confusion within the community while also increasing the amount of work required of software developers by requiring support for both formats. This prompted the designers of mzData and mzXML, along with instrument vendors, researchers and software developers, to work together to develop a single, unified format for MS data,

mzML [39, 36]. The mzML standard aimed to combine the best features of mzXML and mzData while resolving the difference in flexibility design philosophies. This was achieved by retaining the controlled vocabulary for greater flexibility from mzData and the indexing mechanism from mzXML for enabling fast, random access to specific spectra without needing to read the entire file.

The release of the description of the mzML format also included validation tools, which considered the controlled vocabulary to ensure consistent use of the format, better support for selected reaction monitoring data and immediately available converters, which greatly increased adoption of the format [36]. It also included the option for spectrum-wise and whole file gzip compression to combat large file sizes. The framework for testing, as well as providing a reference implementation, has been supplied by the ProteoWizard software project [40].

Other formats include AnIML, which aims to be a format capable of handling any analytical data however it was first introduced in 2004 and has yet to have an agreed upon and published requirements specification, preventing uptake. The AnIML format does offer some benefits over existing formats, such as inclusion of an audit trail to log descriptions of any changes to the document [41].

Despite the advances made in the MS community for open standardisation, the MSI community largely ignored the proposed formats. Although the standards developed for MS are capable of storing MSI data, it is generally infeasible in practice. The primary reason for this is the lack of pixel coordinate information being incorporated in a given spectrum's metadata. Without such information, MSI software would require the user to provide additional input to be able to arrange the spectra in the correct spatial location to generate ion images. A further significant reason is that imaging data is typically large (easily capable of exceeding 10s of GB for a single image) and so the 33% increased size caused by the use of Base64 encoding makes a significant difference in both storage requirements and data access times. Although the mzML standard enables random access at the spectral level, it does not allow random access at the m/z, or image, level which

can increase image generation times.

With the adaptation of the MRI data evaluation software package BioMap (Novartis) [42] to MSI, Analyze 7.5 [43] quickly became the *de facto* open standard for MSI data. The Analyze 7.5 format was not specifically designed for MSI data, and instead was developed to be a common format for multi-modality imaging data. Because of this, it does not contain the capability to contain any metadata for the instrumentation used or data analysis performed meaning that it cannot conform to the MIAPE standard.

Römpp *et al.* [44] and Schramm *et al.* [45] designed the imzML standard to provide a means for the flexible and efficient exchange of MSI data between different instruments and analysis software. A single dataset stored in the imzML format consists of two files, one an XML file (containing all of the metadata relating to the imaging experiment) and the other a binary file (containing all of the spectral data). The flexibility is achieved from the use of the XML file, which essentially the same as the mzML format, only with additional controlled vocabulary (CV) parameters which cater for the imaging related attributes (such as pixel coordinate and spatial resolution). The efficiency is achieved through the use of the binary storage of the spectral data, reducing the size of the stored data by 33% when compared to mzML. Further, because the format is based on mzML it has the capability to conform to the MIAPE (as well as the proposed MSI MIAPE [46]) standard.

As the introduction of mzXML and mzML had previously highlighted, the inclusion of open source tools for converting, parsing, validating and visualising data in the new format greatly increased uptake as it reduced the amount of time required to implement or incorporate the new format. With the release of imzML, freely available visualisation software (DatacubeExplorer [47] and a modified version of BioMAP [42] to include an imzML parser) and a single converter (for Thermo Scientific RAW to imzML) were made available. None of these tools were made available as open source and the provision of a converter for only a single instrument vendor's format excluded users of other instruments and so initial uptake of the format was slow. Also, no validators were released with

imzML, the rationale being that the XML part of the format can be validated using the existing mzML validators. However, these validators do not take into account the extra CV parameters that are required to describe a valid imaging dataset and so only the structure could be fully validated, not the content.

As the imzML format and tools supporting it have matured, imzML has become accepted as the community standard for sharing MSI data. This has prompted instrument vendors (such as Thermo, Bruker and Waters) to support the format through export functions in their software, easing the initial difficulty in converting data to the open format. Additionally, the acceptance of imzML as a community standard has enabled important community driven projects, such as the release of benchmark datasets [48], development of metabolic atlases [49] and multicenter studies [50].

1.5 MSI Data Analysis Software

Each MSI instrument vendor supplies software for processing and visualising the data acquired on their instruments. A list of all of the preprocessing methods that are supported in each vendor's software is given in Table 1.1. As discussed in Section 1.4, each vendor has their own proprietary data storage format which is the only format that can be read into their software. This means that data acquired on a QSTAR XL (AB SCIEX) cannot be processed using Bruker's flexImaging. This prevents truly comparing data from multiple instruments using these software packages alone and limits the processing algorithms to those implemented within the instrument manufacturers software.

Prior to the introduction to imzML, the only freely available, vendor independent MSI software was BioMAP [42]. A significant limitation of BioMAP in the processing of large MSI data is the limit to the number of m/z channels that can be loaded for any dataset (32768, due to the number format used). Since the advent of imzML, a wide number of third party software packages have been developed and released as open source (MSiReader [51], Cardinal [52] and OmniSpect [53]), freely available (OpenMSI [54] and

DataCubeExplorer [47]), restricted to collaborators (Mirion [55]) or sold (SCiLS Lab [56], MALDIVision [57] and Quantinetix [58]). The preprocessing methods available in each of these (except MALDIVision and Quantinetix) are presented in Table 1.2. As most of these packages support imzML (with the exception of OpenMSI [54] and SCiLS Lab [56]) it is now possible to process data from any instrument that has a corresponding imzML converter, while simultaneously increasing the preprocessing methods available to the analyst through the choice of software. The drawback that still remains is that these software packages do not export processed data or partial results to a format readable by other software packages, meaning that all desired functionality must be included in a single software package.

1.6 Preprocessing Mass Spectrometry Imaging Data

Preprocessing is the most important step in the analysis of MSI data and is often the most overlooked. Any errors that are produced during this stage propagate through the entire analysis and can affect the appearance of a given ion's distribution as well as multivariate and clustering results in the final stage. A major problem is that there is no consensus on which methods are the most appropriate and this isn't helped by instrument vendors providing different algorithms for each stage, without a single common algorithm between them (see Table 1.1).

Coombes *et al.* [72] describe a model of a spectrum using Equation 1.4.

$$f(t) = B(t) + N * S(t) + \epsilon(t) \tag{1.4}$$

where $f(t)$ is the observed signal, $B(t)$ is the baseline, $S(t)$ is the true signal, N is a normalisation factor and $\epsilon(t)$ is the noise. Preprocessing methods aim to correct one or more of these artefacts and are used in combination to reveal the true signal.

Instrument Vendor	Software	Baseline Correction	Smoothing	Normalisation	Peak Detection
AB SCIEX	oMALDI / Analyst	None	Gaussian [59] 3 Weighted Points [59]	TIC [59] Mass Window(s) [59]	IWA [59]
	TissueView	Derivative [60]	User Defined Filter [60]	Mass Window [60]	
Bruker	flexImaging / flexAnalysis	Convex Hull [61] Top Hat [61] Local Median [62]	Savitzky-Golay [61] Gaussian [61] Chemical Noise [62]	RMS [61] TIC [61] Median [62] Mass Window(s) [61]	Centroid [62] SNAP [62] Sum [62]
Thermo	ImageQuest / Xcalibur	Curve fitting [63]	Moving Mean [63] Gaussian [63]	TIC [64] Mass Window [64]	Genesis [65] ICIS [65] Avalon [65]
Waters	High Definition Imaging / MassLynx	Curve Fitting [66]	Savitzky-Golay [66] Moving Mean [66]		Normal [66] Apex [66]

Table 1.1: Preprocessing methods available in MSI instrument vendor's software. TIC - total ion current. RMS - root mean square. Mass window equivalent to a single reference peak. IWA - Intensity-weighted average to calculate centre of gravity.

Software	Data Format	Baseline Correction	Smoothing	Normalisation	Peak Detection
BioMAP [42]	Analyze 7.5 [67] imzML	Derivative [67]	Sinc [67] Savitzky-Golay [67]	Reference Image [67]	
Cardinal [52]	Analyze 7.5 [68] imzML [68]	Local Median [68] Local Minimum [68]	Moving Mean [68] Gaussian [68] Savitzky-Golay [68]	TIC [68]	Simple [68] Adaptive [68] LIMPIC [68][69]
DataCubeExplorer [47]	Analyze 7.5 [70] Datacube [70] imzML [70]				
Mirion [55]	imzML [55] raw [55] udf [55]				Unknown Method
MSiReader [51]	Analyze 7.5 [71] mzXML [71] mzML [71] imzML [71]	msbackadj() [71]		TIC [71] Mass Window [71] Custom Data [71]	Parabolic Centroid [71] mspeaks() [71]
OmniSpect [53]	Analyze 7.5 [53] imzML [53] mzXML [53] netCDF [53]				
OpenMSI [54]	OpenMSI [54]				
SCiLS Lab [56]	Bruker	Unknown Method		TIC [56]	Unknown Method

Table 1.2: Preprocessing methods that exist in third party MSI software. The raw file format supported by Mirion is ThermoScientific data only. It is worth noting that the user manual for Mirion and SCiLS Lab were not easily available (without permission from the author and without purchasing respectively) and so the list of available functions may not be accurate.

1.6.1 Dead-time Correction

In the MCP detector installed in the QSTAR XL, any impact event that is detected at > 100 mV is counted as an ion [32]. Following this, there is a period of time (on the order of several nanoseconds) where any further impact events are not recorded, this is the 'dead-time' of the detector. Intense peaks become distorted by a reduction in the right side of the peak, shifting the centroid to the left (lower m/z) of the real value and suppressing the peak height. A number of publications have focused on tackling this issue in SIMS [73, 74, 75], however this issue is not isolated to SIMS and affects any instrument that utilises an multi-channel plate (MCP) as the detector.

To reduce the effects of dead-time, the detector in the QSTAR XL contains four collection anodes which are connected to independent time-to-digital converters (TDC). Thus, approximately a quarter of the ions will strike each anode, reducing the probability that an ion impact event will occur during a period of dead-time [76].

1.6.2 Smoothing

Smoothing is applied to mass spectra in order to compensate for small fluctuations (noise) in the signal ($\epsilon(t)$ in Equation 1.4). A comprehensive review of smoothing algorithms applied to mass spectrometry is provided by Yang *et al.* [77].

To calculate the new value of a data point in the smoothed spectrum, window based smoothing algorithms consider k data points either side, forming a 'window' of size $(2k + 1)$. All smoothing algorithms presented below are window based. The general form of weighted window based smoothing is given in Equation 1.5. The simplest is the moving mean where $w_i = 1$ for all i.

$$y[n] = \frac{1}{2k+1} \sum_{i=-k}^{k} w_i x[n-i] \tag{1.5}$$

where x is the spectrum, k is the window size and w is the weights vector to be applied.

In Gaussian smoothing, shown in Equation 1.6, the weights to be applied across a

window are calculated from a Gaussian function.

$$y(t) = \int_{-\infty}^{\infty} x(\tau) \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(t-\tau)^2}{2\sigma^2}} d\tau \tag{1.6}$$

where σ is the standard deviation of the Gaussian function.

Savitzky-Golay smoothing filters, shown in Equation 1.7, remove noise from the signal while retaining peak shape and height by local least-squares polynomial approximation [78]. These properties make Savitzky-Golay smoothing extremely attractive and has made it the most widely used smoothing algorithm in Analytical Chemistry [79, 80].

$$y[n] = \sum_{i=-k}^{k} h_{0,i} x[n-i] \tag{1.7}$$

where $h_{0,i}$ are elements of $\mathbf{H}$ which has size (N + 1) by $(2k + 1)$ and is calculated by

$$\mathbf{H} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T \tag{1.8}$$

where the matrix $\mathbf{A}$ of size $(2k + 1)$ by (N + 1) consists of the elements

$$\alpha_{n,i} = n^i \tag{1.9}$$

where $-k \leq n \leq k$ and $i = 0, 1, ..., N$.

1.6.3 Baseline Correction

The 'baseline' refers to an elevated intensity in one or more *m/z* regions that is not attributable to a resolvable ion peak. There is no agreed consensus as to the cause of the baseline artefact, however common theories include detector overload events and chemical noise from matrix fragmentation.

An estimate of the baseline is calculated and then subtracted from the spectrum using Equation 1.10.

$$y[n] = x[n] - B[n] \tag{1.10}$$

As with smoothing, multiple methods exist for baseline correction but limited discussion on which is appropriate in any given situation is present within the literature.

A simple estimate for the baseline is to calculate the minimum value within a window (similar to the window based smoothing algorithms presented previously), occasionally called the local minimum, as shown in Equation 1.11.

$$B[n] = \min(\{x[n-k], x[n-(k-1)], ..., x[n], ..., x[n+(k-1)], x[n+k]\}) \tag{1.11}$$

The minimum function can be replaced with other functions such as the median, which provides an estimate of the baseline at the level of the noise, shown in Equation 1.12.

$$B[n] = median(\{x[n-k], x[n-(k-1)], ..., x[n], ..., x[n+(k-1)], x[n+k]\}) \tag{1.12}$$

A more sophisticated method for determining the baseline is the TopHat method, which estimates the baseline from the mathematical morphology operations erosion followed by dilation (the combination of which is called opening) as shown in Equation 1.13. When applied to one dimensional signals, such as a spectrum, erosion is equivalent to the local minimum (shown in Equation 1.11) and dilation is equivalent to the local maximum.

$$B = x \circ S = (x \ominus S) \oplus S \tag{1.13}$$

where $\circ$ denotes the mathematical morphology operation opening, $\ominus$ denotes erosion, $\oplus$ denotes dilation and S is the structuring element, a vector where every value is 1 and its length is the window size.

1.6.4 Normalisation

Normalisation methods aim to remove spectrum-to-spectrum variability within a given image to reveal a more accurate representation of the underlying distribution of ions. A thorough investigation into commonly employed normalisation techniques was conducted by Deininger *et al.* [81]. These methods are also used to attempt to provide a quantitative value for the amount of a given ion at a given location, often in the form of a concentration. In reality however, these are only determining the amount detected at a given location due to the complexity of the sampling process and the possibility of interfering matrix effects. Normalisation coefficients, f, are calculated individually for each spectrum and are applied using Equation 1.14.

$$y[n] = \frac{1}{f(x)} x[n] \tag{1.14}$$

The most widely used normalisation method is total ion current (TIC) normalisation, also called Manhattan norm or ℓ_1-norm, where each spectrum is normalised to the sum of all counts within the spectrum given in Equation 1.15 where $p = 1$. Euclidean normalisation, or ℓ_2-norm, is given by Equation 1.15 where $p = 2$. In the case where $p \to \infty$, Equation 1.15 approaches the maximum norm, or the ℓ_∞-norm.

$$f(x) = \left(\sum_{n=0}^{N} |x[n]|^p \right)^{\frac{1}{p}} \tag{1.15}$$

Median normalisation, given in Equation 1.16, provides an approximate normalisation to the baseline of the spectrum. Median normalisation is robust to the effects of preprocessing methods such as smoothing [81]. Haqqani *et al.* used median normalisation to correct run-to-run variation in LC-MS label free proteomics [82]. In this study, only the peptides which were matched between all samples were considered and each spectrum was normalised to the ratio of the median intensity of the matched peptides in all runs to the median intensity of the matched peptides within the current spectrum.

$$f(x) = median(x) \tag{1.16}$$

Noise level normalisation, shown in Equation 1.18, builds on median normalisation to determine an estimate of the noise present within the data [81].

$$x'_i = x_i - x_{i-1} \tag{1.17}$$

$$f(x) = median(|x' - median(x')|) \tag{1.18}$$

One normalisation method that is common to the majority of the commercial tools presented in Table 1.1 is 'mass window', given in Equation 1.19. The choice of which mass window(s) to normalise to depends on the experiment at hand, but is often used in MALDI to normalise to matrix peaks [83] or a deuterated analogue of a molecule of interest [84].

$$f(x) = \sum_{m \in M} x[m] \tag{1.19}$$

where M is a vector containing the indices of the mass window(s).

Källback *et al.* compared the use of normalisation approaches for the quantification of different concentrations of imipramine (0-16 pmol) deposited onto a rat lung section [85]. In this study the normalisation to imipramine-D_3 (which was mixed with the MALDI matrix and applied to the sample) produced the best r-squared value (0.985) implying that this resulted in the closest to linear response between concentration and detected ion intensity. When performing logarithmic regression on the same data, RMS normalisation provided the best fit, with a correlation coefficient of 0.937.

1.6.5 Peak Detection

The aim of peak detection is to determine the location, height and widths of peaks (corresponding to ideally one, but often more than one, ion) within a mass spectrum. There are a wide variety of peak detection methods presented within the literature and within available software (see Tables 1.1 and 1.2 for a selection). The main reason that so many exist is that the described method incorporates one or more preprocessing methods (such as smoothing or baseline correction) as well as the peak detection step, and so in reality they are preprocessing workflows rather than unique peak detection methods.

In the case of the peak detection methods supplied in the vendor's software (and closed source software in general), without an accompanying document or publication that describes the details, it is impossible to replicate the results on data acquired on different instruments. As these methods determine which features are or are not recorded in the reduced dataset, this has a significant impact on the results of any subsequent processing such as multivariate analysis (MVA).

The principal peak detection method involves calculating the first derivative of the spectrum and looking for positive-to-negative zero-crossing points, which indicate the gradient was previously positive and then negative, corresponding to local maxima. The presence of noise in the data will increase the number of local maxima and thus increase the number of 'peaks' detected by this method. Applying smoothing methods prior to this gradient based peak detection method removes these false positives. The benefits of this method is that it doesn't require any prior knowledge to determine the peaks within the data.

A related method moves from left to right across the whole spectrum looking for a positive gradient change (or more simply, a change in intensity) exceeding a threshold to denote the start of a peak and then the inverse to signify the peak end point. The peak apex can then be identified as the maximum value within the calculated peak window.

Peak Filtering

With the peak detection methods described above, small fluctuations will remain in the data and so it is common to provide some limitations on the peaks that are retained. This can be in the form of a SNR threshold (with different methods of calculating it [77]), an intensity threshold, slopes of peaks, local maximum, shape ration, ridge lines, model-based criterion, peak width, or any combination of the above. The suitability of each of these restrictions depends on the peak detection method used.

Another method of peak filtering is to select a limited number of peaks, typically the x most intense [86]. This number is often a round number (e.g. 100), selected arbitrarily by the user, ensuring that it will produce data small enough for subsequent processing. This then will produce a determinable reduced data set size, allowing the user to ensure that the data is of a suitable size for subsequent processing, however the disadvantage is that the number of expected peaks is required to be known in advance. This is often not the case in exploratory MSI experiments. Furthermore, the most intense peaks can often be uninteresting (for example, the matrix peaks in some MALDI experiments) in terms of biological variance and so if the number of peaks chosen to be retained is chosen too low then the informative peaks may be omitted. Similarly if the number of retained peaks is selected to be an overestimate of the 'useful' peaks within the data then noise will be retained which could have a detrimental effect on further analysis.

1.7 Dimensionality Reduction

MSI experiments can produce extremely large datasets, for example a 4 cm $\times$ 4 cm MALDI target plate imaged at a resolution of 100 μm results in 160k pixels and if 100 kB per spectrum is assumed (6400 m/z-intensity pairs) then the dataset would be approximately 15.26 GB. Imaging the same area at high resolution (10 μm)[87] would result in over 16M pixels, and a potential raw data size of approximately 1.49 TB (1529 GB). Several applications require imaging to be performed on even larger sample areas, such as whole

animal sections [7, 2, 88] or 3D volumes [89, 90, 91, 92, 93, 94, 95, 96, 97].

The data size of a single spectrum is dependent on the mass resolving power of the instrument and the mass range of interest. Instruments such as the MALDI-FTICR can have mass resolutions of orders of magnitude greater than that of MALDI-TOF instruments and so when acquiring data over the same m/z range, can produce significantly more data [3, 98, 99]. A common way to reduce the size of data stored per spectrum is to store the data as m/z-intensity pairs rather than storing a value at every possible m/z location to reduce the amount of redundancy. This then introduces a further variable which determines the size of a single spectrum, the number of species detected (which is, in turn, a function of the sample type, the ionisation efficiency and/or the degree of fragmentation).

The problem with large secondary ion mass spectrometry (SIMS) datasets has been tackled by compressing the data to ensure that it fits in RAM [100, 101], but data size limits will continue to exist for algorithms which cannot utilise such compression and will again become problematic as improvements in imaging technology further increase the data size. Alternative data reduction strategies aim to isolate only peaks of interest and eliminate noise to reduce the amount of uninformative data used in further analysis [77, 86]. However, hundreds or thousands of peaks can be detected, and so further reduction is often necessary. Principal component analysis (PCA) is a mathematical technique that can be used to solve this problem by reducing dimensionality while retaining variance within the data [102].

When performing PCA via conventional means, the following steps are typically followed. (i) The dataset (N pixels, M peak intensities) is read into RAM as an $N \times M$ matrix. (ii) The mean spectrum (over the whole dataset) is subtracted from each spectrum in the dataset. (iii) Singular value decomposition (SVD) of the data matrix is then performed to determine eigenvalues and eigenvectors (also referred to as the loadings or coefficients in PCA). (iv) The data are then projected onto the space defined by the eigenvectors to determine the scores. This is exactly how the often used implementation of

PCA *princomp* (as supplied by MATLAB Statistics Toolbox) is performed. Implementations like these require the data matrix, along with multiple additional variables which are the same size as the data matrix, to be stored simultaneously in RAM in order to perform the full calculation. The finite size of RAM can easily be exceeded by the size of MALDI MSI datasets and this implementation severely restricts the size of the dataset that can be processed.

The memory limit can be reached through having a large number of peaks, a large number of pixels, or both and so a tradeoff has to be made that is dependent on the amount of RAM available for analysis [103]. Reducing the m/z dimension is commonly achieved through binning, however peak detection and alignment is a much more robust method of avoiding the loss of information while reducing the data [94]. Dependent on the size of the data, further reduction can be necessary prior to multivariate analysis, and so methods of selectively discarding peaks and pixels have also been developed [104, 86]. This reduces both spectral (potentially merging peaks) and spatial (potentially merging features) resolution to a point which may be deemed unacceptable [105]. An alternative method to solve the memory issues implicit in PCA is to utilise sparse matrix storage [103]. However, it is entirely possible that a data set in sparse matrix form is still too large to be stored in RAM and so in these cases some form of data reduction will still be required prior to the sparse storage such as removing intensity values below a user-defined threshold or performing binning spectrally or both spatially and spectrally.

All of these data reduction methods do provide a useful temporary solution, but as the move is made towards high resolution 3D data sets the problem will again return and so algorithms that are explicitly designed to handle large datasets will become more desirable. The importance of memory optimised algorithms applied to MSI data has been commented upon by Alexandrov and Kobarg [106].

Clustering techniques are becoming an invaluable tool in the processing and interpretation of MSI data sets and have been the focus of many recent articles [107, 102, 106, 97]. PCA has been shown to be a useful technique prior to clustering due to the reduced di-

mensionality and noise suppression [107, 102], however if meaningful peaks are discarded during the data reduction step prior to PCA then the clustering process may provide suboptimal results. In the case of MALDI MSI the largest variance is typically between matrix and analyte regions which will therefore be represented in the first few principal components. If data reduction techniques have been used to remove the matrix pixels and peaks then this will no longer be the case and the first few principal components will instead describe the largest variance within the analyte. PCA is also commonly employed as an unsupervised technique to objectively determine trends within the data [108, 109, 110, 111, 112, 113], which again may result in suboptimal results if the initial data is incomplete. This use of PCA is often considered to be controversial due to the inability to relate negative principal component loading values to experimental m/z signals and so other multivariate analysis techniques such as probabilistic latent semantic analysis (pLSA) and non-negative matrix factorisation (NNMF) have been applied to MSI data [105].

1.8 Parallelisation

As Section 1.6 described, preprocessing is an important part of any mass spectrometry imaging processing workflow. Each spectrum within a given imaging dataset is processed using the same set of operations to remove experimental noise and artefacts. This is typically performed by sequentially applying each operation on each spectrum. However, as the processing of each spectrum is an isolated event, there is no requirement for sequential processing and so the processing can be parallelised.

1.8.1 CPU Architecture

The central processing unit (CPU) is designed to work well with a vast array of applications. With the introduction of dual-, quad- and hexa-core processors it is now possible to perform 2, 4 or 6 separate instructions simultaneously. As each core acts independently

it is possible to perform a different action or process on each core. This is called multiple instruction, multiple data (MIMD).

1.8.2 GPU Architecture

Graphics processing units (GPUs) on the other hand can be comprised of hundreds of cores, operating under single instruction, multiple data (SIMD), where the same action is performed on multiple pieces of data. This limits the algorithms that are applicable to general purpose computation on graphics processing units (GPGPU) to those that require the same operation performed on large amounts of data. As the same preprocessing workflow is applied to every spectrum, this maps well to MSI preprocessing.

1.8.3 Things to consider when comparing CPU and GPU

Care should be taken when comparing performance times of CPU and GPU, with consideration of the CPU and GPU used, how heavily optimised each implementation is and the GPU memory transfer time [114]. Single precision (32-bit) operations are often at least twice as fast as double precision (64-bit) operations [115] and so both algorithms should make use of the same number format for a fair comparison.

The choice of CPU and GPU is of great importance, as the comparison of a high end GPU to a mobile CPU is suboptimal because they are designed for completely different purposes and therefore have different considerations for their operating power and thermal envelope [114].

It is important to include data transfer times to and from the GPU as it can contribute a significant amount of time to the total processing time. In some cases, this can exceed the amount of time required for execution of the CPU algorithm making the GPU implementation prohibitively costly regardless of the speedup achieved in algorithm design and optimisation. Omitting this value artificially inflates the speedup achieved because in practice the data still needs to be transferred to utilise GPGPU.

1.8.4 GPGPU in MS(I)

GPGPU has been employed in MS to provide 8-26 fold speed improvement to MS library searching [116], metabolomic fingerprinting [117] and peptide scoring [118]. The most significant published speed increase through the use of GPGPU is 200x when performing feature detection in proteomics data [119]. However, as with many comparisons, this is slightly overestimated, as it comes from comparing 2 GPUs to a single CPU and does not make clear whether the CPU code was multithreaded.

Kobarg *et al.* [120] developed a GPU implementation of a wavelet transform for preprocessing MALDI MSI data. This resulted in a ≈180x speed increase when compared to a CPU implementation, reducing the processing time for their test dataset from several hours to several minutes which is approaching the limits posed by clinical routine requirements.

Jones *et al.* [121] performed preprocessing, peak detection and reduction to a datacube using CPU based methods before transferring the reduced data to the GPU to perform PCA, NMF and pLSA. The speed increases achieved for NMF ranged from 1.5x to 13.2x depending on how well the number of pixels and the number of peaks retained mapped to the number of cores and memory available on the GPU.

1.9 Introduction to this Thesis

This chapter has provided an introduction to mass spectrometry imaging, data formats, visualisation and analysis software as well as preprocessing and multivariate analysis techniques used. The remainder of the thesis is split into six further chapters which focus on efficient methods for handling the extremely large datasets that are being routinely collected as part of MSI studies and methods for exploring, investigating and displaying the information captured within.

In Chapter 2 software for the conversion of MSI data acquired on any mass spectrometer to the open imzML format is presented. Extensible software for the visualisation,

preprocessing and multivariate analysis of spectral imaging data is also presented, with an investigation into the effects of preprocessing methods on MSI data.

Memory efficient methods for generating ion images, preprocessing entire datasets and performing PCA without requiring the whole dataset to be loaded into memory are presented in Chapter 3. These methods enable even the largest MSI data to be processed using a standard computer while also increasing the number of features, or peaks, retained in the analysis.

In Chapter 4, two algorithms for determining a hierarchical composition, one spatially and one spectrally, of MSI data are presented. These methods are then used to provide greater insight into the composition of a sample when compared to PCA used in isolation.

The way in which imaging data are presented can have a significant impact on the perceived structure, especially when using false colour to display images. This is of great significance in MSI as the majority of data analysis and interpretation is performed by manual inspection of significant ion images. In Chapter 5 the different colour schemes used to present MSI data within the literature are surveyed and applied to the same ion image, highlighting the effect this has on subsequent interpretation of the data. A means for evaluating the perceptual linearity of a given colour scheme is presented and used to evaluate the colour schemes found within the literature.

The software and algorithms developed and presented throughout the thesis were used to analyse data acquired from traumatic brain injury models, the results of which are presented in Chapter 6. This experiment consisted of four MS images which required processing together, resulting in a large combined dataset. Memory efficient methods from Chapter 3 were employed to enable processing and dimensionality reduction of the data. The hierarchical methods presented in Chapter 4 were then used on the reduced data to identify trends within the data.

Finally, Chapter 7 provides discussion on potential future direction for the work presented in each chapter of this thesis.

1.9.1 Publications Arising from this Thesis

- The imzML converter presented in Chapter 2 has been published as: 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.

- The memory efficient principal component analysis algorithm presented in Chapter 3 has been published as: 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 (15), 7146-7153, 2013.

- The work presented in Chapter 5 has been published as: 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.

CHAPTER 2

UNIVERSAL DATA CONVERSION AND ANALYSIS SOFTWARE

2.1 Introduction

This chapter presents software tools for the conversion and analysis of mass spectrometry imaging data acquired on any instrument. These tools are then used to investigate the effects of preprocessing methods on imaging data. Parallelisation of a common baseline correction technique, TopHat, is evaluated by comparing two different implementation algorithms on multiple hardware configurations.

2.2 Data Formats for MSI

Converters translate data stored using a certain specification (a format) into another. Tools exist for the conversion of most mass spectrometry formats to mzML. To compliment this work rather than repeat it, imzMLConverter was designed to use mzML as an intermediary format between the proprietary vendor's formats and imzML, the interface of which is shown in Figure 2.1. This relies on tools such as msconvert (part of ProteoWizard [40]), which can convert data from all major instrument vendor's proprietary formats to mzML.

There are three different ways that mzML file(s) can be generated from imaging data,

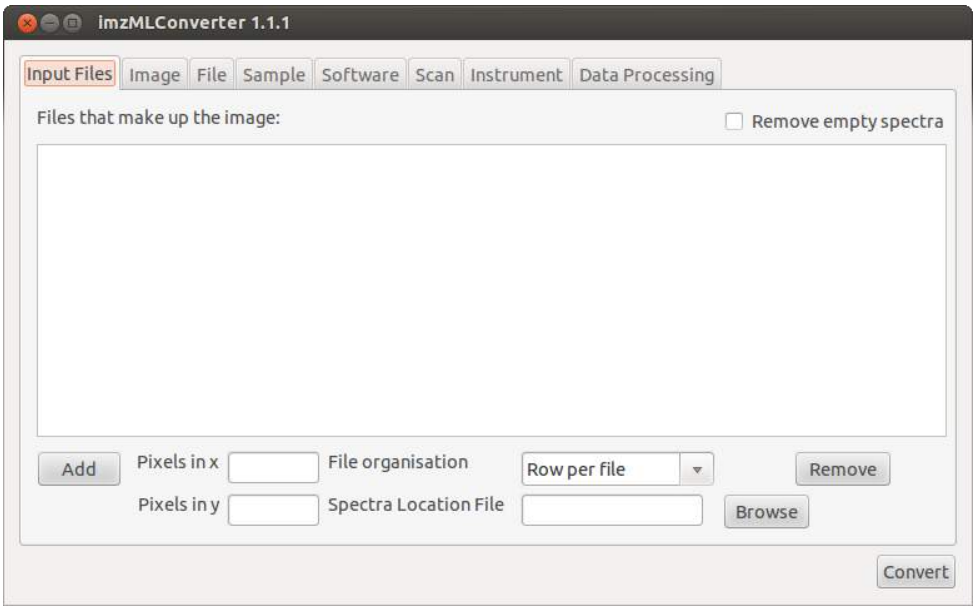


Figure 2.1: imzMLConverter 1.1.1 graphical user interface.

which depend on both the instrument and the imaging parameters used to acquire the data. The first is a mzML file for every row within the image. This option occurs when images are acquired in raster mode, where the sample holder is moved under the sampling probe continuously for each row and pixels are spectra are acquired at fixed time intervals, such as in desorption electrospray ionisation (DESI) and raster mode MALDI. The second is a single mzML file which contains all of the spectra that make up the MS image. The final way is for each spectrum in the image to be stored in its own mzML file. Options in imzMLConverter exist for each of these situations by selecting the 'File organisation' to be 'Row per file', 'Image per file' and 'Spectrum per file' respectively.

The final piece of information required to be able to reconstruct an image from the mzML file(s), is the relative spatial location for each of the spectra. In the case of raster mode imaging where the the imaged area was rectangular, this is trivial and the image dimensions can be automatically detected from the mzML files (number of pixels in x is the number of spectra in each mzML file and the number of pixels in y is the number of mzML files). It is equally trivial when a spot mode image has been acquired over a rectangular area, however the user must define the image dimensions. However, it is becoming increasingly common for instruments to allow the user to select arbitrarily

shaped regions of interest to acquire data from. Two such instruments that allow this are the ultrafleXtreme (Bruker Daltonics) and the Synapt G2S (Waters). The pixel location for each spectrum in an image acquired on an ultrafleXtreme can be determined from the meta information stored in the mzML file, where each spectrum has an associated name with the format of '0_R00X170Y127' (for region of interest 00, x coordinate 170 and y coordinate 127). Setting up an imaging experiment on the Synapt G2S produces a '*.pat' file, an XML based file containing the start and end coordinate of each row of the image and the pixel size. This information can then be used to assign a spatial location to each spectrum in the image.

2.2.1 Compression

One feature of the mzML specification that is omitted from all other available imzML exports or converters is the use of compression. One of the main benefits cited for the use of imzML over other open formats such as mzML is the disk space savings. As shown in Table 2.1, imzML only beats compressed mzML when it too is compressed. The use of compression makes the format more usable for its other major goal, data sharing. The reduction in disk space saves costs in both data storage and data transfer, especially when considering transfer across a network such as the Internet. The optimal choice between compressed or uncompressed data depends on the storage medium and processor used. A solid state disk (SSD) is sufficiently fast in returning data that it would be slower to read a smaller amount of data and decompress than it could be to read the uncompressed data. This is also true for HDDs arranged in RAID configurations, as demonstrated in Figure 2.2. As the transfer time of the storage medium used increases, the benefit of using compressed data begins to outweigh uncompressed in both size and response time, as is shown in the random access, uncompressed data stored on a hard disk drive in Figure 2.2.

Raw Data Format	Raw Size	mzML Size	Compressed mzML Size	imzML Size	Compressed imzML Size
SCIEX (*.wiff)	773 MB	11.50 GB	5.04 GB	8.67 GB	3.80 GB
Waters (*.raw)	3.13 GB	8.38 GB	1.93 GB	6.30 GB	1.46 GB

Table 2.1: Data size of typical MS image as acquired from a QSTAR XL (SCIEX, *.wiff) and a Synapt G2 (Waters, *.raw) in the corresponding proprietary format as well as the open formats mzML and imzML (both with and without compression).

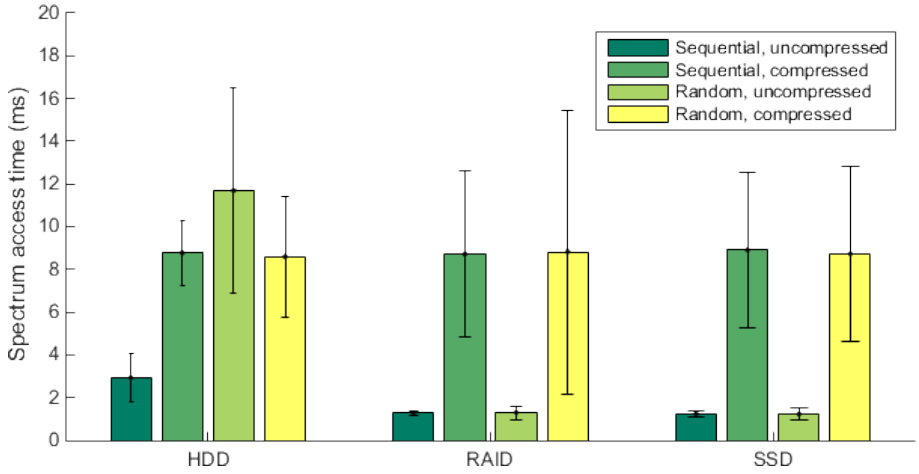


Figure 2.2: Data access times averaged over accessing 1000 spectra either randomly or sequentially from the SCIEX dataset from Table 2.1 stored in compressed and uncompressed imzML. Data stored on either hard disk drive (HDD), redundant array of independent disks (RAID) or solid state disk (SSD). HDD used was 5400 RPM. RAID consisted of two 7200 RPM HDD. SSD was 6 Gbps.

2.2.2 Handling Multiple Spectra per Pixel

Liquid extraction surface analysis (LESA) is a powerful technique capable of analysing complex samples without the need for any sample preparation. The process involves a small (on the order of 1 μL) amount of solvent being aspirated into a pipette tip. The pipette is then positioned at the location of interest, a droplet of solvent is formed on the tip of the pipette and brought into contact with the sample forming a microjunction to extract material from the sample. The droplet is then retracted into the pipette and subsequent electospray ionisation (ESI) mass spectrometry is performed using a mass analyser of choice. Commercial robots for automating the process have been developed, which provide the opportunity to sample at spatially resolved locations, resulting in an image.

When acquiring LESA imaging data on an Orbitrap there are two ways to set up an experiment. One is to acquire all data within the same file resulting in a single chromatogram for the whole imaging experiment, the other is to acquire each injection (and thus each pixel) in separate files. In both cases, a single pixel is comprised of multiple spectra each from an individual scan. Currently no tool exists for generating ion images from these data. To be able to do that the individual scans from each pixel must be combined together (either by summing or by taking the mean) and subsequently converted to imzML.

A tool developed in MATLAB for performing this on an image acquired in a single chromatogram is shown in Figure 2.3. The pixels can be identified as the 'peaks' that are present within the chromatogram (a series of scans where ions were detected) and the baseline is the time between pixels (corresponding to the time where the robot is collecting the solvent and sampling). Due to slight differences in the time taken by the robotic sampler to move to the required position and perform the analysis for each spatial location, no constant pixel offset can be determined which prevents easily determining the the start and end point of each pixel. Assuming that the total ion current in the analyte spectra is sufficiently higher than that of the blank spectra inbetween acquisitions, the

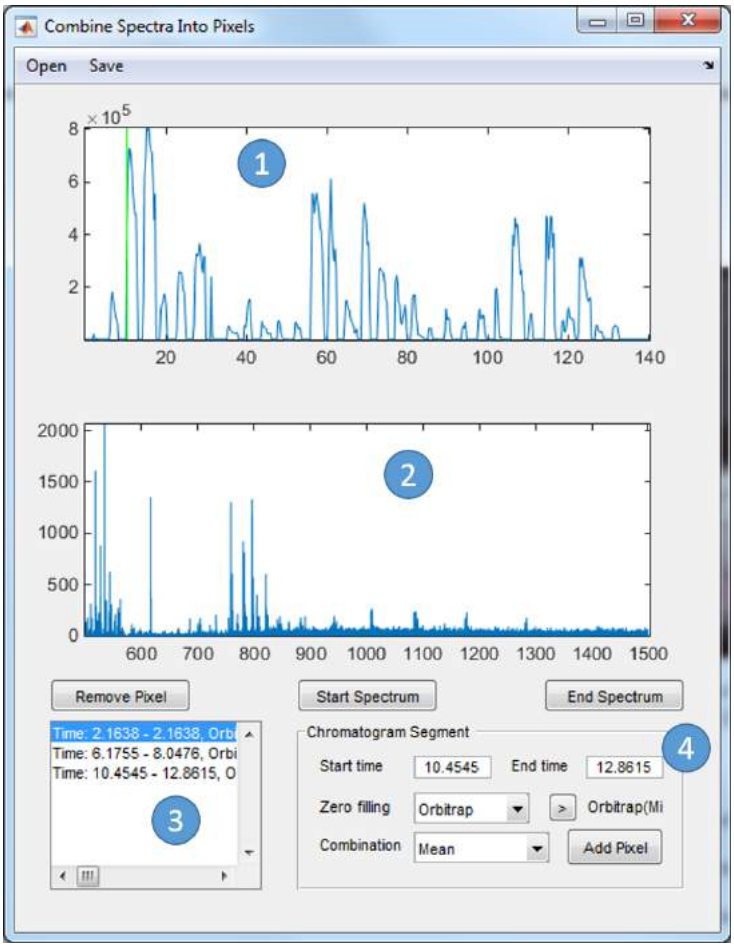


Figure 2.3: Tool for generating single pixel spectra by summing or averaging multiple scans acquired from multiple LESA injections stored in a single chromatogram. 1) Chromatogram. 2) Spectrum at the currently marked point in the chromatogram. 3) List of manually entered pixel start and end times. 4) Options for combining spectra to form a pixel.

pixels can be automatically detected by thresholding the chromatogram at just above the baseline level. In some analyses the microjunction can fail resulting in no solvent being sprayed and therefore no signal detected above the background. If this occurs then the user must manually select the start and end point of the pixel.

Once each pixel has been assigned all of the associated scans, the scans must be combined into a single spectrum by either calculating the mean or sum. To be able to do this, the spectra must have the same m/z axis. The reasoning and methods for achieving this are discussed in detail in Section 2.5.1. Scans are combined into a single spectrum by first applying the chosen method for ensuring a consistent m/z axis to each and then

calculating the mean. This spectrum is then written to an mzML file. The process is repeated for all pixels, updating the mzML file with each addition. The resulting mzML file can then be converted to imzML using imzMLConverter as described in Section 2.2 and further processed in any compatible MSI software.

2.3 Combining Datasets

The ability to combine multiple datasets acquired separately but which together form a single experiment is extremely powerful. Consider the case presented in [4] where different sample preparation methods are being compared but the data were collected as separate mass spectrometry images. To compare these data the analyst would have to load an image, perform any preprocessing necessary, search for an ion image of interest, then repeat for any other dataset being compared. Then the analyst would have to ensure that the intensity scales that the images were presented on were comparable prior to any interpretation. This is a laborious and time consuming process and would have to be repeated for each ion image that was investigated. While this is manageable for an experiment only consisting of two MS images, when trying to perform this on 14 serial sections, such as the data presented in [21], it becomes infeasible.

A feature included within imzMLConverter is the ability to tile and combine multiple imzML files together into a single imzML file, as shown in Figure 2.4, where two separate MSI datasets of sagittal sections of rat brain with different sample preparation methods applied (described in detail in [4]) are tiled horizontally. This then enables the analyst to rapidly compare the spatial distributions and relative abundances of ions in the visualisation software of their choice without needing to open multiple datasets and manually ensure colour schemes and intensity ranges of each ion image generated for each dataset are comparable.

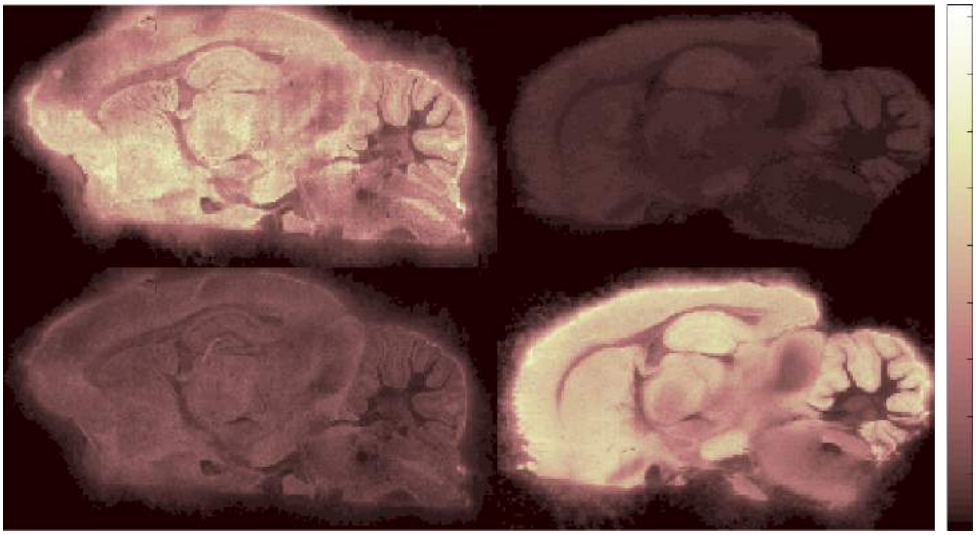


Figure 2.4: Replication of Figure 2a from [4], generated by combining two separate imzML files together and then generating two ion images (PC 32:0 $[M+K]^+$ m/z 772 top row and PC 32:0 $[M+Na]^+$ m/z 756 bottom row).

2.4 Software for analysing MSI Data

SpectralAnalysis was developed in MATLAB, using imzMLConverter as a parser for handling of imzML data, as a universal visualisation and analysis tool for mass spectrometry imaging data, the interface of which is shown in Figure 2.5. The rest of the chapter discusses algorithms which have been implemented into SpectralAnalysis to provide a single analysis tool for MSI data which is capable of not only reading in the data, but visualisation, exploration, preprocessing and multivariate analysis. Such a wide collection of capabilities does not currently exist in any currently available software, where there is a limitation on the instruments supported [56], no preprocessing methods are available [47] or no multivariate analysis algorithms are implemented [51].

2.5 Preprocessing

The purpose of preprocessing is to remove artefacts introduced during the data acquisition stage, to make spectra comparable to one another and to improve the efficacy of peak detection routines. The common preprocessing methods applied in mass spectrometry are

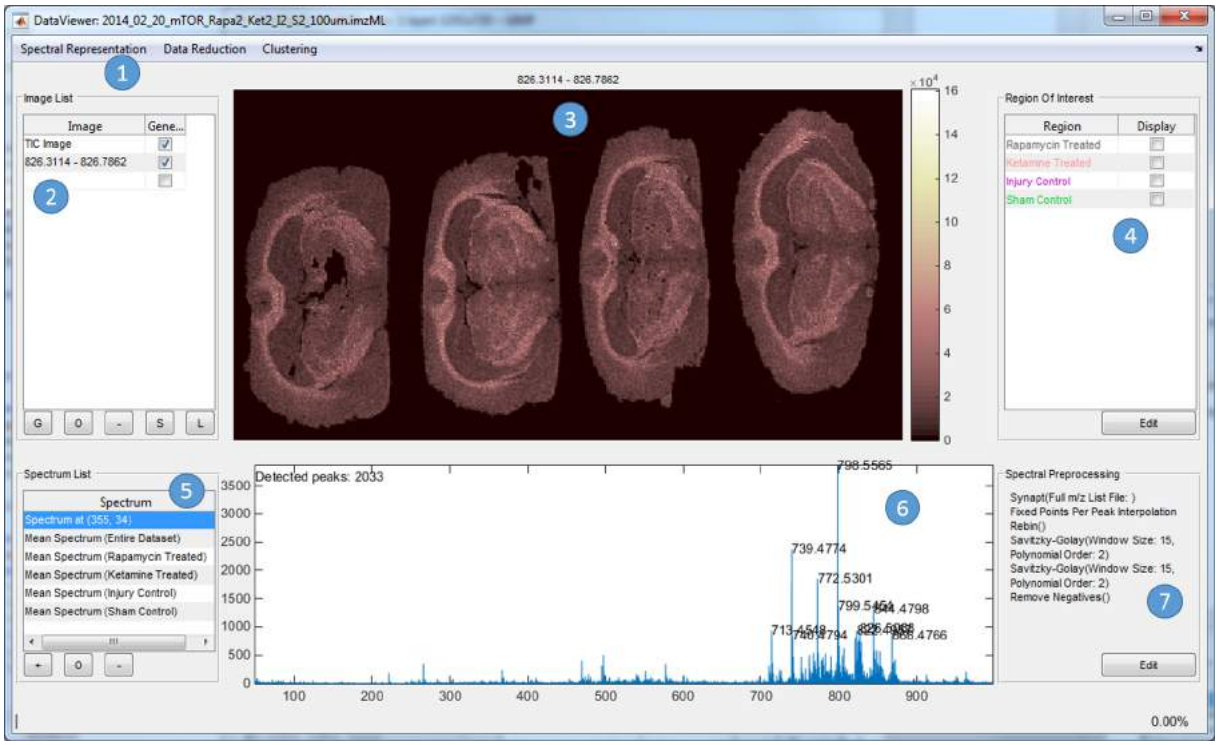


Figure 2.5: Screenshot of SpectralAnalysis interface. 1) Postprocessing options, discussed in Section 2.7. 2) List of previously generated ion images. 3) Current ion image being displayed, with options to generate from a previously saved list or overlay, discussed in Section 5.3.2. 4) List of created regions of interest. 5) List of generated spectra, with options to overlay. 6) Current spectrum view, with peak detection turned on labelling the top 10 intensity peaks out of the 2033 detected, discussed in Section 1.6.5. 7) Preprocessing workflow applied to every spectrum, discussed in Section 1.6.

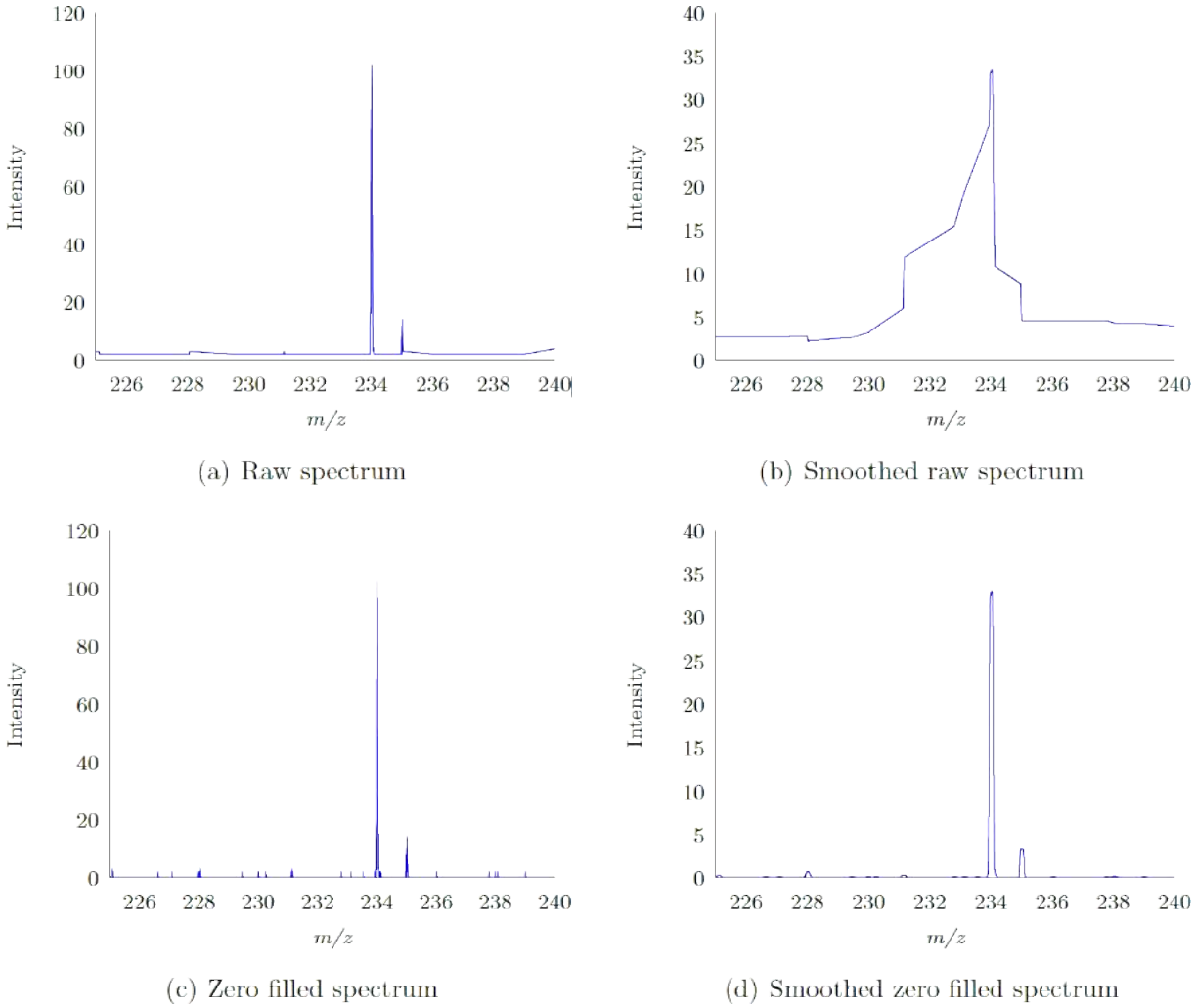
smoothing, baseline correction, normalisation and peak detection. Here we also include commentary on an additional step which is not often discussed, methods for ensuring a consistent m/z axis across a dataset.

The order in which the preprocessing methods are applied has an effect on the resulting data. The widely accepted order for preprocessing is smoothing or denoising followed by baseline correction prior to peak detection [69, 122, 86]. Noise reduction or removal methods such as baseline correction and smoothing aim to improve the peak detection method of choice. In some cases, such as wavelet based techniques, they are implicitly performed.

2.5.1 Ensuring a consistent m/z axis

When comparing, averaging or otherwise mathematically manipulating two or more spectra it is important that they are represented by the same number of m/z bins with the same m/z intervals. This is a common requirement in data reduction routines, where one or more summary spectra, such as the mean spectrum, are used for feature detection [123, 86] or peak alignment [106]. This also has the benefit of enabling spectra to be directly stored as a matrix (a 2D matrix as required for many post processing techniques such as PCA or a 3D 'datacube' for efficient image generation and manipulation).

Some mass spectrometers store the entire raw data for every spectrum and so this then becomes an optional data size and/or noise reduction step. Generally, however, the raw data is stored in sparse form as (m/z, intensity) pairs with zero intensity values omitted, to reduce the size of the data without loss of information. In some cases, zero values either side of recorded value sequences are also stored to remove visualisation artefacts present when they are omitted show in Figure 2.6a. If the data were to remain in sparse form, any window based function (discussed in more detail below) would corrupt the data (as these assume consistent intervals) shown in Figure 2.6.



(a) Raw spectrum

(b) Smoothed raw spectrum

(c) Zero filled spectrum

(d) Smoothed zero filled spectrum

Figure 2.6: Spectrum displayed in sparse format (a), where peak shapes are distorted and artificial baseline is displayed but is not part of the data, which is then corrupted through application of window based smoothing (b). Spectrum displayed with zero values replaced (c) and correctly smoothed (d).

Rebinning

The simplest way to both remove sparsity and to ensure a common axis for all spectra is to generate a new m/z axis with equally spaced m/z values, as shown in Algorithm 2.1. To then project any spectrum onto the new axis, data points that fall between adjacent m/z values on the new axis are summed together to provide an intensity for that interval (commonly called a bin) in the new spectrum, as shown in Algorithm 2.2. The result of this applied with varying bin sizes is shown in Figure 2.7. The bin size chosen can have a number of effects on the data. Selecting a large bin size results in significant denoising but at the cost of broadening the peak width and distorting the peak shape, as demonstrated in Figure 2.7b. Selecting a bin size that is smaller than the sampling frequency of the instrument can result in artefacts such as the introduction of zero values as shown in Figure 2.7d. In this example, the rebinning has essentially split the original peak into three peaks and depending on the subsequent preprocessing methods chosen this could either be rectified or exacerbated resulting in the detection or artificial peaks.

Algorithm 2.1. Rebinning: generate new axis

Require: Minimum, m_{min}, and maximum, m_{max}, m/z values for new m/z axis
Require: Bin size Δm
1: Generate new m/z axis $M_{\text{rebin}} \leftarrow \{m_{\text{min}}, m_{\text{min}} + \Delta m, m_{\text{min}} + 2\Delta m, ..., m_{\text{max}}\}$

Algorithm 2.2. Rebinning: apply new axis

Require: Spectrum S with m/z axis M
Require: New m/z axis M_{rebin}
1: $j \leftarrow 0$
2: $h \leftarrow \Delta m/2$
3: Create new spectrum S' to have same size as M_{rebin}
4: **for** $i \leftarrow 0$ to $|S| - 1$ **do**
5: **if** $M(i) < m_{min}$ **then**
6: continue
7: **end if**
8: **if** $M(i) > m_{max}$ **then**
9: break
10: **end if**
11: $t_1 \leftarrow M(i) - h$
12: $t_2 \leftarrow M(i) + h$
13: **while** $j < |M_{\text{rebin}}|$ and $M_{\text{rebin}}(j) < t_1$ **do**
14: $j \leftarrow j + 1$
15: **end while**
16: **if** $j < |M_{\text{rebin}}|$ and $M_{\text{rebin}}(j) < t_2$ **then**
17: $S'(j) \leftarrow S'(j) + S(i)$
18: **end if**
19: **end for**

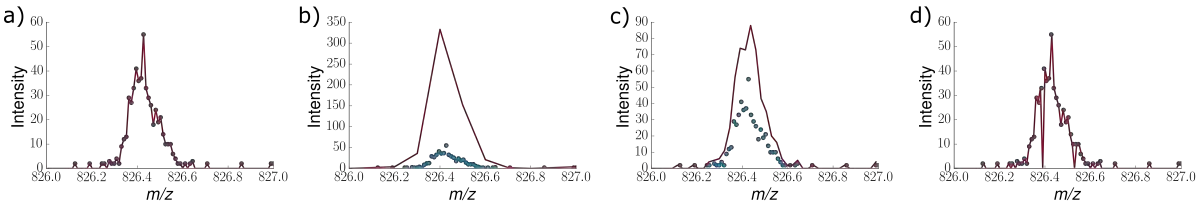


Figure 2.7: The effects of different bin sizes applied to the peak at m/z 826 in a single spectrum. The recorded data from the mass spectrometer shown as blue circles. Red line shows the standard representation of a mass spectrum (linear interpolation between each data point as a guide for the eye) with a) the raw data b) data binned at 0.1 m/z c) data binned at 0.023 m/z d) data binned at 0.01 m/z.

Interpolated Rebinning

The next method again generates an axis with equally spaced m/z values (using Algorithm 2.1), however instead of summing data points that fall within the bins, interpolation is performed. Multiple interpolation methods exist, the effects of which are demonstrated in Figure 2.8. Quadratic interpolation distorts the signal severely, especially in the case where Δm is smaller than the detector sampling intervals (in this example when $\Delta m = 0.01$), and so is inappropriate for this style of data due to the heavily modified peak shape. The differences between the spectra acquired using the other interpolation methods are more subtle.

This method has a number of advantages over the simple method above, at the slight cost of computation speed. The peak heights in the new spectrum are more comparable to that of the raw spectrum, which may be important for quantification studies. Also, the dropping to zero and splitting of peaks artefact is no longer present in the smallest bin size and instead the new spectra now closely match the the raw data.

The caveat of this method is that if the input data does not accurately capture the shape of all peaks within the spectrum then this distortion will remain in subsequent processing. Consider the spectrum presented in Figure 2.6a where peak shapes are distorted and an artificial baseline is present due to the sparse representation of the data. Performing interpolated rebinning on this data will not remove the artefacts, but instead will capture them in the output spectrum. These will then remain in subsequent process-

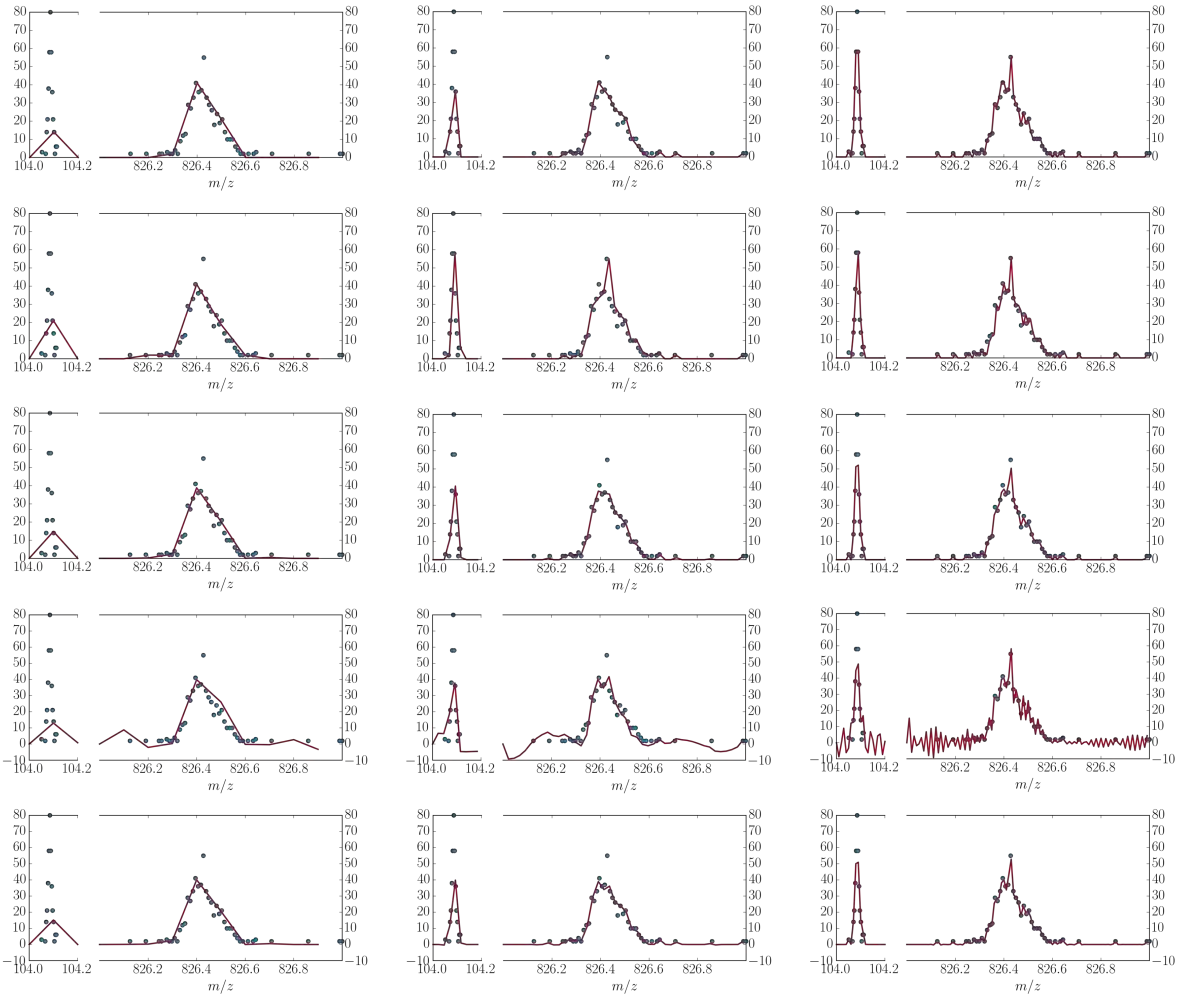


Figure 2.8: The effects of different interpolation methods (row 1 - nearest neighbour, row 2 - zero order spline, row 3 - linear, row 4 quadratic spline and row 5 cubic spline) with different bin sizes (column 1 $\Delta m = 0.1$ m/z, column 2 $\Delta m = 0.023$ m/z and column 3 $\Delta m = 0.01$ m/z).

ing causing erroneous results. Some data are stored as 'semi-sparse' where 0 values are placed either side of a series of non-zero values to correct for the visualisation artefacts presented in Figure 2.6. Interpolated rebinning will perform as desired in such situations.

Rebin in the detector domain

The drawback of both of methods described above is that this linear rebinning in *m/z* is not representative of how the data are acquired, and so a bin size that may be appropriate in the high *m/z* region may be too large in the low *m/z* region. This can be seen when comparing the peak at *m/z* 104 to the peak at *m/z* 826 in Figure 2.8. Instead of equally spaced bins in *m/z*, an axis can be generated with equally spaced bins in the domain of the detector (time in the case of TOF instruments) and then be transformed into the *m/z* domain (by a quadratic relationship for converting time to *m/z*).

To reconstruct the spectrum exactly as it was acquired, the detector sampling intervals must be know or be calculated from the data as shown in Algorithm 2.3 for TOF data. This is achieved by first converting the *m/z* axis into time using the relationship given in Equation 1.3. Then the difference between the time points is calculated. These values should all be multiples of the detector sampling interval, as shown in Figure 2.9b. Due to slight rounding errors, the selection of the minimum time interval is not sufficient, so the modal interval is taken as the sampling interval. The calculated intervals can then be used to create a global axis for a given imaging experiment using Algorithm 2.4.

Algorithm 2.3. Calculate detector sampling interval for QSTAR from spectrum

Require: m/z axis M
1: $T \leftarrow \sqrt{M}$
2: $\Delta T \leftarrow T(1, ..., |T|) - T(0, ..., |T| - 1)$
3: $\Delta T' \leftarrow \{\}$
4: **for each** Δt in ΔT **do**
5: **if** $\Delta t < \frac{3}{2} \min \Delta T$ **then**
6: Insert Δt into $\Delta T'$
7: **end if**
8: **end for**
9: $\delta \leftarrow \text{mode}\{\Delta T'\}$

Algorithm 2.4. Generate new axis for QSTAR

Require: Minimum, $m_{\min}$, and maximum, $m_{\max}$, m/z values for new m/z axis
Require: Detector sampling interval δ
1: $t_{min} \leftarrow \sqrt{m_{\min}}$
2: $t_{max} \leftarrow \sqrt{m_{\max}}$
3: Generate time axis $T \leftarrow \{t_{\min}, t_{\min} + \delta, m_{\min} + 2\delta, ..., t_{\max}\}$
4: $M_{\text{qstar}} \leftarrow T^2$

Set union of all *m/z* bins

In some cases the detector sampling interval is not constant, shown in Figure 2.9f, and so the above method is not appropriate. A universally applicable method for determining an m/z applicable for every spectrum being considered is to perform the set union on each m/z axis in the dataset, as shown in Algorithm 2.5. To generate a complete m/z axis using this method something, be it actual signal or noise, must be detected at every possible m/z bin in at least one spectrum in the image, which becomes increasingly likely as the image size increases. This method does not rely on interpolation and so ensures that the data is exactly as it was when it was recorded. There are two potential drawbacks to this method, it requires processing every spectrum within the entire dataset, which can be slow for large images, and does not ensure any consistency in the spacing of the m/z bins which has implications on subsequent window based preprocessing methods. The later drawback is exacerbated if the assumption that something is detected in every possible m/z bin is not met.

Algorithm 2.5. Set union of all m/z bins

1: $M_g \leftarrow \{\}$
2: **for each** spectrum S in dataset **do**
3: Get m/z axis M for current spectrum S
4: $M_g \leftarrow M_g \cup M$
5: **end for**

Ensure constant number of bins per peak

It is a well observed phenomenon that peaks broaden as a function of increasing m/z. In TOF instruments this is partially related to the quadratic relationship between time and m/z, however even when rebinning is performed in the detector domain, the number of bins that a peak spans increases with m/z. For example, consider the data presented in

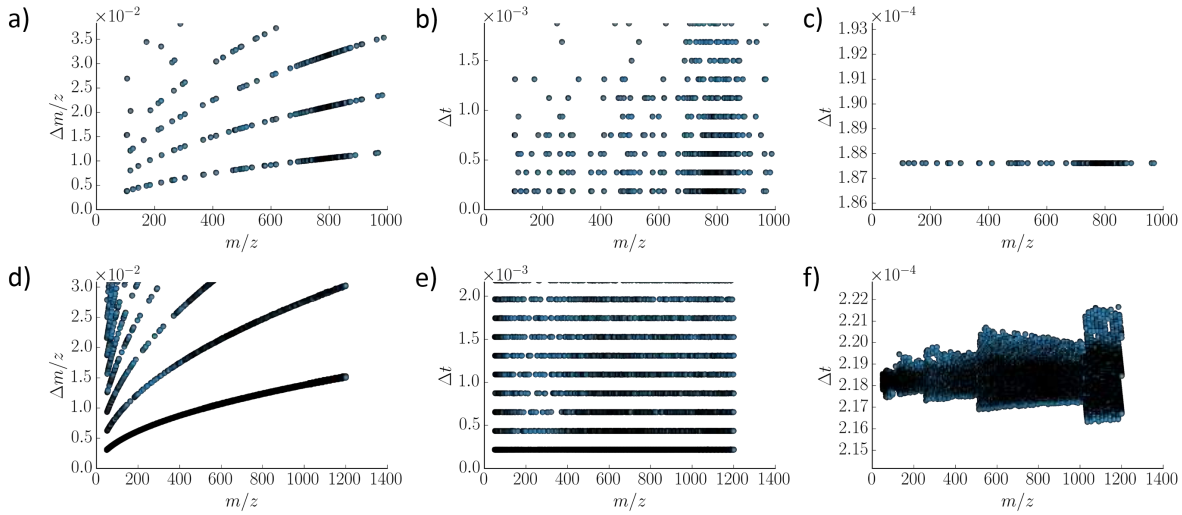


Figure 2.9: Comparison of the detector sampling in QSTAR XL (top row) and Synapt (bottom row) where a) and d) show difference in m/z between adjacent data points recorded in a single spectrum, b) and e) show the differences converted to the time domain and c) and f) show a zoomed in view of the smallest series of time differences.

Figure 2.8, the peak at 104 spans 14 recorded m/z bins whereas the peak at 826 spans 31 m/z bins. This can have a dramatic effect on window based algorithms (discussed in more detail below), so to prevent this another option is to generate an m/z axis that ensures the same number of bins span a peak at any given m/z. Interpolation, using the new axis, is then performed on each spectrum, with the same caveat described in the 'Interpolated Rebinning' section.

The ideal approach for ensuring a consistent m/z axis when additional window based preprocessing is required would first to be to rebin in the detector domain, to ensure that all zero values are replaced correctly and no artefacts will propagate, and then to ensure each peak spans a constant number of bins, to ensure that subsequent preprocessing is applicable across the whole mass range. In practice however, this can be costly in terms of both computational time and memory and so may not be applicable in certain workflows.

2.5.2 Smoothing

Smoothing aims to remove small, local, fluctuations in intensity, often caused by noise, that prevent peak detection algorithms from functioning optimally. The *de facto* standard

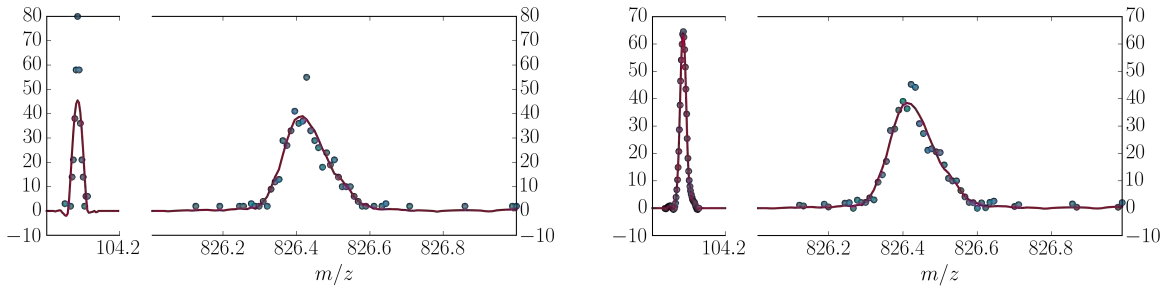


Figure 2.10: Demonstration of the effects of Savitzky-Golay smoothing (window size 15) applied to data on data rebinning in the detector domain (left) and data interpolated to ensure a peak spans approximately 30 mass bins across the whole mass range (right).

smoothing method used is Savitzky-Golay due to its intensity preserving properties [78]. This, along with other commonly used methods such as moving average and Gaussian, are window based techniques, meaning that they consider a set number of data points at once, the 'window', to generate a single data point in the resulting data. The window is then 'slid' along the data to the next point where a new window is considered.

An appropriate window size for window based smoothing methods should ideally be proportional to the number of data points across a peak. As discussed above, the number of points that a peak spans increases with m/z. For example when linearly interpolating with m/z bin size of 0.023, there are 4 points for the peak at m/z 104, but 16 for the peak at m/z 826. When rebinning based on the detector domain, the peaks at 104 and 826 span 14 and 31 m/z bins respectively. Proportionally, the peaks are more similar when the detector based m/z axis is used as opposed to the linear in m/z binning, but there is still a discrepancy in the number of data points per peak. This means that an optimal window size for smoothing at one m/z will cause peak broadening (potentially causing removal or merging of shouldered peaks) at a lower m/z and will have reduced efficacy at a higher m/z, as shown in Figure 2.10a.

This can be remedied in two ways. Firstly, the window based methods could be applied with different window sizes, proportional to the number of data points per peak, across the entire mass range. However, this requires adaptation of all window based preprocessing methods to include this functionality.

A more efficient option is to instead generate an m/z axis consisting of m/z bins that reflect the changing peak width such that when a spectrum is projected onto the new axis all peaks are represented by a constant number of data points. The results of applying Savitzky-Golay smoothing to a detector based m/z axis and a constant number of data points per peak m/z axis are shown in Figure 2.10. When the detector based m/z axis is used, the chosen window size for the smoothing function is appropriate for the peak at m/z 826, but causes peak broadening and a reduction in the height of the peak at 104, whereas when the same number of data points span both peaks, neither peak is broadened and the heights are retained, minus noise. The FWHM of the peak at m/z 104 in Figure 2.10a is 0.03, compared to 0.02 in Figure 2.10b which is equivalent to the mass resolving power being reduced to 3500 from 5200 (calculated at m/z 104.08537).

2.5.3 Baseline Correction

Baseline correction aims to remove an experimental artefact often attributed to chemical noise to aide peak detection and increase comparability between spectra. The effect is more pronounced when acquiring data over a large mass range and is a common feature in protein imaging. The type, or lack thereof, of baseline present in the data is dependent on the both the instrument and experimental parameters, such as the mass resolving power, mass range, the laser power (inducing and subsequently increasing fragmentation), the analyte and the matrix used.

The fact that different styles of baseline can occur is largely omitted from the literature, with methods often being presented as suitable for all spectra acquired using a given ion source. Any direct comparison of baseline correction methods without taking into account the style of baseline it was developed for is therefore unfair. Probably the most common style of baseline in the proteomics community is a monotonically decreasing baseline, caused by the combination of low mass resolving power and the increased number of ions detected at the low mass range (commonly referred to as chemical noise, encompassing the matrix related ions, less desirable biological molecules and associated fragmentation).

Multiple methods for correcting for this have been proposed, most notably the convex hull method, which is featured within flexAnalysis (Bruker Daltonics). This method works well for the monotonically decreasing (or increasing, although this is not a style of baseline observed in MS) case, however if the sign of the gradient of the baseline alters more than once then sections of the baseline will not be removed. To get around this issue, Coombes *et al.* [124] proposed a monotone local minimum, whereby the minimum intensity within a user defined window is taken as the baseline at that point in the spectrum.

Extending upon this idea, the TopHat baseline correction removes the assumption that the baseline correction is monotonically decreasing and instead performs the mathematical morphological opening operation on the spectrum (with a user defined structuring element width, ideally the width of the broadest peak to avoid removing signal). This method will essentially set all regions of the spectra that are larger than the structuring element width to zero. This can potentially result in amplifying noise in comparison to the reduced signal, thus reducing the apparent SNR of certain peaks.

A final option included within SpectralAnalysis, that is not necessarily a baseline correction method in the sense that it removes chemical noise, is the removal of negative values is included as an optional compensation for artefacts introduced in other preprocessing methods such as cubic interpolated rebinning and Savitzky-Golay smoothing.

2.5.4 Normalisation

Normalisation is a relatively controversial topic in mass spectrometry imaging with a significant amount of debate still ongoing. A detailed review of common normalisation methods is provided by Deininger *et al.* [81]. Since then an additional method for normalisation has been proposed by Fonville *et al.* [104]. A visual comparison of these normalisation methods applied to a sagittal section of rodent brain is shown in Figure 2.11.

In the raw image there are quite apparent experimental artefacts in the form of criss-cross patterns. These patterns are removed in all methods except median normalisation,

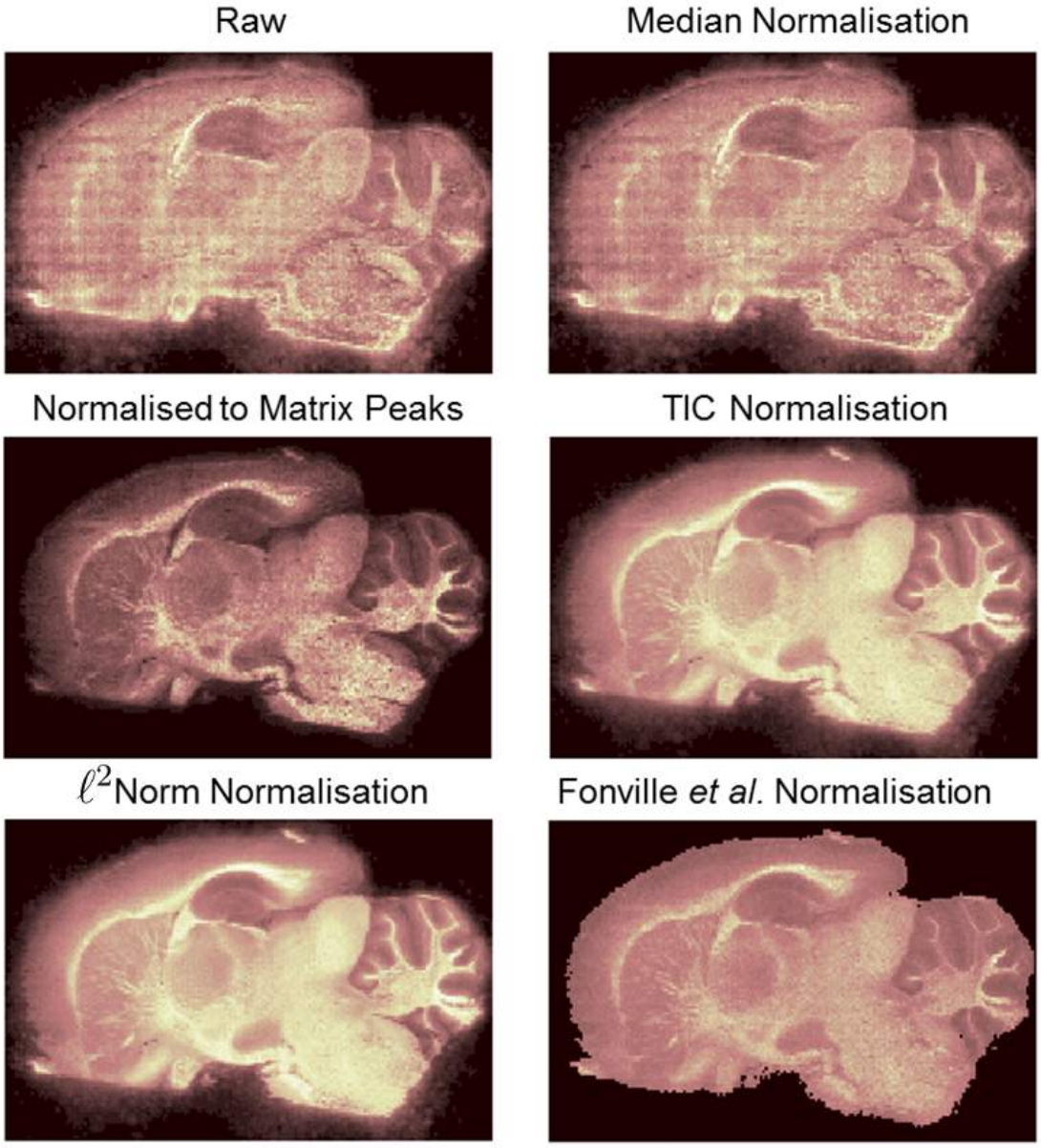


Figure 2.11: Comparison of normalisation techniques applied to the same ion image (m/z 810 in a sagittal section of formalin fixed rat brain [4]).

which normalises to an approximate measure of the intensity of the baseline. In this dataset there is no baseline, resulting in a similar median value for every pixel (which in this case becomes a measure of noise rather than the baseline, and is approximately 2). In extremely sparse datasets, the median will be 0, causing a divide by 0, and as such is an inappropriate normalisation method for such data.

The other normalisation methods make assumptions about the nature of the data. In MALDI MSI, normalising to matrix peaks assumes that the matrix sound be constant across the image and so by normalising to the matrix peaks the aim is to compensate for any heterogeneity of the matrix distribution. When considering only matrix regions, the sum of all detected matrix ions (fragments, clusters and adducts) could potentially provide a good normalisation factor. However, once an analyte is incorporated suppression effects can cause ions to be detected differently, or not at all, and so this method becomes less suitable. As this relies on the matrix peaks this method is only applicable to MALDI data, however the matrix peaks could be replaced with other experimental constants prevalent in other techniques such as solvent peaks in DESI or LESA, however with similar caveats.

The TIC (and similarly the L^2) normalisation method makes the assumption that at every pixel location the same number of ions should be detected. In homogeneous single compound samples this would hold true, however any form of heterogeneity renders this assumption inappropriate. It could be argued that within a given area there is only a given amount of charge present required for the formation of ions and so despite the heterogeneity this method is applicable. Due to varying proton affinities of molecules present, suppression effects and reactions that may occur within the plume (for example charge transfer or metastable fragmentation) this assumption is unlikely to hold true. Fonville *et al.* [104] attempts to provide a more robust method of normalisation that does not suffer from the issues listed above, by only considering signal from the analyte when constructing the scaling factor for each pixel. However, considering the heterogeneity within the analyte this is still not an ideal solution.

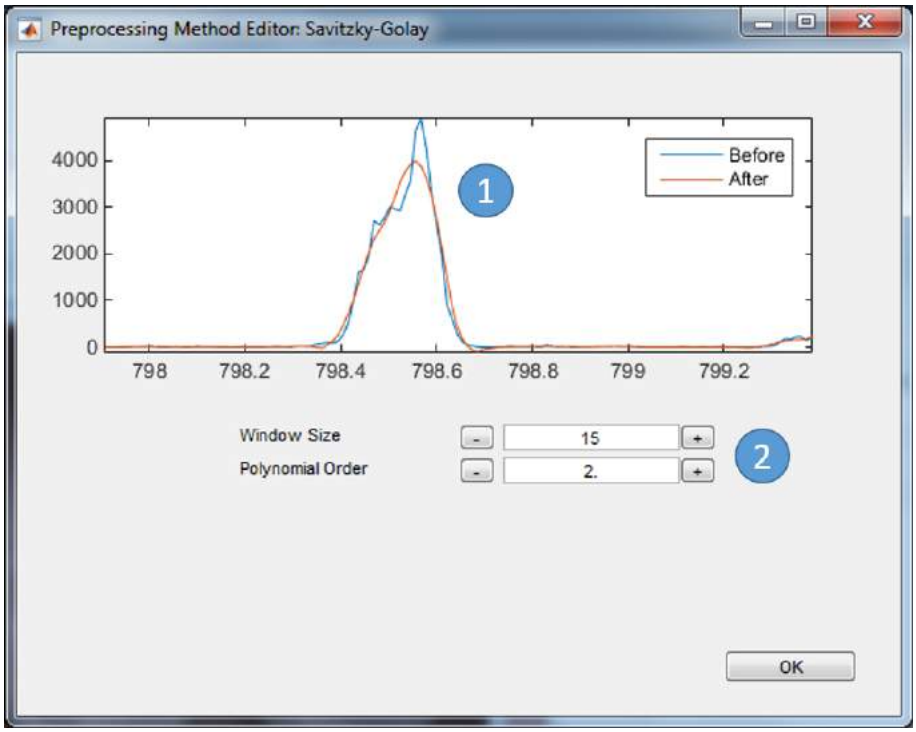


Figure 2.12: Screenshot of the real time update of preprocessing effects. 1) Raw spectrum ('Before') overlayed with the preprocessed spectrum ('After'). 2) Preprocessing method's parameters. Changing these values updates the 'After' spectrum so their effect is visualised in real time.

2.5.5 Tailoring preprocessing to data

As has been discussed and demonstrated previously, the suitability of certain preprocessing methods largely depends on the nature of the data to be analysed. To increase the compatibility with a larger number of instruments, the preprocessing methods included in each vendor's software (and discussed above) are included in SpectralAnalysis. Taking this a step further and allowing the effects of the preprocessing methods to be visualised in real time, enables the user to select appropriate methods and associated parameters for removing experimental artefacts and noise, removing the 'black box' nature of many software packages. The interface for performing this is shown in Figure 2.13.

Furthermore, it is possible to create a custom 'preprocessing workflow' that includes one or more preprocessing methods to be applied in a user specified order. This not only gives the user great flexibility over the transformations applied to each spectrum, but also enables the recreation of previously published routines such as LIMPIC [69] as well as

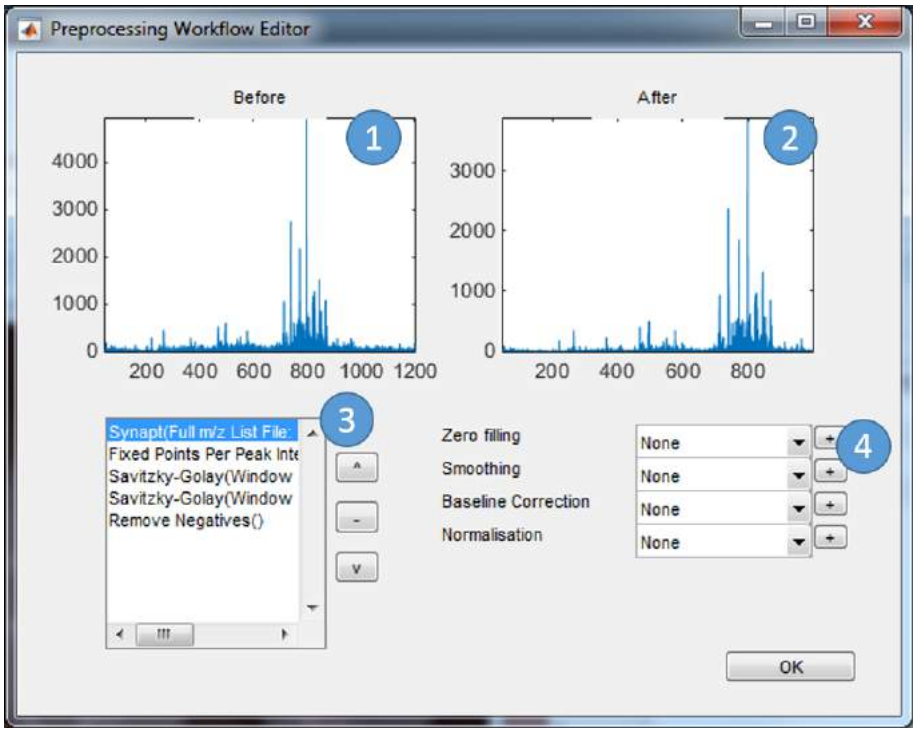


Figure 2.13: Screenshot of the workflow generation. 1) Spectrum prior to any preprocessing applied. 2) Spectrum after full preprocessing workflow applied. 3) Preprocessing workflow, a list of methods and parameters in a set order. 4)

in-house workflows. This allows rapid evaluation and incorporation of newly developed preprocessing workflows without the need for additional software and could provide a route for methods to be published alongside articles or submitted as part of the review process.

2.5.6 Image Generation and the Effect of Preprocessing on Image Quality

Multiple methods for generating ion images are implemented within mass spectrometry imaging software. The results of applying these different methods are shown in Figure 2.14. When ion images are generated without the application of any preprocessing, the resulting images are more noisy and the different methods have greater variation. Through applying preprocessing to each of the spectra, the apparent differences are lessened and only relative intensity differences remain. A difference between methods that extract a single value compared to those that sum over a mass range is still evident. This is likely

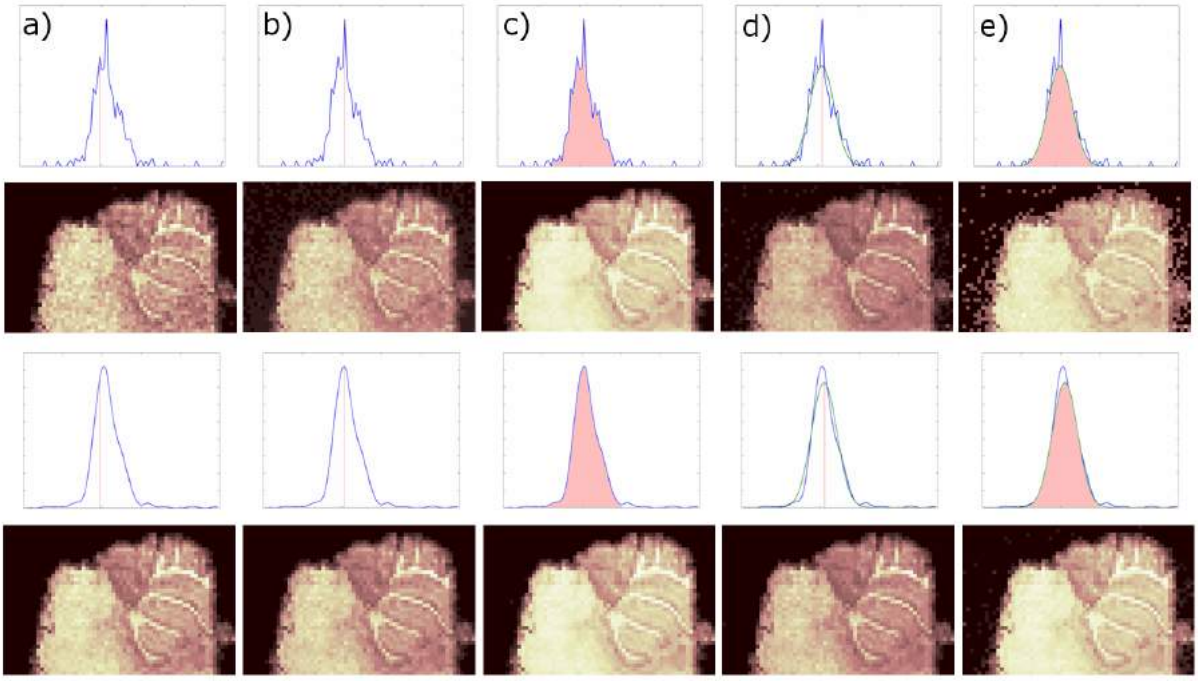


Figure 2.14: Comparison of ion images generated with various methods with and without preprocessing applied. a) Extract values at specific m/z location determined from total spectrum. b) Maximum in m/z range. c) Integrate over m/z range. d) Maximum in fitted function (Gaussian function used). e) Integrate over fitted function (Gaussian used). Row 1) Raw data shown in blue trace. Extracted value(s) shown in pink. Fitted function shown in green trace. Row 2) Generated ion images from raw data in Row 1). Row 3) Preprocessed data shown in blue trace. Extracted value(s) shown in pink. Fitted function shown in green trace. Row 4) Generated ion images from raw data in Row 3).

due to the effect illustrated in Figure 2.15 where the distribution of intensities on the left side of the peak is different to that of the right side, potentially due to unresolved ions having different spatial distributions.

2.6 Speeding up Preprocessing through the use of GPGPU

General purpose computation on graphical processing units (GPGPU) can provide significant speed improvements for highly parallelisable problems. Processing of mass spectrometry data falls in this category due to the requirement of performing exactly the same baseline correction, smoothing and peak detection on each spectrum in the dataset.

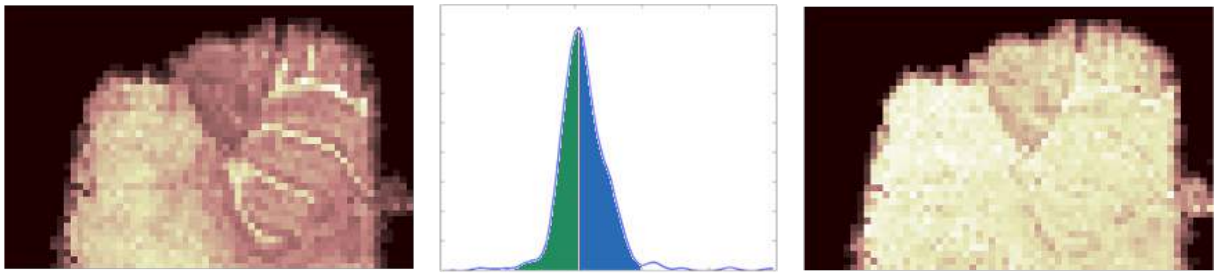


Figure 2.15: Comparison of ion images generated by summing values a) highlighted in green and c) highlighted in blue.

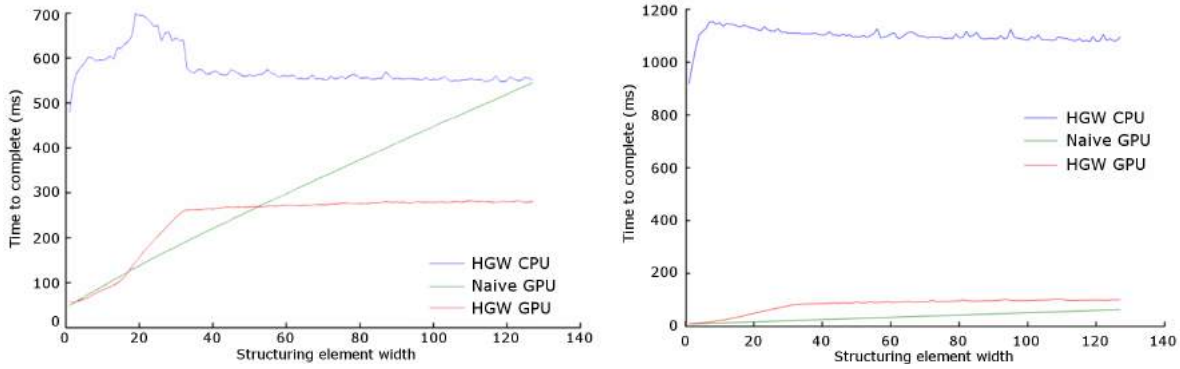


Figure 2.16: a) 3.40 GHz Intel i7-2600 running Windows 7 with an NVIDIA Quadro 2000. Transfer adds 20-30ms. b) 2.40 GHz Intel Xeon E5645 running Linux with an NVIDIA Tesla C2070. Transfer adds 20-40ms.

To demonstrate the speed improvements achievable the Top-Hat algorithm for baseline correction was implemented using the Herk-Gil-Werman (HGW) algorithm [125] for both CPU and GPU. Two iterations of Savitzky-Golay (with a fixed window size of 5) were applied to a dataset (500 pixels and 20714 m/z channels) followed by baseline correction using the Top-Hat algorithm (with varying structuring element widths) on two different hardware configurations.

As would be expected, the naïve and HGW implementations executed on a slower clock speed CPU results in worse performance, but still the HGW implementation outperforms the naïve in both cases. Again, when comparing GPU implementations, the higher performance GPU (double precision floating point performance, Tesla C2070 515 GFlops vs Quadro 2000 240 GFlops; cores, Tesla C2070 448 vs Quadro 2000 192) outperforms the lower performance GPU. However, interestingly, the relationship between the naïve and the HGW implementations differ between the two GPUs. When using the Quadro 2000

and the structuring element is less than 50, the naïve implementation performs as well as or outperforms the HGW implementation, after which point the HGW implementation remains constant time and successively outperforms the naïve implementation. This switch over point takes longer to reach when using the Tesla C2070 (structuring element size of approximately 150) due to the higher number of cores available. As the optimal structuring element size is related to the width of the largest peak within the data, values of above 150 are unlikely to be required whereas it is possible that values of greater than 50 will be needed. Thus, depending on the hardware available, and the problem at hand, the optimal implementation may differ.

The most beneficial way to reduce the processing time is to utilise different implementations based on the parameters used for each processing method and the size of the data that is to be processed. Speed improvements are also dependent on the hardware available and should be optimised and selected on a per computer basis and then applied to all subsequent processing.

2.7 Multivariate Analysis

Once peaks have been detected, the data can be reduced down to a 'datacube' [86]. As hundreds or thousands of peaks are routinely identified for a single dataset, manual investigation to determine trends within the data is laborious and impractical. Multivariate analysis techniques have been shown to be a powerful tool for aiding interpretation of these complex datasets. Despite this, very few freely available software packages include such techniques and those that do only include one or two [54, 52]. The efficacy of any single technique used in isolation has recently been brought into question [126].

To address this, SpectralAnalysis includes principal component analysis (PCA), non-negative matrix factorisation (NMF), maximum autocorrelation factor (MAF) [127] and probabilistic latent semantic analysis (PLSA) [128]. Selected factors from each of the techniques applied to a MALDI MS image of a sagittal section of rat brain [4] are shown

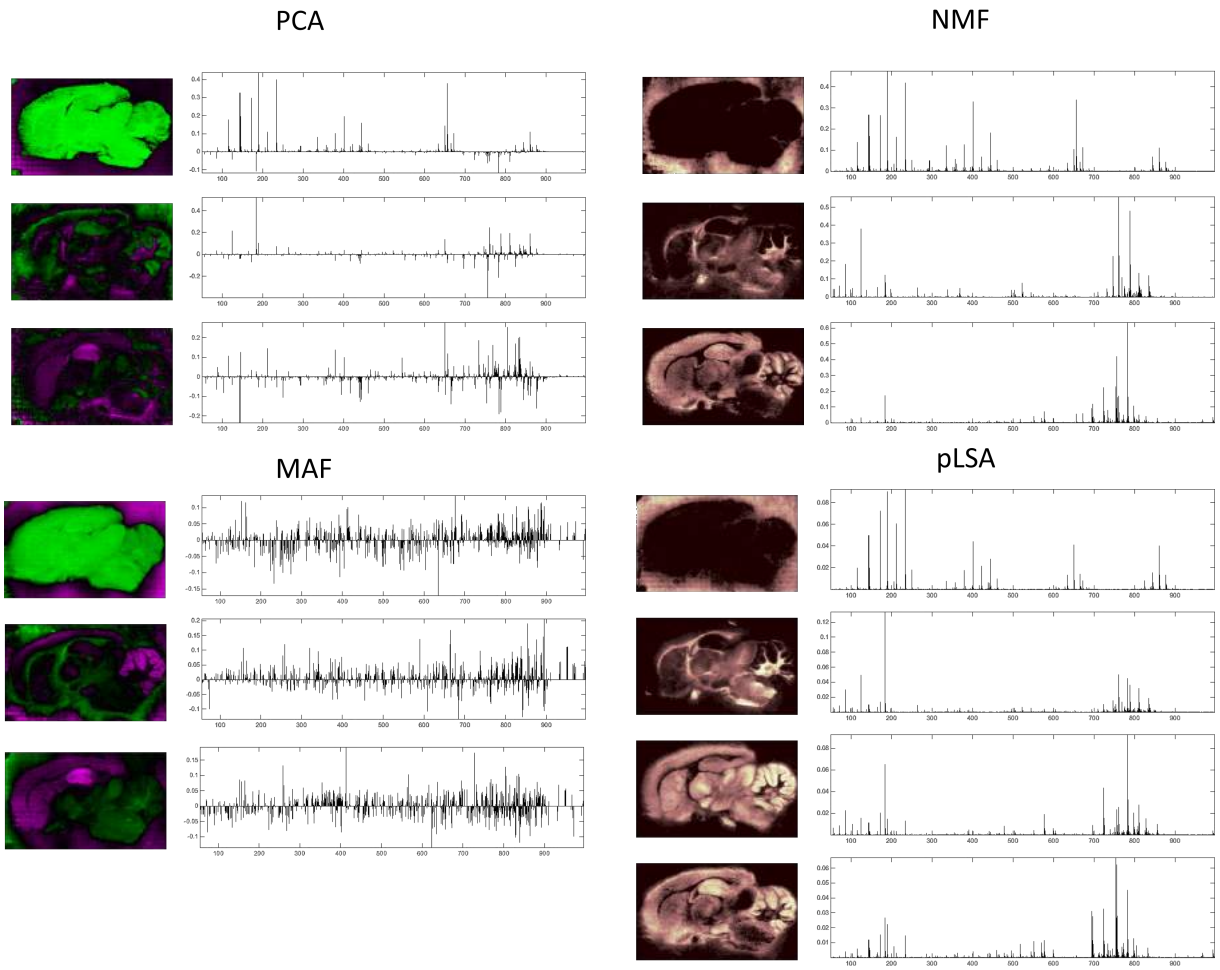


Figure 2.17: Selected factors from principal component analysis (PCA), non-negative matrix factorisation (NMF), maximum autocorrelation factor (MAF) [127] and probabilistic latent semantic analysis (PLSA) [128] applied to a MALDI MS image of a sagittal section of rat brain [4].

in Figure 2.17. The first factor from each technique was selected as the factor that showed a difference between the matrix region and the tissue region. The resulting spectral profiles in PCA, NMF and pLSA are relatively similar, showing the contribution of the matrix peaks. The second factor in PCA and MAF and the second and third factor in NMF and pLSA were selected to show differentiation between grey and white matter. The final factor in PCA, MAF and pLSA was selected to highlight the hippocampus region. The hippocampus region in the MAF factor has significant contrast, whereas it could easily be missed in the PCA and pLSA and wasn't present at all in any NMF factor. This further emphasises the power of employing multiple techniques to analyse MSI data.

2.8 Extensibility

SpectralAnalysis was developed with extensibility in mind, providing a platform for visualisation and processing that it is simple to include additional data format readers, preprocessing, multivariate analysis and clustering algorithms without the requirement of developing any user interface or data visualisation code. A block diagram visualising the core components that can be extended is shown in Figure 2.18. A 'Parser' handles the reading of a given file format, for example imzML, to get meta information such as the image dimensions (width, height, depth) and whether the data is stored in sparse format or dense (to determine the need to ensure a consistent m/z axis as discussed in Section 2.5.1) as well as to read parts of the data from disk. Extension of this allows data in different formats (such as older MSI formats like Analyze 7.5) as well as file formats associated with other imaging modalities, as discussed in Section 2.9. The 'DataRepresentation' determines how the data is to be handled, either in memory or left on disk (discussed in Chapter 3), and could be extended to include additional capabilities such as a hybrid of the two (cached data in memory, majority remain on disk). 'Preprocessing' and associated subcomponents include all of the features discussed in Section 1.6 and can be extended to include methods or algorithms that are currently omitted or to de-

velop new algorithms and make use of the real-time visualisation of its effects on spectral data. 'Postprocessing' includes all multivariate analysis techniques shown in 2.7 and can be extended to include additional algorithms. This provides a platform for rapid testing of algorithms at every stage of the analysis process on multiple modality datasets with instant visualisation of the results.

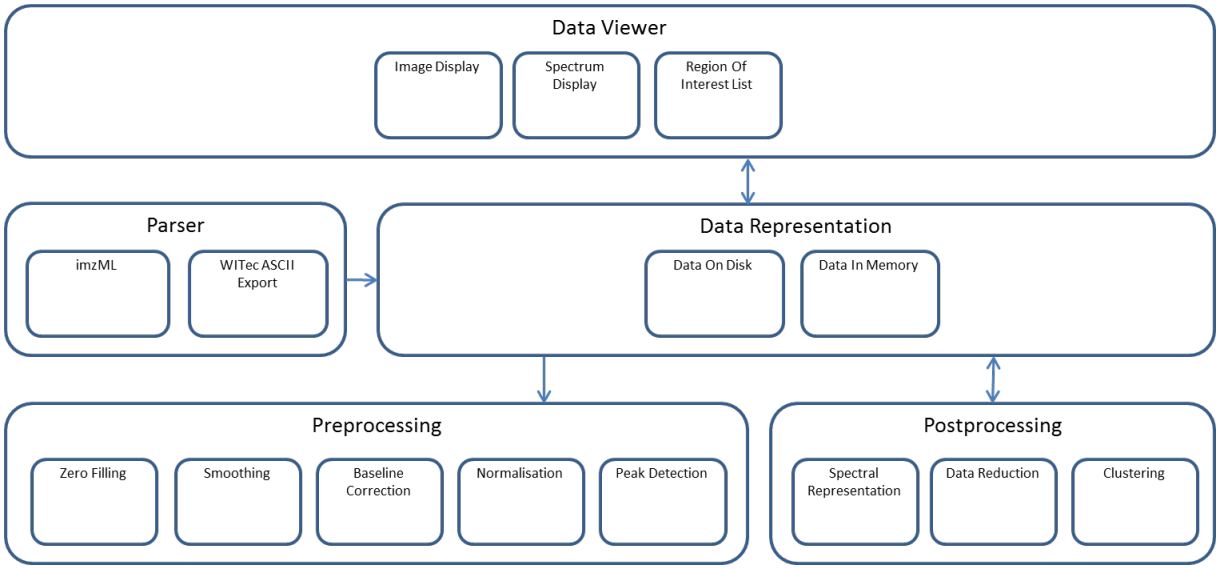


Figure 2.18: Block diagram describing the software interaction. Easily extensible sections are 'Parser', 'Preprocessing' and 'Postprocessing'.

2.9 Multimodality Data

It is becoming increasingly desirable to incorporate multiple additional techniques into the analysis of mass spectrometry imaging data. This can range from simply including histology images to determine co-localisation with anatomy, through inclusion of additional MSI data (either from the same instrument or a complementary one) [129], to the inclusion of other spectral imaging modalities such as Raman [130]. To cater for this scenario, SpectralAnalysis was written in such a way that enables any spectral data to rapidly be incorporated, allowing any of the core functionality (such as preprocessing and multivariate analysis, discussed below) to be performed without alteration. A selection of data from different modalities processed using SpectralAnalysis is given in Figure 2.19.

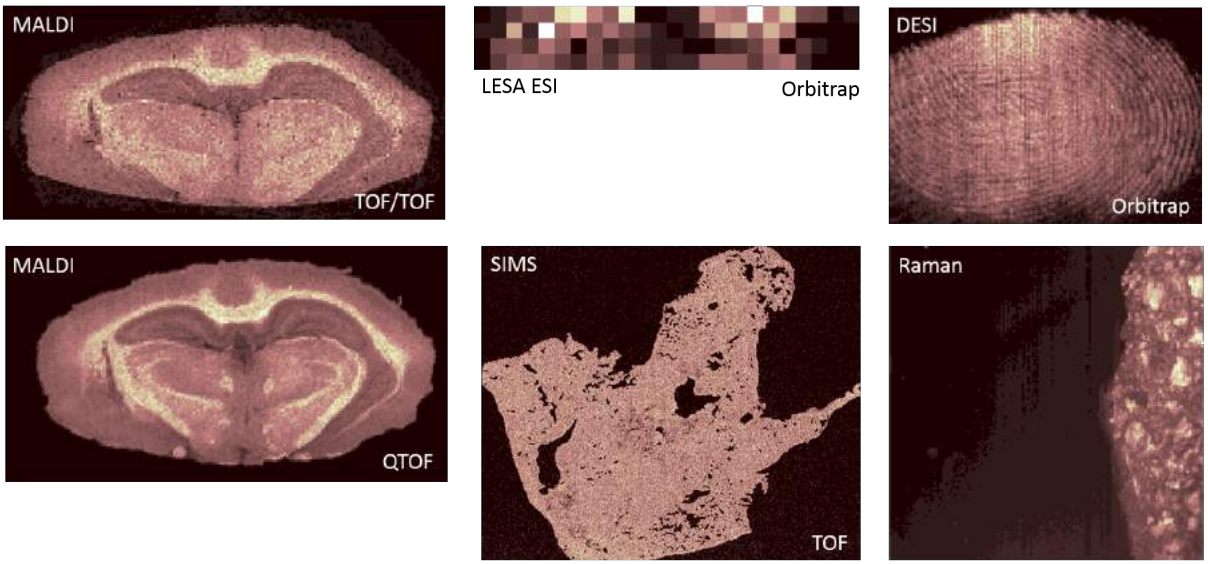


Figure 2.19: Demonstration of different imaging modality data processed using SpectralAnalysis. In the MALDI TOF/TOF and Raman data every pixel consisted of a spectrum with the same length and channel labels as every other pixel in the dataset (referred to as 'continuous' in the imzML specification). The LESA ESI Orbitrap, DESI Orbitrap, MALDI QTOF and SIMS data were stored in sparse format and so each spectrum was stored as m/z, intensity pairs resulting in different length spectra at each pixel (referred to as 'processed' in the imzML specification, despite the data being raw data).

Although some preprocessing techniques are only suitable for specific styles of data, the end goal largely remains the same and the majority of algorithms for smoothing, baseline correction and peak detection included are technique independent providing a powerful platform for multimodality processing, investigation and visualisation.

2.10 Analysing Subsets of Large Datasets

As instruments develop and increase both the mass resolution and the lateral resolution, the data size is rapidly increasing, with raw data easily capable of exceeding 10s to 100s of GB for a single MSI dataset. Furthermore, the move towards clinical applications with increasing cohort and sample numbers, with replicates, is significantly increasing this challenge. The vast majority of MSI software loads the data to be processed into RAM before any visualisation or analysis can be performed. This then introduces a restriction on the size of the data that can be processed based on the hardware of the

computer being used, and as the data size is rapidly outpacing the hardware specifications this is becoming an increasing problem and may render some software/hardware/data combinations unusable.

This problem has been addressed previously by enabling the ability to load and analyse a subsection of the dataset, limiting the number of pixels, the mass range, or both [47, 55]. SpectralAnalysis also includes this option and expands upon it by allowing the user to select an arbitrarily shaped region of interest as well as an optional mass range limit to be loaded into memory, the interface for this is shown in Figure 2.20. In order to do this the axis must be consistent and so any of the techniques discussed in Section 2.5.1 can be employed to ensure this, where in some cases the parameters can be specified to also contribute to the reduction of data (such as rebinning) at the cost of potentially discarding information.

2.11 Conclusions

SpectralAnalysis provides a unique, and currently the most exhaustive, collection of algorithms for preprocessing and subsequent multivariate analysis of spectral imaging data. This, combined with the flexibility of the extensibility to include additional algorithms, results in a platform suitable for comparisons of preprocessing methods on MSI data acquired on any instrument.

Due to the capability of handling multiple spectral imaging modalities, each of which capture different information about the sample, SpectralAnalysis would be an excellent platform on which to develop and integrate image fusion techniques, such as those used in satellite imagery. Image fusion aims to combine data from multiple sources to gain information that was not present in each source in isolation. For example, the combination of a high spatial resolution, panchromatic image with a multispectral, but low spatial resolution image resulting in a high spatial multispectral image. This could be developed and applied to MSI to combine, for example, the high spatial resolution of SIMS data

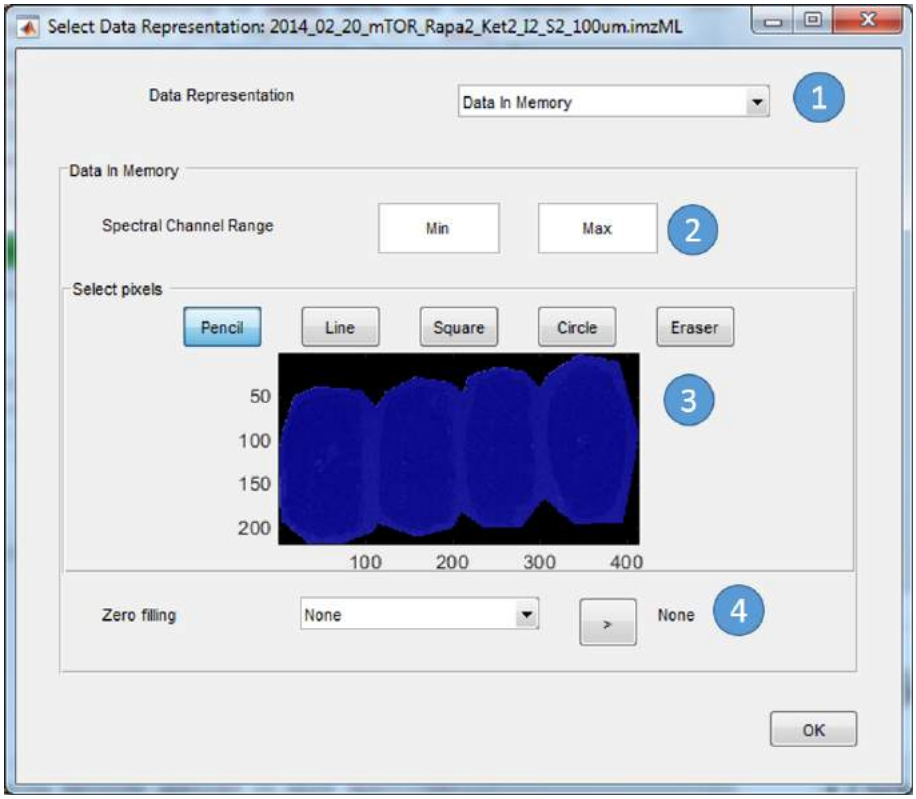


Figure 2.20: Interface presented when opening a dataset in SpectralAnalysis providing the user with options for how to handle large data. 1) Option between loading the data into memory or leaving the data on disk and using memory efficient methods (discussed in Chapter 3). 2) Limit the spectral domain. 3) Select a region of interest to load data within only that area. 4) Ensure consistent m/z axis to enable loading the data into memory and for optional data size reduction as discussed in Section 2.5.1.

with the high mass range of MALDI data.

CHAPTER 3

MEMORY EFFICIENT DATA HANDLING AND ANALYSIS

3.1 Introduction

The size of a single mass spectrometry image is rapidly increasing. Datasets can easily reach 100s GB, which far exceeds the available RAM in a typical computer. This presents the analyst with only one option if they wish to analyse their data, they must reduce the data such that it fits within memory while also ensuring sufficient free memory remaining to perform any subsequent analysis. This can be achieved through one or more of the methods discussed in Section 2.10, however these either limit the analyst's view of the dataset as a whole or discard data (and potentially analytically useful information) which can be detrimental to the analysis. This issue of datasize is compounded when multiple MS images are combined as in Section 2.3, requiring the analyst to compromise even further to be able to visualise the data. This chapter describes memory efficient methods that enable datasets vastly exceeding the size of the available RAM to be visualised, preprocessed and analysed using multivariate analysis.

3.2 Memory Efficient Ion Image Generation

Generation of individual ion images does not require the whole dataset to be loaded into memory. Instead only the data points that fall within the peak boundaries (m_{min} and m_{max}) are required so that one of the methods described in Section 2.5.6 can be applied. This can also be taken a step further, and since the calculation of an intensity at a given pixel is completely independent of all other pixels, only one spectrum is required in memory at a given point in time. This significantly reduces the amount of memory required, as a single spectrum ranges from 100s kB to 1-2 MB, compared to the 10s of GBs for the whole dataset. In many cases it is desirable to preprocess the data prior to ion image generation. The algorithm presented in Algorithm 3.1 presents a memory efficient method of generating ion images from preprocessed data. Each spectrum is loaded in sequentially, preprocessed and then the data points within peak limits are extracted and an intensity is generated based on the image generation method of choice. The spectrum can then be removed from memory before the next is loaded in. This reduces the amount of memory required to the size of a single spectrum, plus the size of the ion image(s) to be generated, which is orders of magnitude smaller than the whole data, allowing TBs of data to be visualised on even the most memory constrained systems.

Algorithm 3.1. Memory Efficient Ion Image Generation

Require: Peak limits m_{min} and m_{max}
Require: Preprocessing workflow $W \leftarrow \{w_1(S), ..., w_n(S)\}$ containing n preprocessing methods from Section 1.6
Require: Image generation method $G(S, m_{\text{min}}, m_{\text{max}})$ from Section 2.5.6
1: $I \leftarrow \{\}$
2: **for each** spectrum index i in dataset **do**
3: Read in spectrum S_i from disk
4: **for each** preprocessing method w in W **do**
5: $S_i \leftarrow w(S_i)$
6: **end for**
7: Get spatial location of spectrum from header information (x, y, z)
8: $I(x, y, z) \leftarrow G(S_i, m_{\text{min}}, m_{\text{max}})$
9: **end for**

3.3 Memory Efficient Spectral Representation Generation

Peak detection is often performed on spectral representations of the data [86]. As above, these only require a single spectrum to be loaded into memory at once and can be generated in a memory efficient manner using Algorithm 3.4. It is possible to generate multiple representations at once, requiring only a single pass through of the data, by including additional update methods after line 8 of Algorithm 3.4. This provides a memory efficient method of generating all spectral representation proposed by McDonnell *et al.* [86] for optimal peak detection in a given dataset.

Algorithm 3.2. $U_t(S_R, S)$ Update method for total/mean spectrum generation

Require: $|S_R| = |S|$
1: **for each** index i in S_R **do**
2: $S_R(i) \leftarrow S_R(i) + S(i)$
3: **end for**

Algorithm 3.3. $U_b(S_R, S)$ Update method for basepeak spectrum generation

Require: $|S_R| = |S|$
1: **for each** index i in S_R **do**
2: $S_R(i) \leftarrow \max(S_R(i), S(i))$
3: **end for**

Algorithm 3.4. Memory Efficient Spectral Representation Generation

Require: Preprocessing workflow $W \leftarrow \{w_1(S), ..., w_n(S)\}$ containing n preprocessing methods from Section 1.6
Require: Spectral representation update method $U(S_R, S)$ from Algorithm 3.2 or 3.3
1: $S_R \leftarrow \{\}$
2: $n \leftarrow 0$
3: **for each** spectrum index i in dataset **do**
4: Read in spectrum S_i from disk
5: **for each** preprocessing method w in W **do**
6: $S_i \leftarrow w(S_i)$
7: **end for**
8: $S_R \leftarrow U(S_R, S_i)$
9: $n \leftarrow n + 1$
10: **end for**
11: If mean spectrum required, $S_R \leftarrow S_R/n$

3.4 Memory Efficient Datacube Generation

By combining the above methods it is possible to reduce the MS image to a 'datacube' in a memory efficient manner, using Algorithm 3.5. In this case, only a single spectrum and

the datacube is required to be in memory at any one point in time. This allows reduction of data to peak lists without a limitation applied to the number of peaks retained.

Algorithm 3.5. Memory Efficient Datacube Generation

Require: Preprocessing workflow $W \leftarrow \{w_1(S), ..., w_n(S)\}$ containing n preprocessing methods from Section 1.6
Require: Image generation method $G(S, m_{\min}, m_{\max})$ from Section 2.5.6
1: Calculate spectral representation S_R using Algorithm 3.4
2: Peak pick on S_R using chosen method from Section 1.6.5 to get m/z limits $M_{\min}$ and $M_{\max}$
3: $D \leftarrow \{\}$
4: **for each** spectrum index i in dataset **do**
5: Read in spectrum S_i from disk
6: **for each** preprocessing method w in W **do**
7: $S_i \leftarrow w(S_i)$
8: **end for**
9: **for each** peak index j in $M_{\min}$ **do**
10: $D(i, j) \leftarrow G(S_i, M_{\min}(j), M_{\max}(j))$
11: **end for**
12: **end for**

The algorithm as it is presented reduces and loads the data into memory, however this can also be used to write the reduced data to disk by altering line 10 in Algorithm 3.5 to be a disk write instead of a matrix update. In this case only a single spectrum is required to be in memory, making this process feasible on memory constrained systems where the datacube is larger than that of the RAM. The methods for handling large datasets described in Section 2.10 can then be employed to visualise portions of the data.

3.5 Memory Efficient Principal Component Analysis

It is not always the case that data reduced to a datacube is small enough to fit into RAM while also leaving enough memory to perform multivariate analysis, using a technique such a principal component analysis (PCA). One option, described in Section 2.10, is to load only part of the dataset and perform PCA, however this relies on *a priori* knowledge that may not exist and limits the analyst's understanding of the data to the portion loaded.

A method for performing PCA on large databases, where it is frequently impossible to load an entire dataset due to memory limitations, has been previously been reported [131]. The method is based on the formation of two "summarisation" matrices

$$\mathbf{L} = \sum_{i=1}^{N} \mathbf{x}^{(i)} \tag{3.1}$$

$$\mathbf{Q} = \sum_{i=1}^{N} \mathbf{x}^{(i)} (\mathbf{x}^{(i)})^T \tag{3.2}$$

where the column vector $\mathbf{x}^{(i)}$ is the i^{th} data entry and the sums are over the N entries in the dataset. From the summarisation matrices, the covariance matrix can be computed as

$$\Sigma = \frac{1}{N}\mathbf{Q} - \frac{1}{N^2}\mathbf{L}\mathbf{L}^T. \tag{3.3}$$

The important feature of this formulation is that the summarisation matrices can be formed incrementally and require only one data point to be in memory. This can readily be adapted to MSI data, where the summarisation matrices can be formed incrementally from each spectrum, corresponding to a pixel (or voxel), shown in full detail in Algorithm 3.6 and diagrammatically in Figure 3.1. This is especially advantageous for datasets containing large numbers of pixels. Note that the full covariance matrix must be constructed and very high-dimensional datasets may still prove to be intractable. For this reason, we employ peak detection methods in order to reduce the dimensionality of the data.

3.5.1 Description of the algorithm

With reference to Figure 3.1 and Algorithm 3.6, the algorithm can be decomposed into the following steps:

Determine peak list (lines 1-2): The preprocessing workflow, W, was comprised of rebinning in the detector domain (Section 2.5.1) and Savitzky-Golay smoothing (with window size 25). Peak picking was performed using the gradient method (Section 1.6.5) on the basepeak spectrum (Algorithm 3.3).

Read in spectrum (lines 7, 24): Spectra, consisting of (m/z, intensity) pairs, were accessed sequentially from the binary portion of the imzML format and loaded into RAM.

Pre-process spectrum (lines 8-14, 25-31): After loading, spectra were preprocessed using the workflow W described above. Each peak in the spectrum was reduced to a single intensity value, using the image generation method, G, extract value at peak apex as described in Section 2.5.6.

Update summarisation matrices (lines 15-16): The summarisation matrices are updated with the current peak reduced spectrum S_p

$$L \leftarrow L + S_p \tag{3.4}$$

$$Q \leftarrow Q + S_p S_p^T \tag{3.5}$$

Calculate Σ (line 19): Once the summarisation matrices have been updated with all spectra, the covariance matrix can be computed as

$$\Sigma \leftarrow \frac{1}{N} Q - \frac{1}{N^2} L L^T \tag{3.6}$$

Eigendecomposition of Σ (lines 20-21): Eigendecomposition was performed by first reducing the covariance matrix to a tridiagonal matrix followed by QR decomposition of the tridiagonal matrix to calculate the eigenvalues and eigenvectors.

$$U \Lambda U^T = \Sigma \tag{3.7}$$

where U contains the eigenvectors (also referred to as the loadings or coefficients in PCA) and the diagonal of Σ contains the eigenvalues.

Calculate Scores (line 32): The scores of the data points against the principal components are computed point-by-point (only one spectrum is required in memory). The score of the spectrum from pixel i against principal component j is computed as

$$P(i,j) \leftarrow (S_p - \frac{L}{N})^T U_j \tag{3.8}$$

where U_j is the j^{th} column of U (the j^{th} principal component) and L/N is subtracted in order to mean-centre the data. Score images can then be generated for a principal component of choice by arranging the score values in the same two dimensional grid as the spectra were collected.

In certain cases, it may be necessary to construct the correlation matrix instead of the covariance matrix. This is necessary in cases where the variables are on different scales and so is not generally applicable for MSI data, but may be useful for liquid chromatography ion mobility spectrometry-mass spectrometry (LC-IMS-MS) data where elution time, drift time and m/z are on different axes [132]. The correlation matrix ρ can be formed as

$$\rho = \Sigma \circ \frac{1}{\sigma\sigma^T} \tag{3.9}$$

where $\circ$ denotes the element-wise (Hadamard) product of the matrices and $\sigma = \sqrt{\text{diag}\,(\Sigma)}$ is a vector containing the standard deviation of each dimension (peak). Overwriting the covariance matrix with the correlation matrix will ensure that no extra memory is required and then the subsequent steps of eigendecomposition of ρ ($U\Sigma U^T = \rho$) and scoring can be followed as described previously, with Equation 3.8 replaced by

$$P(i,j) \leftarrow \left[\left(S_p - \frac{L}{N}\right) \circ \frac{1}{\sigma}\right]^T U_j \tag{3.10}$$

3.5.2 Numerical Optimisations

Whilst this method does not require the whole dataset to be loaded simultaneously, the full covariance matrix ($M \times M$, where M is the number of peaks) must be formed. This limits the dimensionality (number of peaks) that can be processed. However, there are several numerical optimisations that can be made in order to increase this limit. In the

Algorithm 3.6. Memory Efficient PCA

Require: Preprocessing workflow $W \leftarrow \{w_1(S), ..., w_n(S)\}$ containing n preprocessing methods from Section 1.6
Require: Image generation method $G(S, m_{\min}, m_{\max})$ from Section 2.5.6
Require: Number of components to generate c
Calculate spectral representation S_R using Algorithm 3.4
Peak pick on S_R using chosen method from Section 1.6.5 to get m/z limits $M_{\min}$ and $M_{\max}$
$L \leftarrow 0$
$Q \leftarrow 0$
$N \leftarrow 0$
for each spectrum index i in dataset **do**
    Read in spectrum S_i from disk
    for each preprocessing method w in W **do**
        $S_i \leftarrow w(S_i)$
    end for
    $S_p \leftarrow \{\}$
    for each peak index j in $M_{\min}$ **do**
        $S_p(j) \leftarrow G(S_i, M_{\min}(j), M_{\max}(j))$
    end for
    Update summarisation matrix L using Equation 3.4
    Update summarisation matrix Q using Equation 3.5
    $N \leftarrow N + 1$
end for
Calculate covariance matrix Σ using Equation 3.6
Reduce Σ to tridiagonal matrix
Perform QR decomposition of tridiagonal matrix resulting in eigenvalues Λ and eigenvectors U
$P \leftarrow \{\}$
for each spectrum index i in dataset **do**
    Read in spectrum S_i from disk
    for each preprocessing method w in W **do**
        $S_i \leftarrow w(S_i)$
    end for
    $S_p \leftarrow \{\}$
    for each peak index j in $M_{\min}$ **do**
        $S_p(j) \leftarrow G(S_i, M_{\min}(j), M_{\max}(j))$
    end for
    Update row i of scores matrix P using Equation 3.8
end for

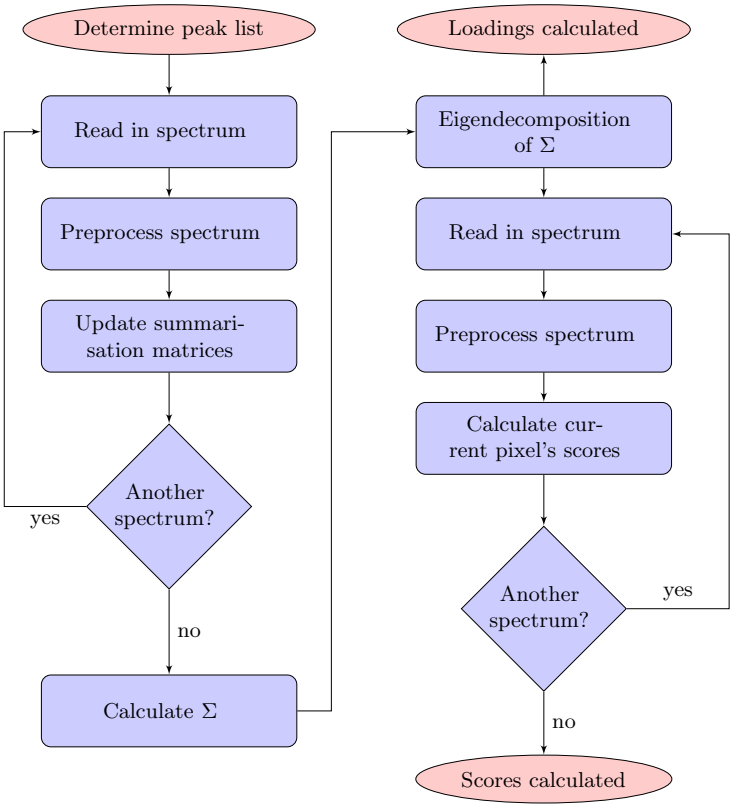


Figure 3.1: Workflow for memory efficient principal component analysis, only requiring a single spectrum plus the summarisation matrices in RAM at any single point in time.

computation of Q, the product $S_p S_p^T$ is symmetric and hence Q is also symmetric and one need only compute and store the upper triangular part of Q. The formation of both $S_p S_p^T$ and LL^T can be performed element-by-element, updating the relevant variable (Q and Σ respectively), removing the need to allocate any extra memory for storing temporary variables the same size as the covariance matrix. Furthermore, the covariance matrix can be computed "in-place" of Q to further reduce the memory requirements. With the covariance matrix stored in packed triangular form, the necessary eigendecomposition can be performed first by tridiagonalisation, and then decomposing with optimised numerical methods from LAPACK designed for this case [133]. It is also possible to discard latter columns of U once the number of principal components (P) to retain has been determined, reducing U from $M \times M$ to $M \times P$.

Figure 3.2: Data sizes that *princomp* (MATLAB Statistics Toolbox version 7.12.0.635) can process using 8 GB RAM shown as the area under the curve, demonstrating the compromise between either number of pixels (or samples) or peaks retained in the mass spectrometry data reduction step. All combinations of number of peaks and pixels shown can be processed with the new workflow.

3.5.3 Memory usage comparisons with current methods

The maximum size of data that could be processed with *princomp* was determined by selecting a value for the number of peaks, M, and then iteratively altering the number of pixels, N, (using binary search) to produce a random data matrix of size $N \times M$ and then performing *princomp* on the matrix. If this resulted in an 'Out of Memory' error then the number of pixels was reduced, otherwise the number of pixels was increased. When the maximum number of pixels was determined for a specific number of peaks, the number of peaks was altered and the process was repeated. The results of this from a computer with 8 GB RAM are presented in Figure 3.2.

Assuming a large, high resolution image (of 32M pixels [121]) is to be processed using *princomp*, only 8 peaks could be retained. With such small numbers of peaks, it is feasible to examine all ion images manually and there is no longer any requirement to reduce the dimensionality prior to further analysis. Furthermore, and more importantly, reducing a complex dataset to a small number of peaks will likely discard a significant amount of informative data. For data sizes which exceed the memory limit, the data must be reduced prior to PCA by discarding some information, introducing a compromise between the number of pixels and the number of peaks to retain, and is often remedied through the use of data reduction techniques that discard peaks, pixels or both [86, 104]. Discarding pixels generally requires some form of prior knowledge about the dataset, specifically which pixels are relevant. This has the opportunity to introduce analyst bias if the pixel selection method requires user input to aid the generation of a mask to separate matrix related pixels and sample related pixels. Discarding peaks is a much more widely accepted form of data reduction, but dependent on the method used it can introduce the

same disadvantages as discarding pixels. However, it should be noted that if uninteresting or uninformative peaks and pixels can be discarded in an unbiased and correct way then the results of PCA will provide much more insightful information on the nature of the variations in the data and potentially reveal variation that was previously masked by the uninformative differences.

An alternative method to reduce the amount of memory required by the PCA algorithm is to use the NIPALS implementation of PCA [134]. This algorithm iteratively calculates the principal components in order of largest variance explained and so only the required principal components can be calculated, reducing computation time and memory requirements. However, the requirement to store the data matrix as well as the residual matrix (which is the same size as the data matrix) in memory means that for large datasets this method will still be prohibitive and the proposed method will still out perform NIPALS in terms of memory savings.

3.5.4 Verification

To verify that the new workflow produced the same results as simply using *princomp* a dataset with a sufficiently small number of pixels and detected peaks was chosen, and the coefficients and scores produced were compared. The difference between the two resulting coefficient matrices is on the order of 10^{-6} ($10^{-4}\%$). The matrix $\mathbf{Q}$ (Equation 3.2) summarises $\mathbf{X}\mathbf{X}^T$, where $\mathbf{X}$ is the data matrix [131] and any differences between these two matrices propagate through the algorithm. Such small differences do not significantly affect any of the observed results; the correlation between any principal component's coefficients calculated by the two methods is 1, therefore they describe identical distributions. A visual comparison of score images produced using both methods is shown in Figure 3.3 with the memory requirements at each step of the process shown in Table 3.1. Although the memory sizes included in Table 3.1 focus on the most common case where the number of pixels is larger than the number of peaks, memory savings will still be achieved in the inverse case but will be less significant. For a system with 8 GB RAM, a dataset

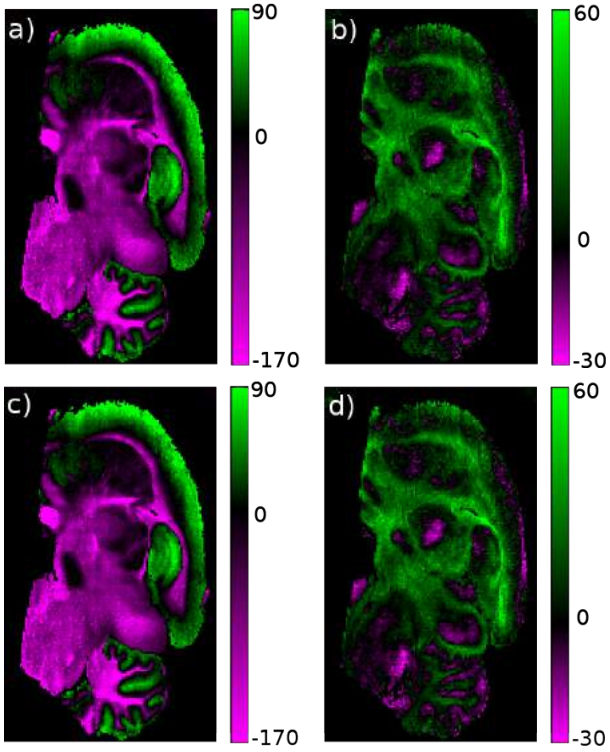


Figure 3.3: Comparison of principal component score images of a MALDI MS image of a single rat brain section using *princomp* (a, b) and the memory efficient PCA method (c, d). a) and c) show principal component 5 (demonstrating a significant amount of variance between grey and white matter regions) and b) and d) show principal component 19 (demonstrating that information is still contained in high principal components). Score images produced with either method are identical, independent of which principal component is calculated.

with 100000 pixels and assuming the first 50 principal components will be calculated, the maximum number of peaks that can be included in the proposed workflow is greater than 26000. This is in contrast to the maximum of 2666 peaks included when using *princomp* shown in Figure 3.2.

Verification on large data sets was performed by simulating a full 3D MALDI MSI experiment by replicating the single section data used previously. Binning the data at 0.2 *m/z*, the standard bin width used in BIOMAP (Novartis), resulted in 4751 bins (31 kB per spectrum) and a data size of 744 MB (for 20535 pixels). Only two full slices of this size could be retained and processed using *princomp* and 8 GB of RAM. Assuming 12 μm sagittal sections are taken of a rat brain of size 2cm x 1cm, the distance between the two full sections would be 325 μm, over which distance the internal structure can change

Step	*princomp*	Memory (MB)	Proposed Method	Memory (MB)
1)	Data matrix ($N \times M$)	2288.82	$\mathbf{x}^{(i)}$ ($M \times 1$)	0.02
			$\mathbf{L}$ ($M \times 1$)	0.02
			$\mathbf{Q}$ ($[M(M+1)/2] \times 1$)	34.34
2)	Data matrix ($N \times M$)	2288.82	$\mathbf{L}$ ($M \times 1$)	0.02
	Mean centred data ($N \times M$)	2288.82	Σ ($[M(M+1)/2] \times 1$)	34.34
3)	Data matrix ($N \times M$)	2288.82	$\mathbf{L}$ ($M \times 1$)	0.02
	Mean centred data ($N \times M$)	2288.82	Σ ($[M(M+1)/2] \times 1$)	34.34
	$\mathbf{U}$ ($N \times M$)	2288.82	$\mathbf{U}$ ($M \times M$)	68.66
	$\mathbf{S}$ ($M \times M$)	68.66	$\mathbf{\Lambda}$ ($M \times 1$)	0.02
	$\mathbf{V}$ ($M \times M$)	68.66	Working ($[4M-4] \times 1$)	0.09
4)	Data matrix ($N \times M$)	2288.82	$\mathbf{L}$ ($M \times 1$)	0.02
	Mean centred data ($N \times M$)	2288.82	Σ ($[M(M+1)/2] \times 1$)	34.34
	$\mathbf{U}$ ($N \times M$)	2288.82	$\mathbf{U}$ ($M \times M$)	68.66
	$\mathbf{S}$ ($M \times 1$)	0.02	$\mathbf{\Lambda}$ ($M \times 1$)	0.02
	$\mathbf{V}$ ($M \times M$)	68.66	Scores ($N \times P$)	38.15
	Scores ($N \times M$)	2288.82		
	Max. RAM Usage (MB)	9223.96	Max. RAM Usage (MB)	141.18

Step	*princomp*	Proposed Method
1)	Read dataset into RAM	'Read in spectrum'
	Pre-process and reduce if necessary	'Pre-process spectrum'
		'Update summarisation matrices'
2)	Mean center data matrix	'Calculate Σ'
3)	SVD of data matrix	'Eigendecomposition of $\mathbf{\Sigma}$'
4)	Data matrix projected using eigenvectors	'Calculate Scores'

Table 3.1: Memory size requirements and the corresponding intermediate variable sizes at each step of the principal component analysis algorithm using N (number of pixels) = 100000, M (number of peaks) = 3000 and assuming P (principal components to calculate) = 50 (assuming 99% variance is explained in the first 50 principal components, however commonly fewer principal components are required in practice). Steps 1-4) for *princomp* correspond to steps i-iv) as described in the introduction and are summarised in the lower table. Steps 1-4) for the proposed method refer to the named algorithm steps described Section 3.5.1.

significantly. Using the method of selecting informative peaks by Fonville *et al.* [104] to reduce the number of bins (which results in 564 informative bins at 4 kB per spectrum and 88 MB per slice), 23 sections would be able to be processed, with 30 μm between each section. Clearly acquiring data with higher lateral, axial or mass resolution would increase the data size, regardless of applying peak picking or not, and therefore decrease the number of sections possible while increasing the distance between sections. However use of the proposed method would enable the entire brain to be processed at any lateral or axial resolution, either with or without discarding peaks detected on either binned or raw data acquired using a QSTAR XL or QSTAR Elite over the mass range of m/z 50-1000, shown in Figure 3.4.

3.5.5 Application to real world data

Following verification, the new methodology was applied to a dataset containing multiple serial sections taken from a single mouse brain which was too large to be handled with *princomp* [21]. The new methodology was used to reduce the 3D dataset without the requirement of discarding peaks or pixels. Clustering was then performed, using the k-means algorithm ($k = 7$, selected as optimal [135] from $k = 2...10$ shown in Figure 3.5), on the reduced dataset, revealing 3 clusters on the tissue region and 4 clusters describing the matrix region. Prior to visualisation image registration was performed on ion images of m/z 826 using StackReg as part of the Fiji package [136] which was modified to enable exportation of the calculated affine transform. The affine transform was then imported into MATLAB and applied to all image stacks prior to visualisation. All 3D data were visualised using vol3d v2 [137]. All three clusters in the tissue region are visualised in Figure 3.6, highlighting white matter in yellow (including the corpus callosum and arbor vitae), grey matter (including the cerebral cortex) in blue and a tissue edge region in red.

Molecules which correlate with each cluster were determined by calculating the Pearson product-moment correlation coefficient (given in Equation 3.11) between every image produced at each m/z bin in the raw data and a binary image of the cluster of interest

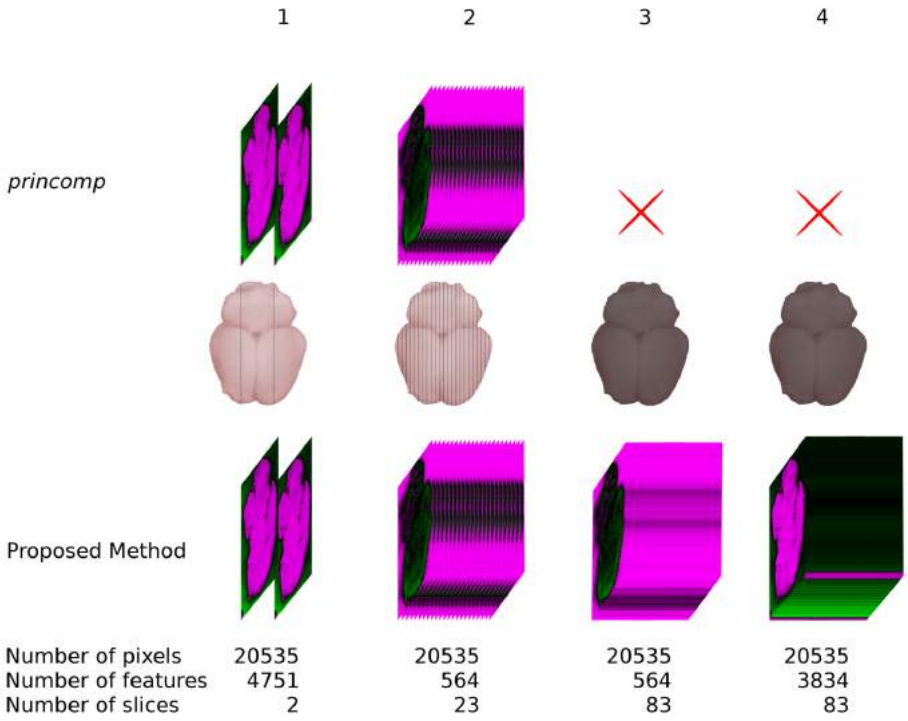


Figure 3.4: Simulated 3D MALDI MSI data of a 2 cm x 1 cm rat brain through repetition of a single 12 μm section image and the corresponding first principal component score image when using *princomp* (top) and the proposed method (bottom). Column 1 shows the maximum number of sections (2 sections) that could be retained if the data was binned at 0.2 m/z (4751 bins), the standard bin width used in BIOMAP (Novartis), and then analysed with *princomp* with 8 GB RAM. Distance between sections would be 325 μm. Column 2 shows the maximum number of sections (23 sections) that could be retained if informative peaks were extracted [104] (564 extracted m/z bins) prior to PCA using *princomp*. Distance between sections would be 30 μm. Column 3 shows the PCA results if the entire brain was sectioned and analysed (83 sections) and informative peaks were extracted [104]. The red cross indicates *princomp* failed due to memory limitations. Column 4 shows the PCA results if the entire brain was sectioned and analysed (83 sections) and all detected peaks (3834 detected peaks) were retained when determined from the entire data set.

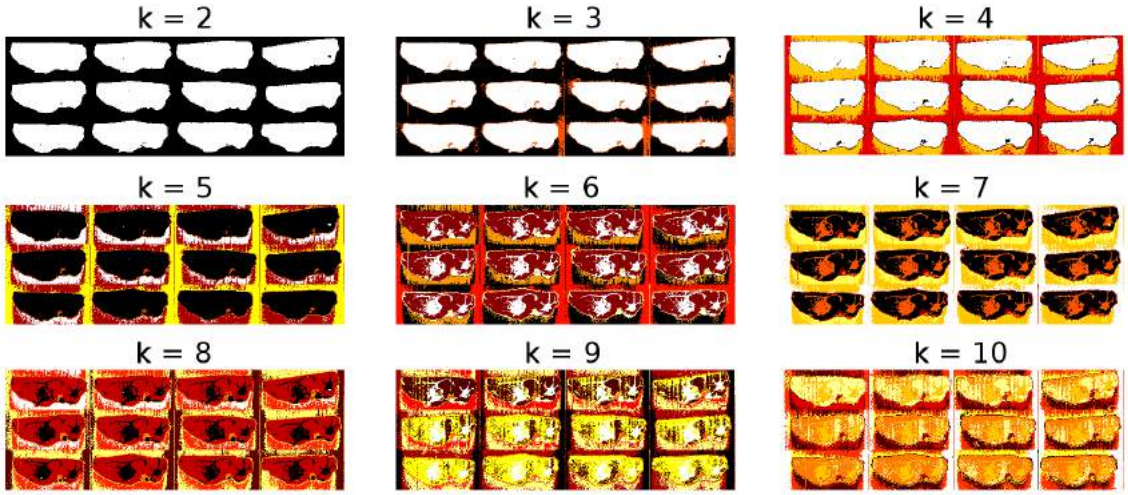


Figure 3.5: k-means applied to PC 1-40 scores of serial mouse brain sections with varying values for $k = 2...10$.

[97]. The distribution of the highest correlating molecules with each cluster, tentatively identified as PC 36:1 [26] for the white matter (yellow) cluster, PC 32:0 [26] for the grey matter (blue) cluster and haem [88] for the tissue edge (red) cluster, in both 2D and 3D are shown in Figure 3.7.

$$\rho_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}} \tag{3.11}$$

The selection of only informative m/z bins [104] would have reduced the data set sufficiently to be analysed with *princomp*. Clustering with k-means ($k = 3$) on the reduced dataset produced the same clusters as shown in Figure 3.6, however a review of the peak lists showed that $m/z\,616$ was discarded by this reduction method. This peak was actually found to contribute significantly to the red cluster in Figure 3.6. Examination of ion images of this peak showed that it was found predominantly in a region in which commonly detected endogenous species were observed to be of unusually low ion counts. This regional suppression is likely to have been caused by contamination with blood, during either the organ excision, or sectioning/mounting procedures. The noisy background resulted in this peak being discarded. This observation would have been difficult to make without an

unsupervised tool such as PCA (specifically one which can handle retention of the entire peak list), requiring manual inspection of every possible ion image and comparing to the reduced dataset while considering the anatomy of the sample.

The benefits of applying clustering algorithms to a large (50 GB) 3D kidney dataset have been described recently [97]. The described data contained a very large (512495) number of pixels and so peak picking was performed prior to clustering to reduce the data sufficiently, which involved discarding peaks if they appeared in less than 1% of the spectra. An alternative method of data reduction would be to use the method presented here, which can cope with arbitrarily many pixels, to reduce the entire dataset of 7677 m/z bins to a small number of principal components that explain at least 99% of the variance. This would reduce the chance of discarding informative peaks that are only present in a small, localised feature that is smaller than 1% of the image size (5124 pixels). As such, this work provides new ways to evaluate the effects, suitability and robustness of peak picking on larger datasets.

Spatial binning [103] may be a useful tool for the reduction in memory requirement (as well as enhancing imaging signal-to-noise and increasing contrast) however this comes at a cost of spatial detail. For cases where spatial binning is performed solely for the benefit of memory reduction the method proposed here will prove valuable. A recent article stated that the choice to only retain 650 m/z bins was "a pragmatic desire to use manageable covariance matrices"[138], however the method presented here has demonstrated that covariance matrices far exceeding 650 m/z bins can be handled with ease, and use of the proposed method would enable handling of the entire 11280 processed m/z bins. Binning in the m/z domain combined with a sparse data matrix representation in order to compute PCA on a 3D ToFSIMS dataset was recently described [89]. In order to preserve the sparsity, and therefore the memory reduction achieved, mean centering and scaling were omitted, however the work presented here does not require the entire dataset to be stored in RAM and so mean centering and scaling can be applied if required.

The iterative accessing and processing of the data increases the computation time

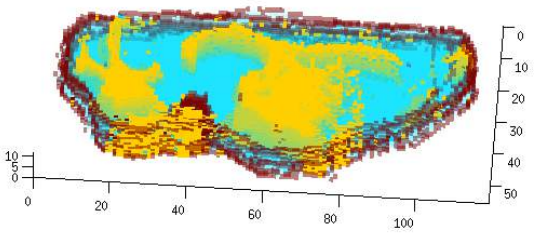


Figure 3.6: 3D representation of on tissue clusters determined with k-means (k=7) applied to a MALDI mass spectrometry image of 12 serial sections of mouse brain after being reduced by PCA (with 99.14% of the variance retained in 40 principal components).

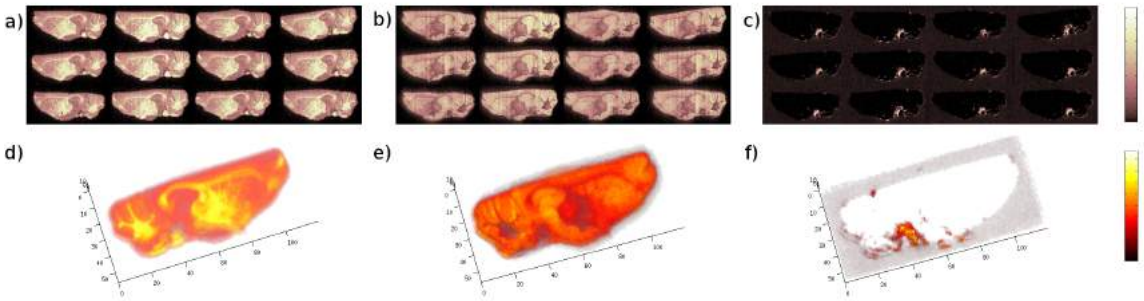


Figure 3.7: (a-c) Selected ion images which correlate highly with each on tissue cluster distribution determined from k-means of serial sections, shown in 2D representation. (d-f) 3D representations of ion images where the alpha channel (transparency) is proportional to the intensity. (a,d) m/z 826 (PC 36:1 $[M+K]^+$). (b,e) m/z 734 (PC 32:0 $[M+H]^+$). (c,f) m/z 616 (haem $[M+H]^+$).

required to process each dataset when compared with methods which retain the entire dataset in RAM, like *princomp*. However, if the dataset has already been pre-processed and reduced to peak lists, either through using automatic data reduction methods described by [86] and [104] or by manually selecting known peaks [139], then the amount of processing required is reduced by eliminating the 'Pre-process spectrum' steps from the workflow shown in Figure 3.1. For small MSI datasets, where the number of pixels and detected/retained peaks falls below the curve in Figure 3.2, standard implementations of PCA will outperform (in terms of speed) the approach described here. As a comparison, the time required to load, pre-process and perform *princomp* on the single rat brain image (shown in Figure 3.3) was 5.7 minutes whereas the time required to perform the new methodology was 13.2 minutes. The reason that the time is doubled is due to the requirement to read in and process the data a second time to calculate the scores. Eliminating the pre-processing step reduced the time to 2.7 minutes and 8.2 minutes for *princomp* and the proposed method respectively. Processing time for the 3D dataset using the proposed method was 24 minutes. Even if the processing time is lengthened, as the size of the data increases (in both the number of pixels and in mass resolution) the need for data processing routines that can handle increased data sizes becomes more apparent. Furthermore, general purpose programming on graphical processing units (GPGPU) could be employed to reduce the time taken as the generation of the summarisation matrices is highly parallelisable [121].

This method inherently provides the ability to compute the covariance and correlation matrices in a more memory efficient way, enabling rapid determination of co-localised m/z peaks on larger data sets than previously reported [140]. This type of investigation requires prior knowledge of an m/z of interest and so would be beneficial in pharmaceutical studies where the m/z of the drug is known and molecules, such as metabolites, which co-localise with the drug are of specific interest.

Despite this article being focused solely on MSI data, the method presented here can be applied to any analytical technique that has sufficiently large data such that memory

limitations have become problematic.

3.5.6 Memory Efficient Scaling

The results of PCA are extremely sensitive to the type of scaling (or lack thereof) applied to the data [141]. Tyler *et al.* [141] proposed 4 different scaling methods, auto-scaling (each peak divided by its standard deviation), root mean scaling (each peak divided by the root mean of its intensities across the image), filter scaling (dividing each peak by the standard deviation of a set number of pixels near that peak, excluding pixels that have zero counts) and shift variance scaling (dividing each peak by its standard deviation in the shift matrix where the shift matrix is calculated by subtracting X from a copy of itself which has been shifted by one pixel vertically and/or horizontally). Of these, auto-scaling, root mean scaling and shift variance scaling can be implemented in a memory efficient manner through the use of iteratively generated means as demonstrated in Algorithm 3.7. This method only requires a single spectrum, plus the scaling spectrum, to be stored within memory at any one time.

Algorithm 3.7. Memory Efficient Scaling Generation

Require: Preprocessing workflow $W \leftarrow \{w_1(S), ..., w_n(S)\}$ containing n preprocessing methods from Section 1.6
Require: Image generation method $G(S, m_{\min}, m_{\max})$ from Section 2.5.6
Calculate spectral representation S_R using Algorithm 3.4
Peak pick on S_R using chosen method from Section 1.6.5 to get m/z limits $M_{\min}$ and $M_{\max}$
$N \leftarrow 0$
$M \leftarrow 0$
$M' \leftarrow 0$
for each spectrum index i in dataset **do**
    Read in spectrum S_i from disk
    for each preprocessing method w in W **do**
        $S_i \leftarrow w(S_i)$
    end for
    if Shift variance scaling **then**
        Read in and preprocess spectrum as above for the spectrum that is horizontally and/or vertically shifted S'_i
        $S_i \leftarrow S_i - S'_i$
    end if
    $N \leftarrow N + 1$
    $\delta \leftarrow S_i - M$
    $M \leftarrow M + \frac{\delta}{N}$
    $M' \leftarrow M' + \delta(S_i - M)$
end for
if Auto-scaling or shift variance scaling **then**
    $scaling \leftarrow \sqrt{\frac{M'}{(n-1)}}$
elseRoot mean scaling
    $scaling \leftarrow \sqrt{M}$
end if

3.6 Conclusions

Data processing is an essential and extremely challenging aspect of mass spectrometry imaging research. The highly multivariate nature of the technique poses challenges in both the limits of data size and the ease of information extraction from imaging experiments. We have presented a means of handling the complete data without discarding potentially useful information. These methods will become increasingly important as efforts towards complete and unsupervised review of large 3D image datasets continues. The memory efficient workflow described here provides, for the first time, a means of performing PCA on extremely large MS image data. The methods also allow for a comparison of data reduction techniques and a means of comparing discarded (or retained) peaks with a complete list. This will facilitate ongoing efforts in data reduction tools. Although the methods here can be performed on complete data, data reduction may still prove important in increasing throughput, where smaller sets can be processed in significantly shorter time frames. Speed of data processing remains an ongoing challenge and current bottle neck in the adoption of this technique in clinical settings. The methods described here will ensure a robust selection of data in such instances, increasing the likelihood of MALDI MSI being adopted in clinical laboratories.

CHAPTER 4

HIERARCHICAL COMPOSITION OF MSI DATA

4.1 Introduction

The ability to determine structure within a mass spectrometry image and to be able to relate that to *a priori* knowledge such as histology, labelled atlases, molecular databases and previously acquired data is becoming increasingly important as mass spectrometry imaging moves closer to the clinic.

A similar challenge exists in computer vision, where automatic recognition of objects and their corresponding category is desired. This can be achieved by having a huge library of categorised images (sometimes referred to as a dictionary) and then matching each of these to the image to identify the most likely match. However this is extremely inefficient in terms of both storage of the image library as well as computationally in the matching process. A more efficient approach has been proposed where each object is learnt as a hierarchy of smaller parts which may be shared across multiple objects [142]. This approach significantly reduces the size of library of objects as each higher level can be stored as a series of references to each of the parts in the lower levels that combine together to produce the high level object. The matching process is also more computationally efficient, as the search space can be systematically reduced through the removal of branches containing parts that do not match.

Although this technique has not yet been applied to mass spectrometry imaging,

there exist features within the data that make this an attractive option to pursue. The hierarchical nature and shared parts maps well to mass spectrometry imaging data. For example, a hierarchy could be used to represent the way in which a given molecule is detected, with the intact molecule at the root of the hierarchy, then adducts, clusters and fragments forming the branches on the next level, repeating down the hierarchy for subsequently smaller fragments. While these features of the data could be considered redundant in that they represent the same molecule detected in different states, they have been shown to aid identification [143, 26]. As MSI data is essentially a series of ion images, many of which show similar or partially similar spatial distributions, the shared parts data representation used in computer vision could capture this redundancy well.

A further benefit is that compositional hierarchies provide an efficient way to represent relationships between multiple objects in a system and can be generated in an unsupervised and incremental manner. Unsupervised methods of analysis are generally preferred in MSI due to the large amount of data generated in a single experiment. Once generated, hierarchical representations enable fast indexing and matching of features in unseen data sets which would provide a rapid way to summarise and annotate new data based on previous observations.

This chapter presents algorithms that generate compositional hierarchies from mass spectrometry imaging data, which specifically take into account the additional m/z dimension of the data. In the computer vision literature the parts are generated through the use of edge detection and are built bottom up, however in MSI there is not just a single image that is being processed and so edge detection would need to be performed on every ion image for a direct application of those methods. It is therefore proposed that the spatial parts are learnt in a top down approach, by segmenting each level based on the spectral dimension resulting in a spatial hierarchy. A bottom-up approach is also presented, where ion image groups are sequentially combined to form larger groups, forming a spectral hierarchy potentially capturing isotopic patterns, adducts, clusters and fragments.

4.2 Generation of a Spatial Hierarchy

A MS image of mouse brain [21] was reduced to a data matrix, D (pixels x peaks; 7378 x 2892), using the memory efficient methods described in Section 3.4 where peaks were detected using the gradient method (threshold of 5) on the total spectrum generated using dual pass SavitzkyGolay smoothing (window size = 9). A spatial hierarchy was then generated using Algorithm 4.1. The highest intensity peak in the data is grouped with all other peaks that have a correlation value higher than a user defined threshold (in this case 0.8). Then, the average ion image from all members of the group is calculated and is separated into two clusters using k-means. The data is then split based on the clusters and the process repeated on each subset of the data. If a cluster is smaller than a user defined threshold or a previously processed cluster has a similar distribution (as determined by Equation 4.3) then the hierarchy generation ends for that branch. Then the next highest remaining peak in the total spectrum is determined and the grouping and hierarchy generation is repeated.

This bears some similarity to an already well established method of generating a spatial hierarchy, bisecting k-means, where at each iteration the data is clustered into two groups and then split accordingly [144]. However in the proposed method multiple branches can be generated at each point in the hierarchy, rather than only two, which are identified based on unique spatial distributions, rather than differing spectral profiles.

$$\Sigma_{ij} = \frac{1}{n} \sum (D_i - \mu_i)(D_j - \mu_j) \tag{4.1}$$

where Σ is the covariance matrix, D is the datacube, D_i is the i^{th} ion image in D and μ_i is the mean intensity of ion image D_i.

$$\rho_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}} \tag{4.2}$$

where ρ is the Pearson product-moment correlation coefficient matrix, discussed previously in Chapter 3.

$$s(m_1, m_2) = |\sum m_1 \wedge m_2| - \max\left(\sum m_1, \sum m_2\right)| \tag{4.3}$$

where m_1 and m_2 are binary images, $\wedge$ is binary AND operation and s is a measure of the difference in the number of pixels with the value of 1 in either m_1 or m_2.

Algorithm 4.1. Generation of a Spatial Hierarchy

```
Reduce data to datacube D
M ← {}

function GenerateNextLevel(D, M, ρ_t, t)
    Calculate correlation matrix ρ using Equation 4.2
    Calculate total spectrum T from D
    Sort T in descending order of intensity to produce indices list I
    G ← {}
    while I¬{} do
        Create group g from all members of I that have correlation > ρ_t
        Insert g into G
        Remove members in g from I
    end while
    for each g in G do
        Calculate mean ion image m of all members in g using D
        Cluster m into 2 clusters c_1 and c_2 using k-means
        if |c_1| > t then
            if c_1 not in M and no member m of M has score s(c_1, m) < t from Equation 4.3 then
                Insert c_1 into M
                Extract members of c_1 in D into D'
                G' ← GenerateNextLevel(D', M, ρ_t, t)
                Associate next level in hierarchy G' with g
            end if
        end if
        Repeat above for c_2
    end for
    return G, M
end function
```

4.2.1 Results and Discussion

The mean ion image and spectrum for each of the 11 groups generated at the first level of the hierarchy are shown in Figure 4.1. This identifies unique spatial distributions and corresponding spectra that can be related to experimental artefacts (the matrix region in group 1 and the halo effect in group 8) and anatomy of the sample (white matter in group 9) and also the unknown clustered region from Chapter 3. In Chapter 3, *m/z* 616 (tentatively identified as heam) was identified as a significant contributing factor to this region. This was achieved by performing Pearson's correlation on the cluster distribution and every ion image within the data to identify ions with a high correlation value, which

implies co-localisation with the cluster. The first level hierarchy provides an identical result, with the two members of group 11 being m/z 616 its isotope m/z 617.

The restriction on groups to have at least two member peaks prevents noise being factored into the hierarchy and generating hundreds of groups with only a single ion image. This can further be justified by the fact that any detected ion should have a corresponding isotope peak separated by a nominal mass unit. In some cases the isotope peak may have a different spatial distribution due to unresolved ions, however due to the fact that adducts and fragments are included too it is unlikely that any one ion image will have a truly unique spatial distribution in a given dataset.

The next level of the hierarchy and selected accompanying k-means cluster results used to split the previous level's data are shown in Figure 4.2. The splitting of group 11 only resulted in one cluster that was larger than the threshold (set at 10 pixels), which was associated with everything but the region of interest. However, at level 2 of the hierarchy the same spatial distribution has been identified in group 17 and when k-means is applied to this mean ion image two clusters greater than threshold are generated, one that co-localised with the region to be identified from Chapter 3 and another with the rest of the tissue region. This is due to the fact that as the levels increase the data being processed is getting smaller and more refined (i.e. just the on tissue pixels in level 2 compared to the whole image in level 1) so small features become larger proportionally resulting in better clustering results.

The groups shown in level 1 in Figure 4.2 show the global, unique, spatial distributions present within the data. As the input data is limited to just the on tissue pixels for the next level of the hierarchy, more unique spatial distributions are identified. This provides a rapid overview of the data and can easily be correlated with histology or labelled atlases to provide annotations for the data.

Once a hierarchy has been generated, it is possible to query which groups contain an ion of interest, to determine which other ions co-localise with it. This is demonstrated by selecting a group from levels 1-3 that contains m/z 616 as identified in the exploratory

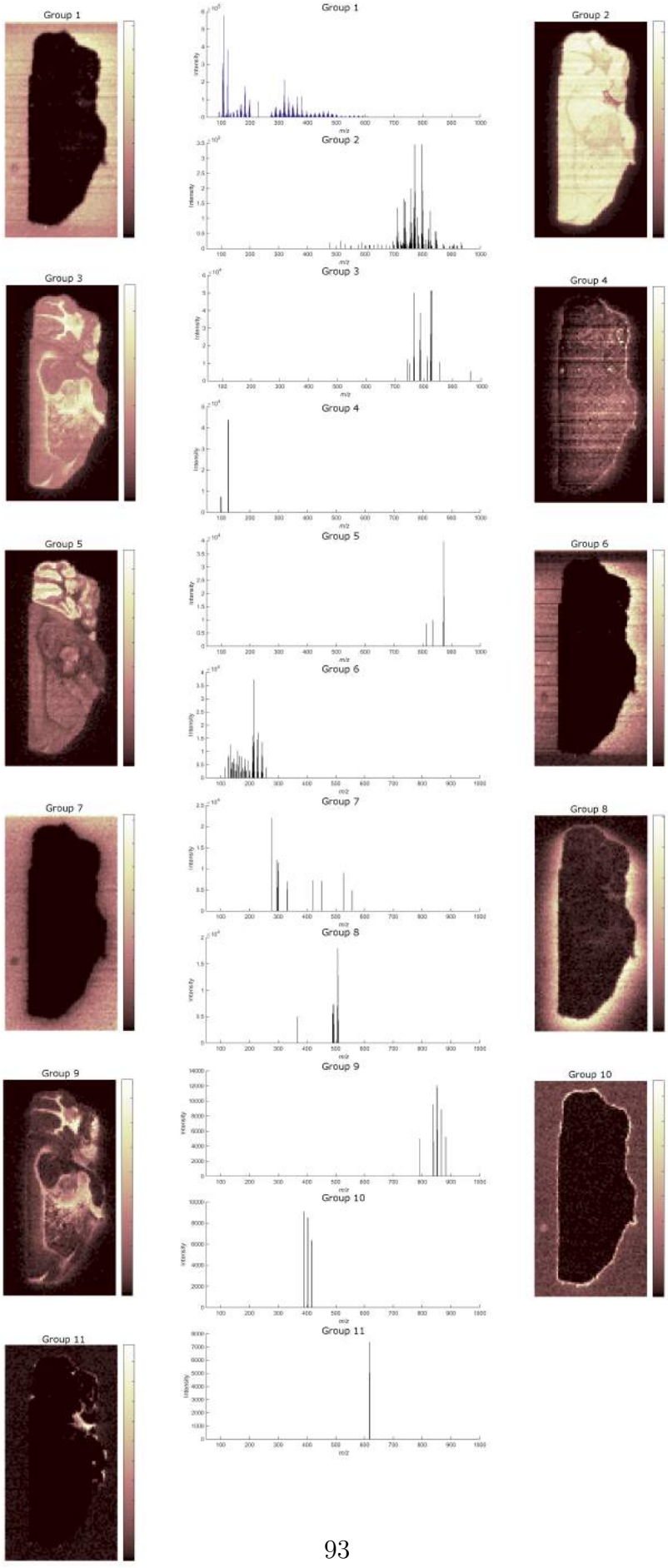


Figure 4.1: Level 1 hierarchy

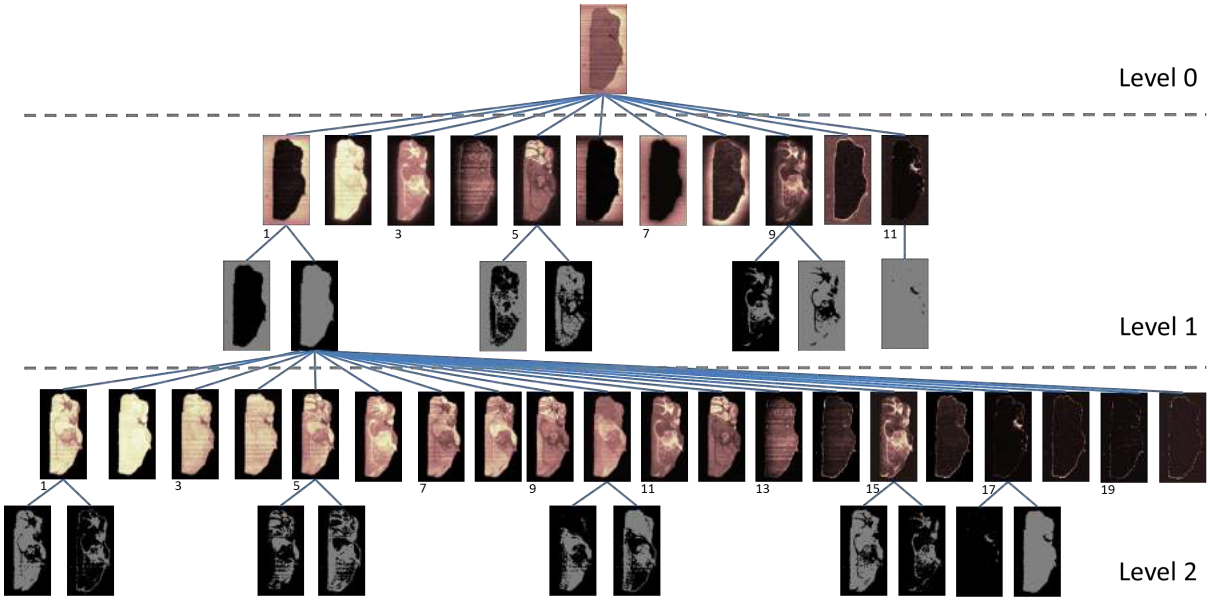


Figure 4.2: First 3 levels of the hierarchy generated using Algorithm 4.1. Masks used to split the data shown in grey. Mean ion image of each group shown at each level of the hierarchy.

analysis in Chapter 3, shown in Figure 4.3. As discussed previously, level 1 produced the same result as performing Pearson's correlation on the entire data, however the groups on levels 2 and 3 identify significantly more peaks that co-localise with this region. These additional ion images were missed previously due to background noise within the matrix region (possibly caused by unresolved matrix peaks) lowering the calculated correlation value.

The method proposed here provides a rapid overview of spatial distributions present within the data (from both the mean ion image of each group and the masks generated from these) and route to determine a series of ion images that co-localise with an ion of interest.

4.3 Generation of a Spectral Hierarchy

The spatial hierarchy generated previously can successfully describe and summarise a given dataset, but is unique to the input data and so cannot be applied to subsequent data easily. To be able to apply to another MS image, either the new image needs to

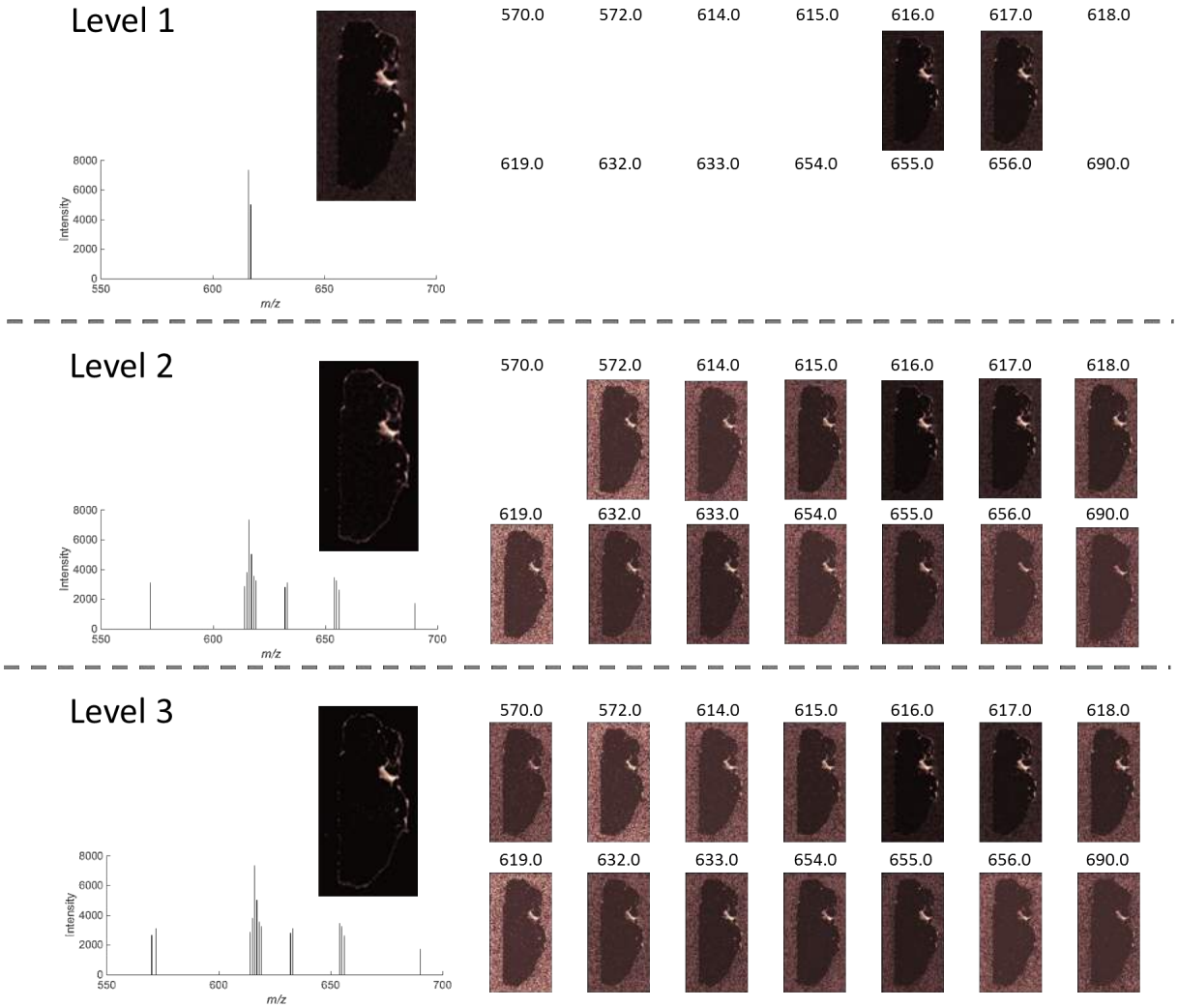


Figure 4.3: Total spectrum and mean ion image (left) from a group from the first three levels of the hierarchy that contains the ion m/z 616. Ion images of each member of each group (right) where the level 1 group contains 2 members, level 2 group contains 13 members and the level 3 group contains 14 members.

be registered to the original image to allow direct comparison or each generated part (or mask) must be scaled, rotated and translated across the corresponding mean ion image of the group to determine if a match is found. This would be computationally prohibitive and thus limits the usefulness of this approach [142]. Furthermore, this would not cater for the case where a tissue section has been torn or otherwise deformed as part of the sectioning and thaw mounting process, as the shape of the feature represented by a given mask may have been significantly altered through this process.

A more beneficial scenario would be to determine a hierarchy that corresponds to the detection states of a given molecule, and then the application of that hierarchy to another dataset would produce a metric to determine the probability that that molecule was also detected in the same states in the new dataset. This would enable rapid evaluation of a library of hierarchies to summarise a dataset, making evaluation and interpretation much simpler for the analyst.

Algorithm 4.2. Generation of a Spectral Hierarchy

Reduce data to datacube D
Variance scale D
Assign each ion image in D to its own group g_0
$G_0 \leftarrow \{\}$
Insert all g_0 into G_0
Calculate correlation matrix ρ using Equation 4.2
$G_1 \leftarrow \{\}$
for each g_0 in G_0 **do**
    Pair g_0 with highest correlating group in G_0 to form g_1
    Insert g_1 into G_1
end for
$l \leftarrow 1$
while $|G_l| > 1$ **do**
    $G_{l+1} \leftarrow \{\}$
    for each g_l in G_l **do**
        $max_{icc} \leftarrow 0$
        $max_g \leftarrow \{\}$
        for each g in $G_{0 \ldots l-1}$ **do**
            $g' \leftarrow g \cup g_l$
            Calculate ICC for g' using ICC(C,1) from Equation 4.4
            if ICC $> max_{icc}$ **then**
                $max_{icc} \leftarrow$ ICC
                $max_g \leftarrow g'$
            end if
        end for
        Insert max_g into G_{l+1}
    end for
    $l \leftarrow l + 1$
end while

A spectral hierarchy using the same reduced MS image generated in Section 4.2 was generated using Algorithm 4.2. In contrast to the spatial hierarchy, the spectral hierarchy

is generated bottom up, starting with each individual ion image forming its own unique group and as the level in the hierarchy increases groups are made by combining each group from the previous level with the group from any lower level that results in a supergroup with the highest correlation. Pearson's correlation can be used, as before, to compare two ion images, however is not suitable as a metric when comparing groups consisting of three or more members.

Intraclass correlation coefficients (ICC) are used to assess the consistency, or the conformity, of measurements of the same quantity made by multiple observers [145]. Therefore, it provides a good metric for use to determine the likelihood that a given set of observations (in this case ion images) have similar spatial distributions, implying they represent the same feature. The equation for calculating ICC(C,1) is given in Equation 4.4, where MS_R is the mean square for rows (in this case a row consists of the intensities for a given pixel in each ion image), MS_E is the mean square error and k is the number of ion images. The models used to construct ICCs assume equal variance [146]. Any two given ion images are unlikely to have equal variance due to variations in signal intensity. For example, naturally occurring isotopes of a given molecule will have identical spatial distributions but the peak magnitudes will have a ratio determined by the natural abundance of constituent isotopic elements, resulting in different variance. To correct for this, each ion image must be variance scaled such that the variance image is 1 prior to hierarchy generation.

$$ICC(C,1) = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E} \tag{4.4}$$

The algorithm presented in Algorithm 4.2 generates a single hierarchy for the entire dataset. The hierarchy can then be pruned at a user defined ICC threshold, resulting in a series of smaller hierarchies which then forms a library representing a dataset. A more efficient implementation, at the cost of generating the global hierarchy, is to include a

constraint on the generation of a group for the current level of the hierarchy and only generate if max_{icc} is greater than a user defined threshold. An example of one such small hierarchy is shown in Figure 4.4.

4.3.1 Applying Spectral Hierarchy to Additional Datasets

As the generated hierarchy does not contain any spatial information, no image registration is required to be able to apply the same hierarchy to an additional dataset. However, to produce accurate results the same ion image must be being considered and compared in each dataset. Due to slight drift in the instrument over time, poor calibration or an artefact of preprocessing methods any two spectra of the same analyte (even when acquired sequentially on the same instrument) may not be truly aligned and so may result in mismatched or unmatched peaks when a direct comparison is performed.

The equation for mass resolving power is given as

$$R = \frac{m}{\Delta m} \tag{4.5}$$

Rearranging this equation allows the calculation of a mass tolerance (Δm) at a given m/z for an instrument with a given mass resolving power results in

$$\Delta m = \frac{m}{R} \tag{4.6}$$

Mass resolving power can be used as a tolerance limit to match peaks between two datasets. Assuming that both datasets are properly calibrated, if two detected peaks, one from each dataset, fall within the mass tolerance of the instrument used then they can be safely matched together because if they had been detected within the same spectrum then they would be unresolvable.

Algorithm 4.3. Applying Spectral Compositional Hierarchy To New Data

1: Generate hierarchy for datacube D using Algorithm 4.2
2: Reduce new dataset to datacube D'
3: $P \leftarrow \{\}$
4: **for each** detected peak p in D **do**
5: Calculate Δm for p using Equation 4.6
6: Find peak p' from D' that has smallest difference to p
7: If $(p - p') < \Delta m$ insert peak pair (p, p') into P
8: **end for**
9: **for each** entry in P that contains a non-unique peak **do**
10: Remove non-unique entry from P based on lowest co-probability value using Equation 4.8
11: **end for**
12: Reduce D and D' to D_r and D'_r respectively based on peak pairing entries in P
13: Create group g' by extracting members of hierarchy H from D'_r
14: Calculate ICC of group g' using ICC(C,1) from [146]

Co-probability

In some rare cases this matching produces multiple pairings containing the same peak from the second dataset. Unless drastically different mass resolution instruments were used, it is unlikely that all of the peaks in the second dataset correspond to the same ions detected as a single peak in the first dataset. To determine which of the peaks in the second dataset is more likely to correspond to the peak in the first dataset, a metric of co-probability is proposed in Equation 4.8.

$$c = \frac{1}{\sum_{i=1}^{N} f(I,i)} \sum_{i=1}^{N} f(I,i) f(J,i) \tag{4.7}$$

$$f(I,i) = \begin{cases} 1, & \text{if } I_i > t \\ 0, & \text{otherwise} \end{cases} \tag{4.8}$$

This determines a measure of similarity, c, between two images (I and J which each have N pixels) as the number of pixels that are above threshold, t, at the same location in both ion images as a fraction of the maximum number of pixels that are above threshold in either image. Co-probability can be used in two ways. Consider the simple case where the two datasets to compare are already registered and so the pixels in each image correspond to comparable anatomy. Co-probability can then be directly applied to the ion image from each image to compared. In the more complex, and also more common, case the analyte

position within the images may be at slightly different angles, relative locations or scales making the direct application inappropriate. In this case, a co-probability 'spectrum' can be generated for each dataset by calculating the co-probability between the ion image to match and all others within the data. If these spectra are similar then it implies that similar peaks have similar relative distributions within the individual datasets. This provides a way to determine the likelihood that a peak pairing is appropriate between unregistered datasets, avoiding the difficult challenge of automated image registration.

Hierarchy Application to Real World Data

Hierarchies were generated using Algorithm 4.2 on one section from a series of 14 serial sections of mouse brain [21]. The generated hierarchies were then applied to data acquired on the adjacent section and ICC values were calculated at each level of the hierarchy, shown in Figure 4.4. The resulting correlation values are comparable for every group, implying that the ion images selected have similar spatial distributions within their respective datasets. It should be noted that this does not ensure that the spatial distributions are identical in both datasets, only that the ion images that form the hierarchy correlate with one another in each of the two datasets separately.

This was then taken a step further and applied to a MS image of rat brain acquired using a different matrix (CHCA instead of para-nitroaniline) [4]. Again, comparable correlation values are produced. This shows that the generated hierarchies are useful for determining whether a given set of biologically relevant ions are present and have similar spatial distributions regardless of the matrix used. When applying the hierarchies that represented ions corresponding to para-nitroaniline to the CHCA dataset then either peaks were missing (determined from the lack of a peak pairing) preventing the hierarchy from being applied or the resulting ICC values differed significantly indicating that the detected peaks were not representing the same structure in the second dataset.

The final dataset that this hierarchy was applied to was an MS image of formalin fixed rat brain acquired with CHCA [4]. The application of the same hierarchy resulted

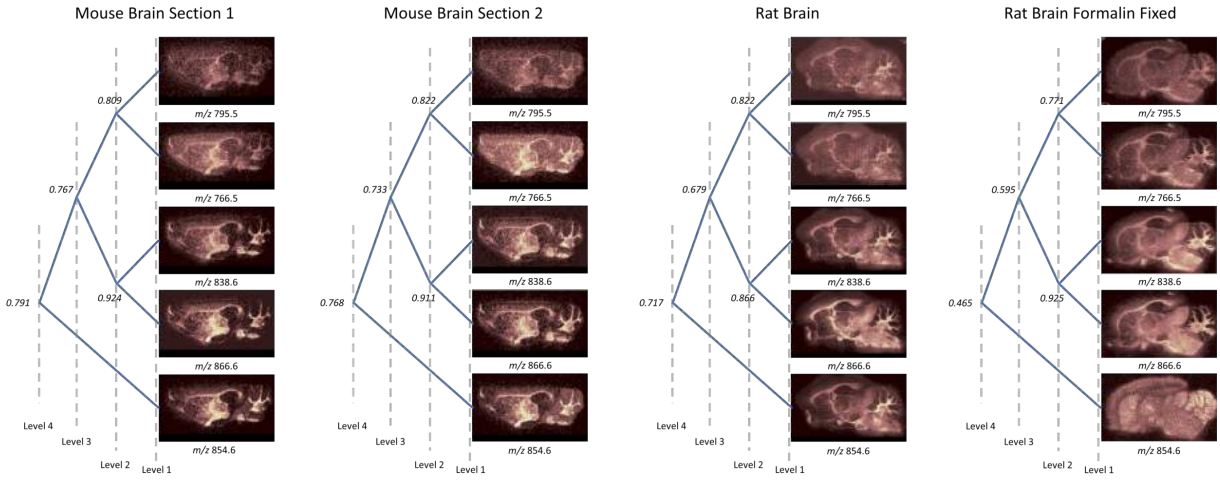


Figure 4.4: Selected level 4 hierarchy generated from a section of mouse brain acquired with para-nitroaniline [21], shown left. Ion images are shown at the leaf nodes and correlation values are shown at all parent nodes. This hierarchy was then applied to each of the additional data sets, serial section acquired in the same manner [21], section of rat brain acquired using CHCA [4] and a section of formalin fixed rat brain acquired using CHCA [4].

in a significantly different ICC value at level 4, shown in Figure 4.4. The lower value is a result of a different spatial distribution of m/z 854.6 when compared with other ion images within the hierarchy. This identifies that there is something different about one of the branches of the hierarchy, indicating that there is something different about the collected data. Although this method does not elucidate the underlying difference, in this case it is likely due to changing adduct formation due to the formalin fixation process resulting in the detection of different ions in different relative abundances.

The hierarchy's mean spectra (discounting pixels which have an intensity of 0 for a given peak) for each of the datasets are shown in Figure 4.5. The first three spectra are comparably similar in their relative abundances of each of the ions contributing to the hierarchy, however significant differences are observed in the spectrum associated with the formalin fixed tissue. When looking at the ion images alone, it was not obvious that there was any difference for the first 4 ion images (m/z 795.5, 766.5, 838.6 and 866.6) as the correlation values and spatial distributions were similar to the other datasets. However, the spectral profile show significant differences, with a shift in the basepeak from m/z 854.6 (unidentified in [4]) to m/z 838.6 (identified as PC 38:1 + Na in [4]). This further

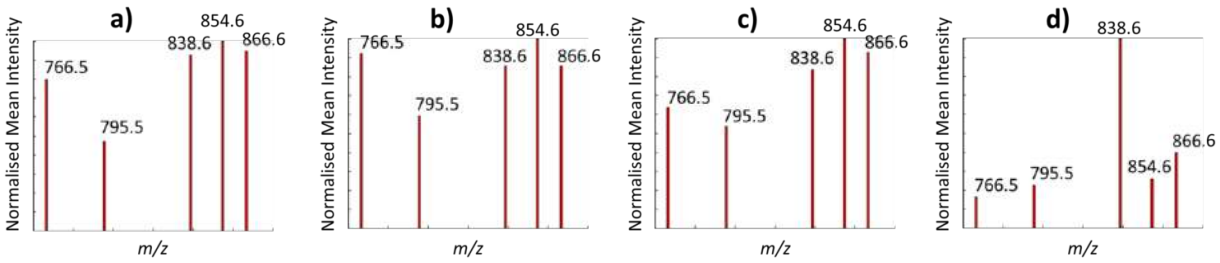


Figure 4.5: Spectral profile of the hierarchy defined in Figure 4.4 for a) a section of mouse brain acquired with para-nitroaniline [21], b) serial section to a), c) section of rat brain acquired using CHCA [4] and d) section of formalin fixed rat brain acquired using CHCA [4]. The spectral profile is calculated by determining the mean non-zero ion intensity for all pixels. Spectra a-c) all have very similar spectral profiles, whereas d) has a clearly different profile, indicating that the hierarchy does not fit this data set.

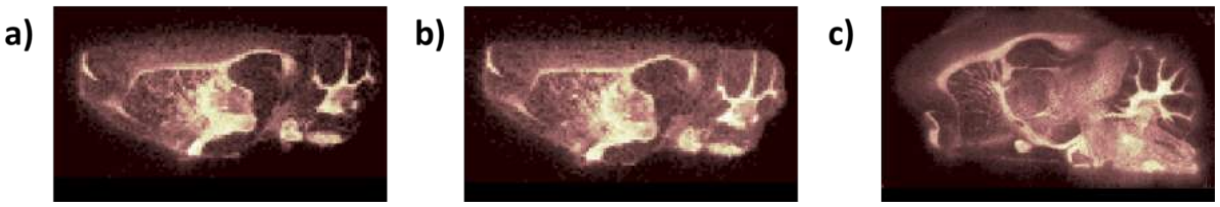


Figure 4.6: Improved signal-to-noise image of the spatial distribution represented by the hierarchy in Figure 4.4 for a) a section of mouse brain acquired with para-nitroaniline [21], b) serial section to a), c) section of rat brain acquired using CHCA [4]. The section of formalin fixed rat brain acquired using CHCA [4] was not included as it was determined that the hierarchy was not applicable.

shows a difference between the first three datasets and the formalin fixed dataset.

As hierarchies are generated from ion images that have the same spatial distribution, summing together the individual ion images results in a distribution map that has increased signal-to-noise (SNR) when compared to any one individual ion image alone, shown in Figure 4.6. This was not performed on the formalin fixed dataset, as the correlation values and spectral profiles were inconsistent and so summing these images would be inappropriate.

4.4 Conclusions

All algorithms proposed here rely on having a pre-generated datacube in memory. The spatial hierarchy could make use of memory efficient methods for the creation of the cor-

relation matrix at each level of the hierarchy. This would increase the number of pixels that could be included in the analysis at the cost of computational speed. As the current implementation requires at least 5 minutes to generate up to the third level of the hierarchy, the additional overhead introduced through the disk access of the memory efficient methods described in Chapter 3 would be prohibitively high. As each node of the hierarchy can be processed independently of all other nodes, it would be possible to parallelise the hierarchy generation using multithreaded or distributed computing implementations. The use of distributed computing with a distributed file system would help mitigate the performance hit attributable to the memory efficient routines, potentially making them useful in their application to the spatial hierarchy generation algorithm presented here.

CHAPTER 5

OPTIMISING COLOUR SCHEMES FOR HUMAN COLOUR PERCEPTION

5.1 Introduction

The typical presentation of mass spectrometry imaging data is in the form of an ion image which can be generated by integrating over a given peak (or m/z range) in every pixel within the image. This produces a 2D spatial distribution of the selected ion, assuming well resolved peaks.

Data visualisation is defined as the process of generation and presentation of a visual representation of data [147]. This is vitally important in MSI, as conclusions are drawn based on the relative intensities of perceived structure(s) present within ion image(s). The intensity scale dictates the relative measured concentration of a given analyte which, as stated above, can span many orders of magnitude presenting a difficult visualisation task.

Due to the large amount of information present in a single MSI data set, presentation of multiple ion images at once is often desirable to determine co-localisation of molecules. This is typically presented as RGB composite images, where each colour channel (red, green and blue) is assigned to a different ion image, with the resulting colours indicating pixel-wise proportional contributions of each analyte.

The style and colour selection used for display has a dramatic effect on the perceived structure within the data [148]. Different aspects of the colour signal communicate dif-

ferent characteristics of the data, and to properly take this into account when defining a colour scheme requires knowledge of physical and psychological properties of colour [148]. To reduce the impact of this, visualisation software typically include multiple colour schemes, with one selected as the default, and rely on the user to select the most appropriate. However, users tend to select visually appealing, vibrant colour schemes such as the "rainbow" colour scheme over more appropriate, 'dull and ugly', colour schemes [147]. The rainbow colour scheme, as its name suggests, is based on the visible wavelengths of the electromagnetic spectrum, which in theory seems like a sensible phenomenon to base a colour scheme on. However, multiple articles over the past few decades have illustrated reasons why it should not be used and the ways it can be actively misleading [148, 149, 150, 151].

In order to evaluate colour schemes effectively, it is important to consider colour spaces, human vision and how they relate to one another. The two major photoreceptor cells found in the retina are rods and cones. Rods are extremely sensitive, capable of functioning at low light levels, and are responsible for scotopic vision (vision in low light conditions), but do not provide any colour differentiation. Cones are responsible for colour vision, with three different types present in a normal eye (S, M and L corresponding to short, medium and long wavelength peak sensitivities respectively). There are substantially fewer S cones than there are M or L cones in the eye. Human perception is less sensitive to changes in hue than luminance [147], with high spatial frequencies being perceived through changes in luminance [149].

Various colour spaces have been defined for different purposes. For a more detailed review and description see [152]. Here the focus will be on RGB, as this colour space maps to electronics, and CIELAB, due to its relation to human visual perception.

RGB is a device-dependent colour space based in the human perception of colours, with each channel roughly corresponding to the region of the visible spectrum that the L, M and S cones, respectively, are sensitive to. Disadvantages of the RGB colour space include its psychological non-intuitivity and the perceptual non-uniformity (low correlation between

the Euclidean distance in the RGB space and the perceived difference in colour) [152]. Despite these issues, RGB is the most commonly used colour space for defining colours as it maps directly to electronic displays.

The main goal of the CIELAB colour space was to provide a perceptually equal space, resulting in Euclidean distances between colours in this space correlating strongly with human visual perception [152]. In this space the three axes correspond to lightness (L), the position between red and green (A, where positive values indicate red and negative values indicate green) and the position between yellow and blue (B, where positive values indicate yellow and negative values indicate blue).

This chapter reviews colour schemes that are used in MSI literature, relating them to human vision and perception while also highlighting any artefacts that can be introduced. Recommendations are made for rules to follow in the selection of colours schemes in order to avoid introducing unnecessary artefacts and to produce perceptually accurate representations of data.

5.2 Methods

MALDI MSI data of a 12 μm sagittal section of mouse brain, which was coated with 15 mL of 10 mg mL^{-1} CHCA via airspray, were acquired using a QSTAR XL (AB Sciex) in positive mode over the range m/z 50-1000. The dataset used is available from www.imzMLConverter.co.uk. Data were converted from mzML to imzML using imzML-Converter and subsequent processing was performed using SpectralAnalysis as described in Chapter 2.

A common m/z axis was calculated for the dataset, as described in Section 2.5.1. Each spectrum was smoothed with a dual pass Savitzky-Golay (window size of 9 bins) filter and summed together to generate the total spectrum. Peak detection was performed on the total spectrum using a gradient method, as described in Section 1.6.5. The intensity of each peak was extracted from each spectrum in turn to generate a datacube and output

to imzML.

Memory efficient principal component analysis was performed as described in Chapter 3, with the preprocessing workflow as described above.

Dichromacy simulations were performed using the Vischeck ImageJ plugin [153].

Articles were selected by searching Google Scholar for three keywords (mass spectrometry imaging). Articles primarily consisting of reprinted figures, such as reviews, and articles containing no ion images were omitted. The first 100 articles meeting these criteria were selected. Articles from the current year were also selected, using the same criteria as above except with the limitation that the year in the search was set to 2014. The first 20 articles meeting these criteria were selected. A complete list of the 120 articles evaluated, and which colour scheme(s) were present in each, is included in Table 5.1.

5.3 Results and Discussion

A colour scheme is a series of colours defined in a chosen colour space (typically RGB) that have a corresponding value such that each value within the data can be replaced with, or mapped to, its assigned colour for display. Accurate representation of the structure in the data largely depends on the type of data being visualised and no colour scheme is ideal in all situations. Types of data can be broadly classified as either nominal, ordinal, interval and ratio. Rough guidelines for each type of data have been discussed elsewhere [148]. This chapter expands on these within the context of MSI data visualisation scenarios.

5.3.1 Ion Images

Ion images are the most commonly displayed form of data within MSI literature, mapping the intensity of a selected ion at every spatial location within the acquired image region. After evaluating 120 MSI articles, it was found that over 30 different colour schemes were used, with some articles containing up to 8 unique colour schemes [7]. A selection of the colour schemes present in the literature were applied to the same unnormalised ion image

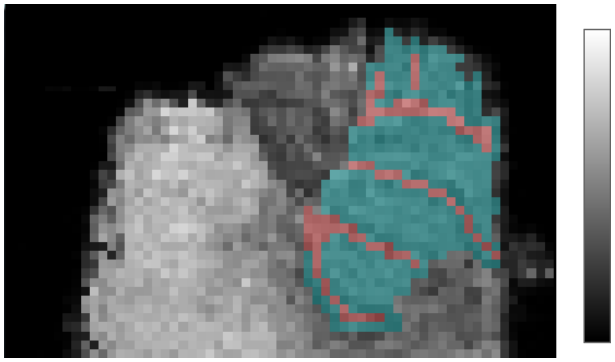


Figure 5.1: m/z 826 with highlighted arbor vitae (red) and cerebellar cortex (cyan).

(mouse cerebellum, m/z 826 with values spanning 0-100 counts) and shown in Figure 5.2. As stated by Rogowitz *et al.* [148], how data are visually represented can greatly influence the perception of the structure within the data. This is especially evident when comparing Figure 5.2a with Figure 5.2o, q and r (at least one of which is present in 34 of the reviewed articles), where contrast is lost between the arbor vitae and the cerebellar cortex (shown labelled in Figure 5.1).

The simplest, most common (appearing in 40 of the surveyed articles), and previously cheapest to publish colour scheme is grayscale (see Figure 5.2a). The benefit of grayscale is that it is a simple task to perceptually order shades of gray based on their lightness, allowing intuitive interpretation of the data. However this comes with a caveat; human perception of brightness is affected by the brightness of surrounding regions [147, 265]. More generally, the perceived appearance of a colour is affected by its surroundings, most notably when they are complementary colours. This effect is referred to as simultaneous contrast.

A grayscale colour scheme, or any colour scheme which only varies luminance, cannot accurately communicate gradual changes (low spatial frequency) in structures present within the data [148]. The number of shades of gray that can be reliably distinguished is dependent on the luminance of the display device used, essentially limiting the number of unique values that can be displayed in a given image [266]. The addition of colour, therefore, is beneficial as it increases the dynamic range of a colour scheme [147].

The most frequently used colour scheme group is the rainbow, or rainbow based, colour

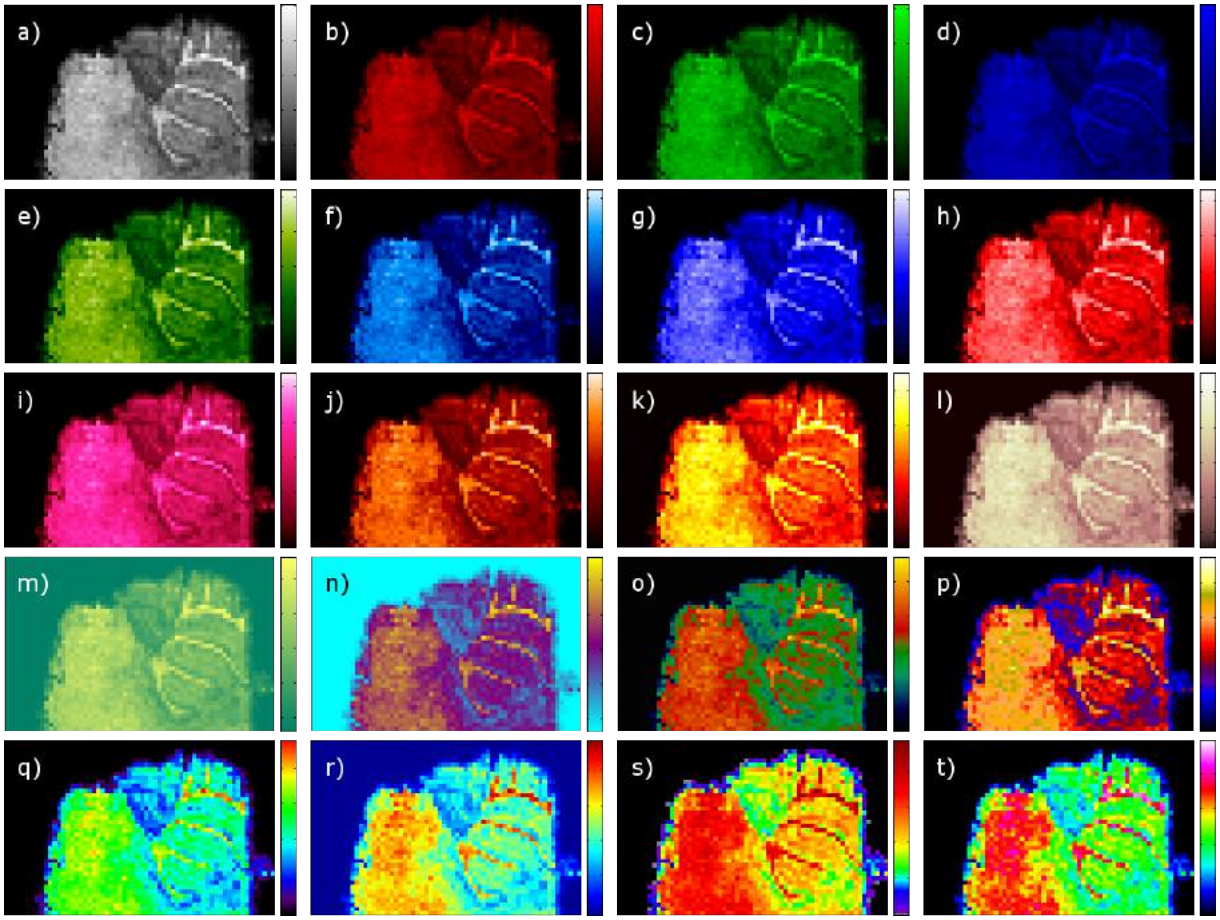


Figure 5.2: Visualisation of the same data (unnormalised m/z 826 from the cerebellum region of a mouse brain) using colour schemes found in MSI literature. Intensity spans from 0 to 100 counts. a) grayscale b) red c) green d) blue e) green to white f) cyan to white g) blue to white h) red to white i) pink to white j) copper to white k) hot l) pink hot m) green to yellow n) cyan to magenta to yellow o) double scale (blue to green, red to yellow) p) temperature based q-t) rainbow based.

Colour scheme	Reference	Count
a)	[154][155][156][157][158][159][160][161][162][163] [164][3][95][165][166][24][167][168][169][170] [171][172][173][174][175][176][177][178][179][180] [181][182][6][183][109][184][185][186][187][188]	40
Inverse of a)	[189][190]	2
b)	[154][129][19][191][7][192][193][194][195][196] [197][198][179][180][181][182][199][200][185][186] [187][201][202]	23
c)	[203][154][160][204][129][205][7][192][193][206] [195][196][197][207][198][179][181][180][182][199] [200][185][186][187][201][202][208]	27
d)	[203][19][7][193][194][206][198][6][199][200] [185][187] [208]	13
Other single colour	[172][19][209][160][204][194][206][207][198][210] [183][199][200][186][202]	15
e)	[211][212]	2
f)	[173][212]	2
g)	[211][213][198]	3
h)	[212]	1
i)	[214]	1
Other single colour to white	[173][182]	2
White to colour	[215][216]	2
j)	[217][161][198][214]	4
k)	[206][218][219]	3
l)		0
Other black-colour-colour	[220][221][222][203][154][164][223]	7
m)	[224]	1
Other colour-colour	[172][95][225][226][164]	5
n)	[157][227][228][229][230][231]	6
o)	[232]	1
p)	[233][234][235][201]	4
q)	[23][227][236][211][237][165][233][129][167][7] [238][207][198][180][239][240][241][242][243][244] [245]	21
r)	[246][247][248][249][250][169][251][193][197][252] [253][201]	12
s)	[254][207]	2
t)	[255][256][19][257][166][258][197][180][252][259] [260][186][202][261][262]	15
Other rainbow	[263][264][244][201]	4
Other	[183][187]	2
Unique single colour		65
Unique single colour to white		11
Unique rainbow		48

Table 5.1: Colour scheme descriptions (letters correspond to the colour scheme displayed at the corresponding location in Figure 5.2) and the articles which used them.

schemes (Figure 5.2q-t) featuring in 48 articles. These colour schemes are so called due to their basis in the colours of each wavelength in the visible portion of the electromagnetic spectrum. There are significant disadvantages with using rainbow colour schemes, which can result in an actively misleading representation of the data [149]. These are excellently demonstrated by Borland *et al.* [149], but will be summarised here. Firstly, there is no perceptual ordering, with experiments showing that people will order rainbow colours in numerous ways [147]. This results in confusion due to the fact that greater-than and less-then relationships are not immediately obvious and must be inferred through memory (error-prone) or consulting a legend (needless distraction). Next, as large regions of the rainbow colour scheme are isoluminant, small detail and sharp features that fall within these regions are obscured. This is evident in Figure 5.2q, where it is difficult, if not impossible, to distinguish between the lower regions of the arbor vitae and the cerebellar cortex. Finally, artefacts are introduced in the form of apparent bands of data resulting from sharp transitions between hues when in reality there are only small differences between the values. This is particularly evident on the tissue boundary in Figure 5.2s. Although this is an artefact of all of the rainbow colour schemes, it is most notable in Figure 5.2s due to the fact that over 50% of the colour scheme is varying shades of red and the remainder includes 6 colour changes.

A further issue with rainbow colour schemes is that perceptual changes in the colours are not uniform, with changes appearing faster in cyan and yellow when compared with green (shown in 5.2q). This further complicates the interpretation of the value of specific colours and can cause false contrast. Attention is drawn to the yellow areas due to their brightness, not necessarily because they are the most important [148].

The most common oversight when considering colours to use when presenting scientific data is the resulting perception by individuals with colour deficient vision [150]. Colour vision deficiency (CVD) refers to the decreased ability, or inability, to see colour and affects 8% of men and 0.5% of women [267]. Anomalous trichromacy, the most common form of CVD, results in one type of cone having a sensitivity shift to one end or the other of the

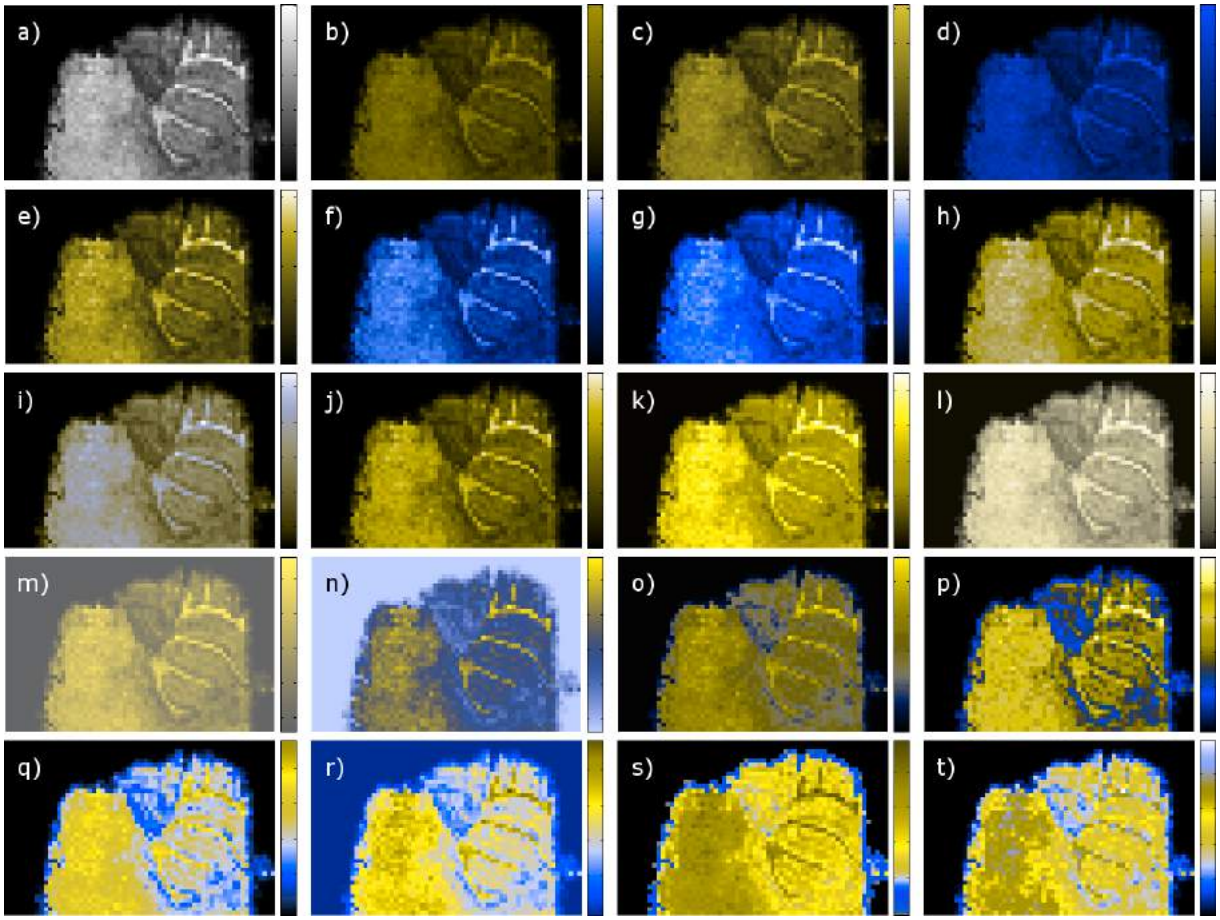


Figure 5.3: Deuteranope simulation of colour schemes found in MSI literature, shown in Figure 5.2. The same colour appears at different values along the colour scale in panels (o-t), rendering it impossible to accurately interpret the data from the visualisation alone.

visible spectrum [268]. Dichromats lack either the L (protanopes), M (deuteranopes) or S (tritanopes) cones, resulting in dichromatic vision. The most severe, and also extremely rare, form of CVD is monochromacy, complete colour blindness.

Simulating deuteranopia, where the M (medium wavelength, roughly centred around green light) cones are absent, for Figure 5.2 further shows the unsuitability of rainbow or rainbow based colour schemes as multiple values are mapped to the same colour. This is especially evident in Figure 5.3t, with the appearance of slight holes in the white matter region of the tissue (dark yellow in Figure 5.3t), as the high intensity pink pixels instead appear as a blue colour, indicating low intensity. This renders the data impossible to accurately interpret for any deuteranope. Similar effects are observed with other forms of dichromacy (shown in Figures 5.4 and 5.5).

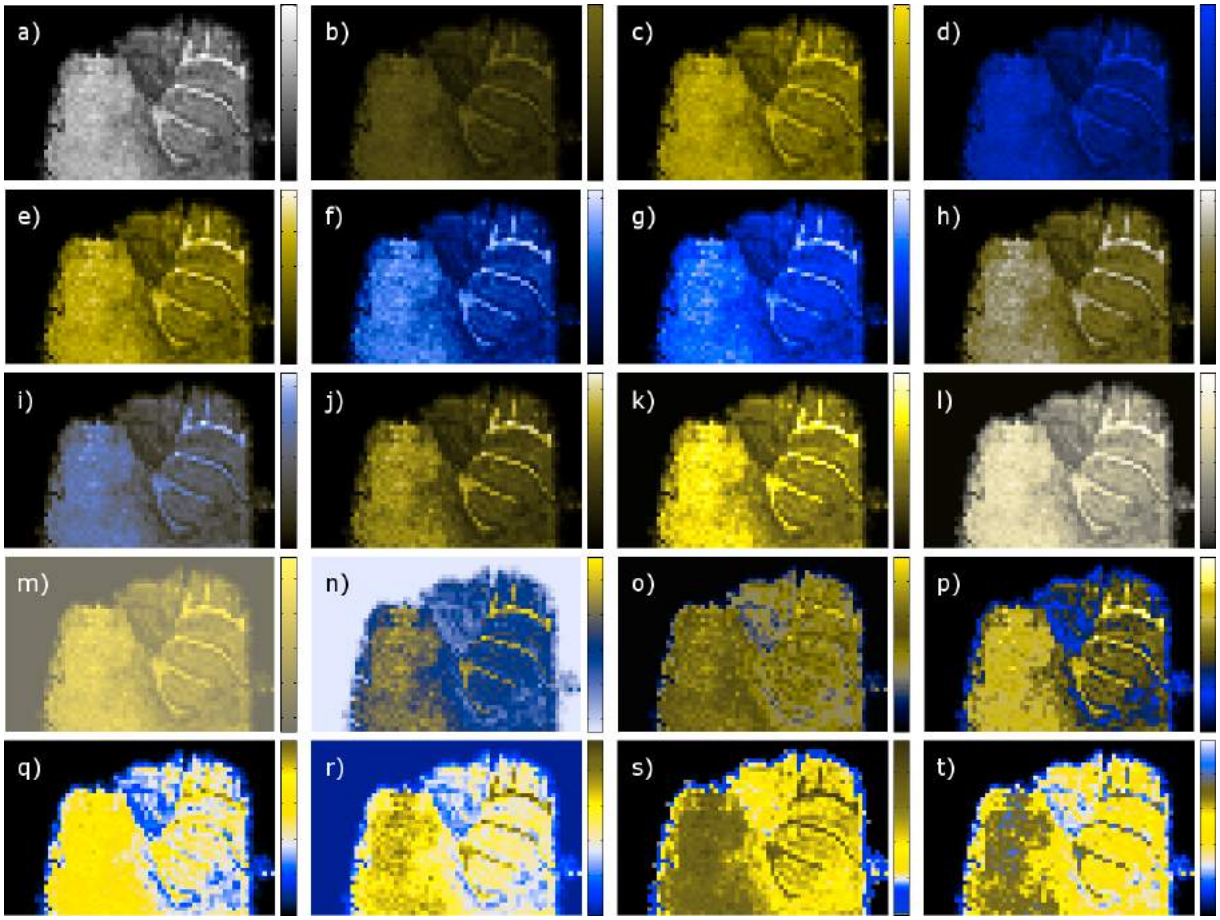


Figure 5.4: Protanope simulation of Figure 5.2.

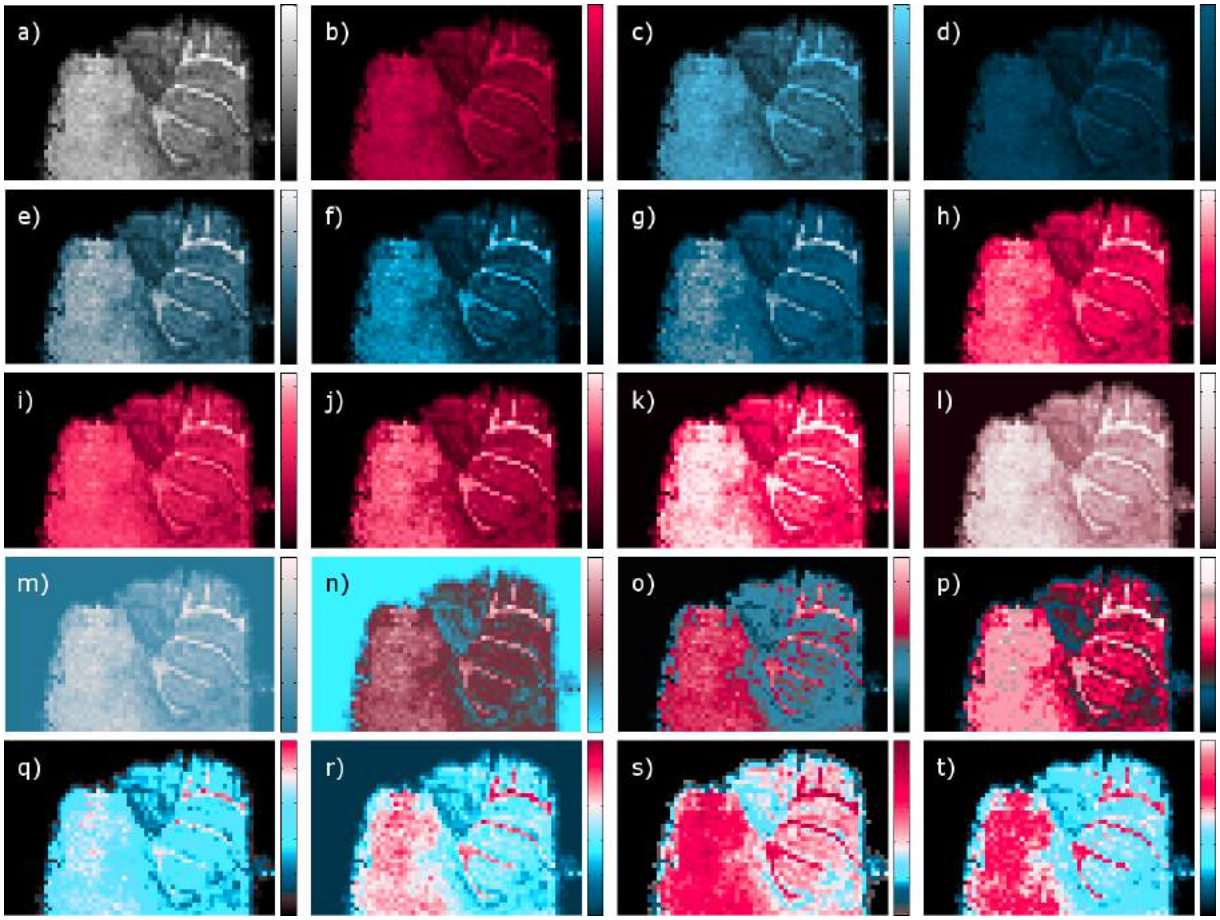


Figure 5.5: Tritanope simulation of Figure 5.2.

The frequency of occurrence of the rainbow based colour schemes, despite their well documented flaws, is likely due to the fact that these colour schemes are often used as default in instrument vendor's software (Figure 5.2t flexImaging (Bruker) and other rainbow, similar to Figure 5.2q, not shown ImageQuest (Thermo Scientific)), in vendor neutral software (Figure 5.2r MATLAB (MathWorks), MSiReader [51], omniSpect [53]) and in commercial software (Figure 5.2r SCiLS Lab (SCiLS GmbH)).

The next most frequently used set of colour schemes is the linear red, green and blue, shown in Figure 5.2b-d respectively, at least one of which appears in 31 of the reviewed articles. These are often presented as a precursor to display of an RGB composite image portraying the spatial distribution of three ions within a single image (discussed in Section 5.3.2). Although the same data is shown, the apparent contrast and brightness is much lower in the blue colour scheme than it is in the green. The distance between colours in the CIELAB space should correspond to perceived changes in colour, however the Euclidean distance was found to have flaws, namely poor performance in the blue region [269]. The CIEDE2000 distance metric was developed to address these issues and provide distances that correlated more closely to human perception. Using CIEDE2000, as implemented by [270], it is possible to determine the distance between the colours representing the minimum and maximum within a colour scheme, where the distance between black and white is 100 and the smallest perceivable difference is approximately 1 [271]. Assuming a colour scheme defined by a straight line in the CIE space, a smaller distance between minimum and maximum results in less perceivable colour changes throughout the colour scheme. Calculating this for the linear forms of Figure 5.2b,c and d) results in a ΔE_{00} of 51.34, 87.80 and 38.45 respectively. As the distance between the black and blue is significantly smaller than the distance between black and green, there are a reduced number of perceptually distinct colours within the blue colour scheme when compared with the green. So, although unintentional, more detail is apparent in whichever ion image is assigned to the green channel.

Calculating the distance between each subsequent colour in each colour scheme pro-

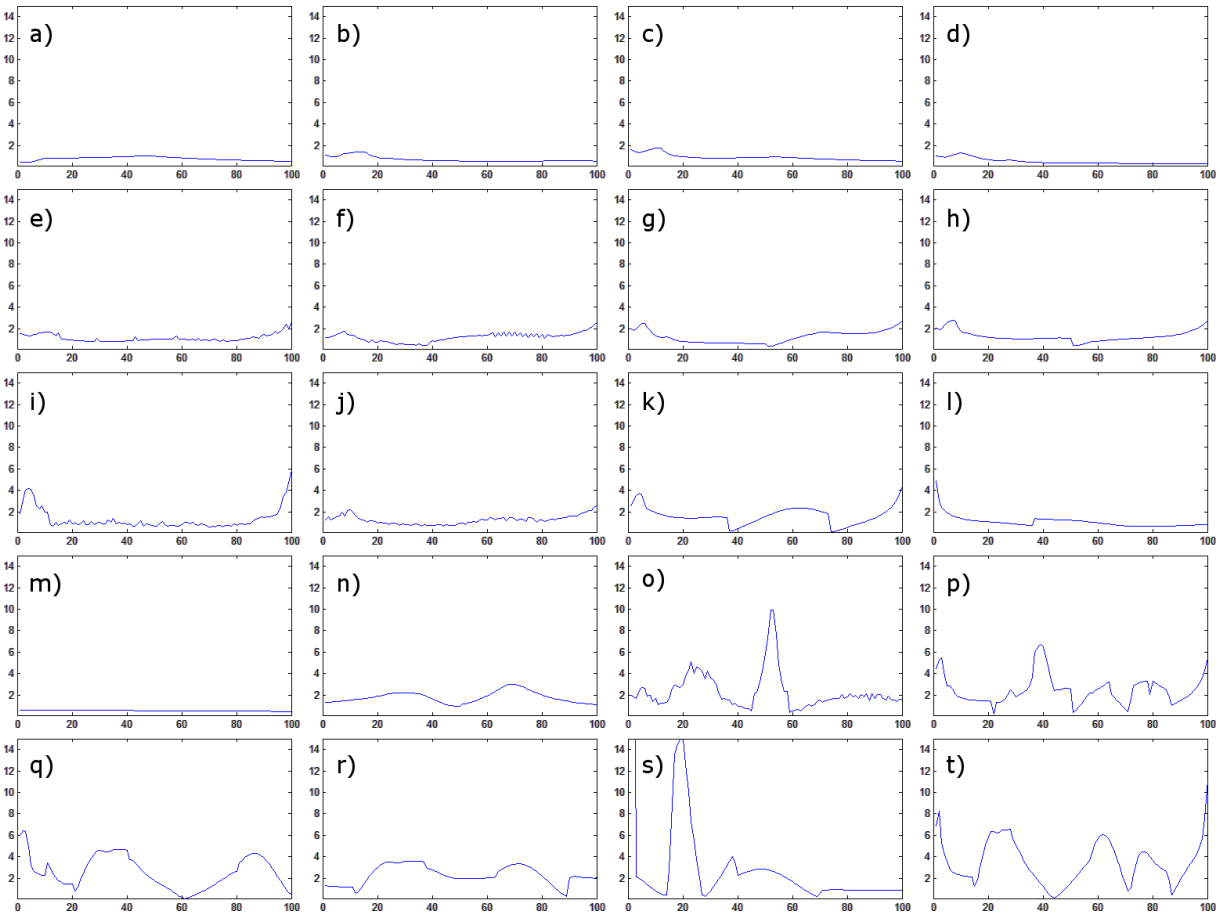


Figure 5.6: Distance between consecutive colours in each colour scheme shown in Figure 5.2 using the CIEDE2000 distance metric, as described in the main manuscript. Each x-axis denotes the index in the colour scheme. All y-axes denote the distance and are limited between 0 and 15, with two data points exceeding this range. The first two values for plot s) are 39.74 and 35.26 respectively.

vides an easy way to evaluate its perceptual linearity. Plots for each of the colour schemes presented in Figure 5.2 are provided in Figure 5.6. Perceptually linear colour schemes have constant distances across the entire range, with the colour scheme in Figure 5.2m being the closest to achieving this. The colour schemes shown in Figure 5.2o, 5.2p (which is the default colour scheme in HDImaging (Waters)) and the rainbow colour schemes (Figure 5.2q-t) show the least linearity.

The application of the colour scheme presented in Figure 5.2n gives the false illusion that there is a tear or part of the tissue missing because parts of the tissue have the same colour as the background.

The presentation of the data using the colour scheme in Figure 5.2o shows significant

contrast between the grey and white matter regions of the tissue due to the change in colour from green to orange. However, this is an artificial sharp boundary introduced at the 50% intensity mark, where values within the 50-55 intensity count range have 2-5x the contrast (according to ΔE_{00}, see the central peak in 5.6o) when compared with other same size ranges. In the CVD simulation (shown in Figure 5.3o) the previously significant contrast between white and grey matter is lost.

As the magnitude is important in the display and interpretation of ion images, equal increases in ion intensity should appear as steps of equal perceived magnitude in the visualisation. Monotonic increase of luminance and saturation have both individually been shown to result in perceived monotonic increase of magnitude [148]. Use of both is necessary to provide a suitable colour scheme, as luminance is important for conveying high spatial frequency information, while hue and saturation convey low spatial frequency structures [148].

By defining colour schemes in the CIELAB space it is much easier to ensure that changes in colour correspond to perceived changes in magnitude. A colour scheme which follows the rules above (monotonic luminance and changing hue) but is defined in the RGB colour space is shown in Figure 5.7b. As shown in Figure 5.7a, the lightness monotonically increases, however final quarter of the colour scheme covers just 5% of the lightness range, resulting in low contrast for values between 75% and 100%. It is difficult to differentiate regions when luminance is equal, even with large chromatic differences [272]. Linearising the lightness increases contrast in this range as shown in Figure 5.7. This improved contrast over small intensity changes is progressively important to distinguish different intensities as the range of values displayed increase in orders of magnitude.

5.3.2 Overlapping Data

It is occasionally beneficial in MSI to overlay data, such as an ion image and a photograph or a corresponding histological image to determine co-localisation of features between the two imaging modalities. This is is done by presenting the ion image as a transparent layer

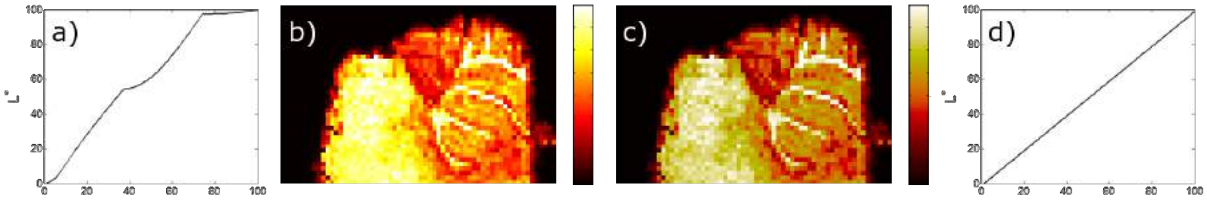


Figure 5.7: Same image displayed with a monotonically (but not linearly) increasing lightness colour scheme and the linearised lightness form. a) Lightness plot for the non-linear lightness colour scheme. b) Non-linear lightness colour scheme. c) Linear lightness colour scheme. d) Lightness plot for linear colour scheme.

over the optical image. However, there are perceptual pitfalls in this approach whereby the contents of the layers can interfere with one another, sometimes to the extent that it is impossible to determine which layer a given object belongs to [272]. For example, if a translucent green object is placed over a blue and a pink object (such as in the case of H&E stain images) the expectation is that the objects would maintain their colour, but with a green tint. This is the case for the blue objects, however the pink objects now appear orange (see Figure 5.8). In order to minimise interference, maximal separation in colour, texture, motion, and stereoscopic depth is required [272]. In the case of MSI the simplest solution is to ensure that the colour scheme(s) selected do not contain any duplicate colours. For example, overlaying an ion image with the rainbow colour scheme in Figure 5.2t onto a H&E stain image would contain duplicate colours, blue, pink, and white, making it difficult to determine which colour belongs to which layer.

Composite images are frequently employed in MSI publications (29 out of the 120 surveyed) to show distributions of multiple ions within the same image. A full list of all composite combinations and accompanying references is included in Table 5.2. Two colour composites are primarily (in 14 articles) displayed using the red and green channels. This has severe implications for any reader who suffers from either protanopia or deuteranopia as the two ions will be difficult, if not impossible, to separate. Safer options for displaying two colour composites are to use green-magenta or blue-yellow.

Display of three colour composites using RGB were found in 12 articles. Such images can be used to determine co-localisation of three ions of interest. This can be a powerful

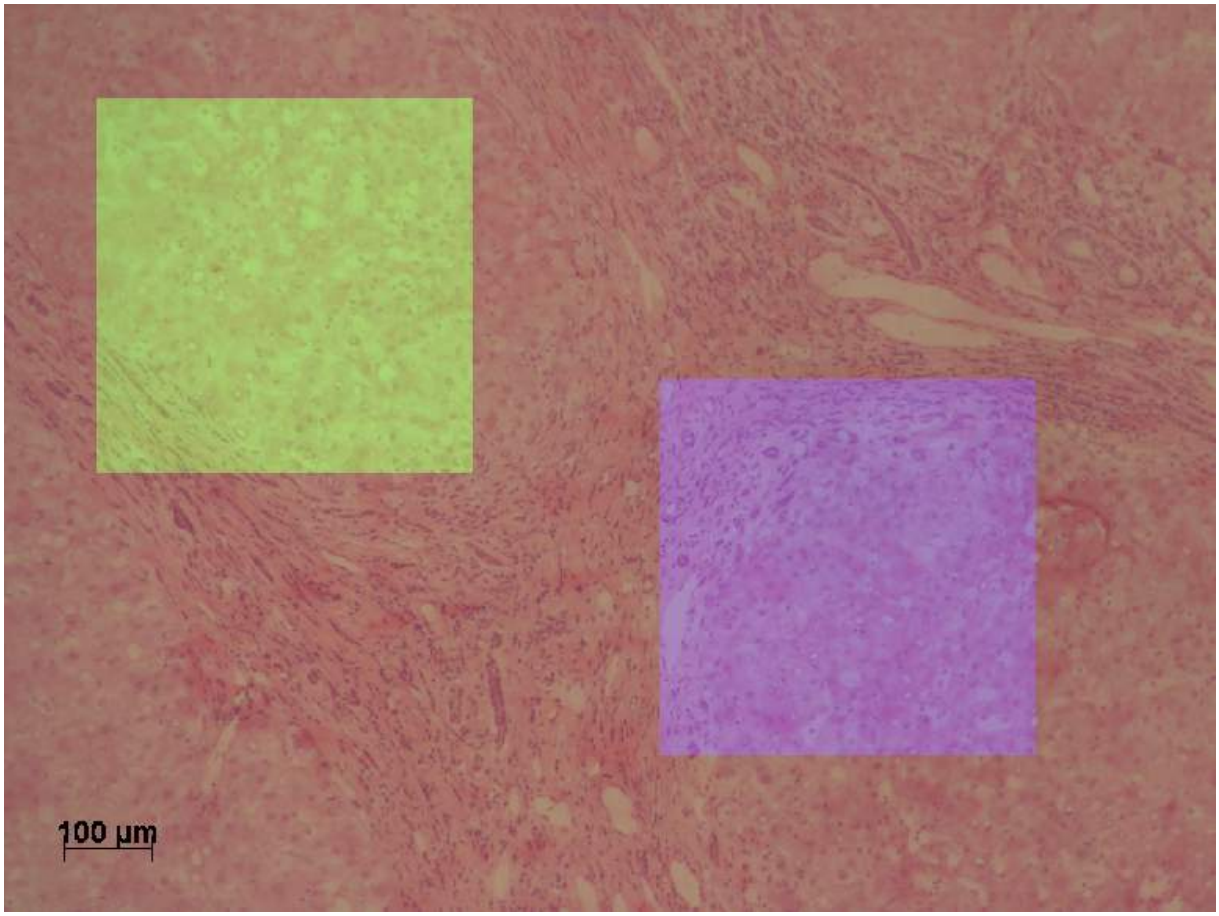


Figure 5.8: Hematoxylin and eosin stain with green and blue box overlays.

Composite	Reference	Count
Red-Green	[154][157][227][256][190][180][182][230][239][155] [167][192][179][181][259]	15
Red-Blue	[157][230][167][195][176]	5
Green-Blue	[230][239][176]	3
Red-Green-Blue	[203][157][228][19][257][7][230][239][167][193] [235][219]	12
Blue-Yellow	[195][176]	2
Other 2 colour composite	[259]	1
Other 3 colour composite	[207][200]	2
4 colour composite	[198][199]	2
8 colour composite	[7]	1
Total		43
Unique		29

Table 5.2: The different colours used to display composite images of two or more ion images.

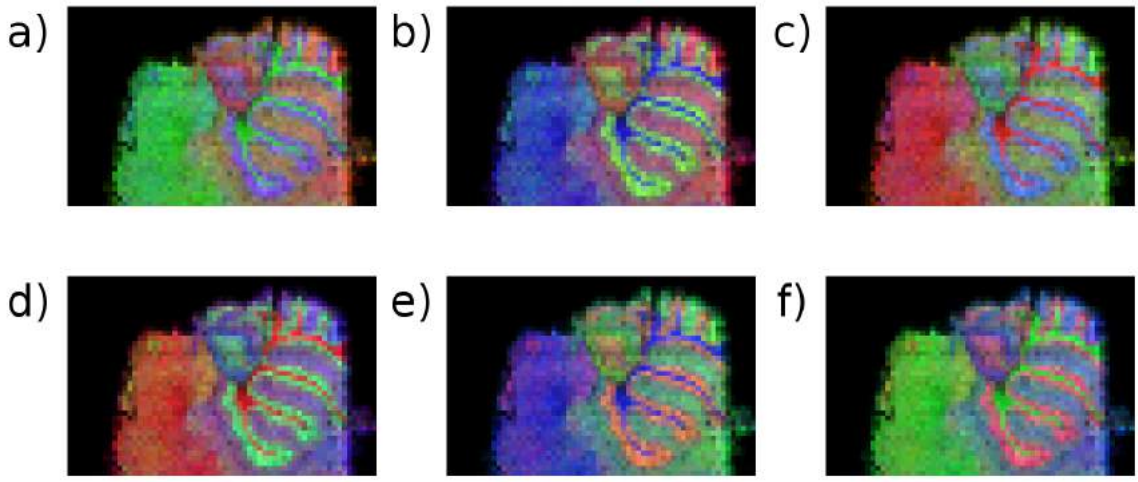


Figure 5.9: Composite RGB images of m/z 713, 826 and 844 where a) R is m/z 713, G is m/z 826 and B is m/z 844 b) R is m/z 713, G is m/z 844 and B is m/z 826 c) R is m/z 826, G is m/z 713 and B is m/z 844 d) R is m/z 826, G is m/z 844 and B is m/z 713 e) R is m/z 844, G is m/z 713 and B is m/z 844 f) R is m/z 844, G is m/z 826 and B is m/z 713.

tool for interpreting data, for example visualising the distribution of a drug compared to anatomy. The order in which the ion images are selected for the red (R), green (G) and blue (B) channels can impact the perceived importance of certain regions due to the brightness of the green channel compared with the red and blue, as is evident in the bright green regions in Figure 5.9. The chosen channel assignment also affects the perceived structure of the data for people with CVD as colours become indistinguishable, shown in Figure 5.10. Unfortunately, there is no simple solution that would result in a representation of the data that could be readily visualised without artefact by everyone. Care should be taken when representing data in this manner by confirming that the order in which the ion images are assigned to colour channels does not significantly alter the interpretation of the data. To ensure that the data is also interpretable to people with CVD, each ion image should be reproduced alongside the three colour composite.

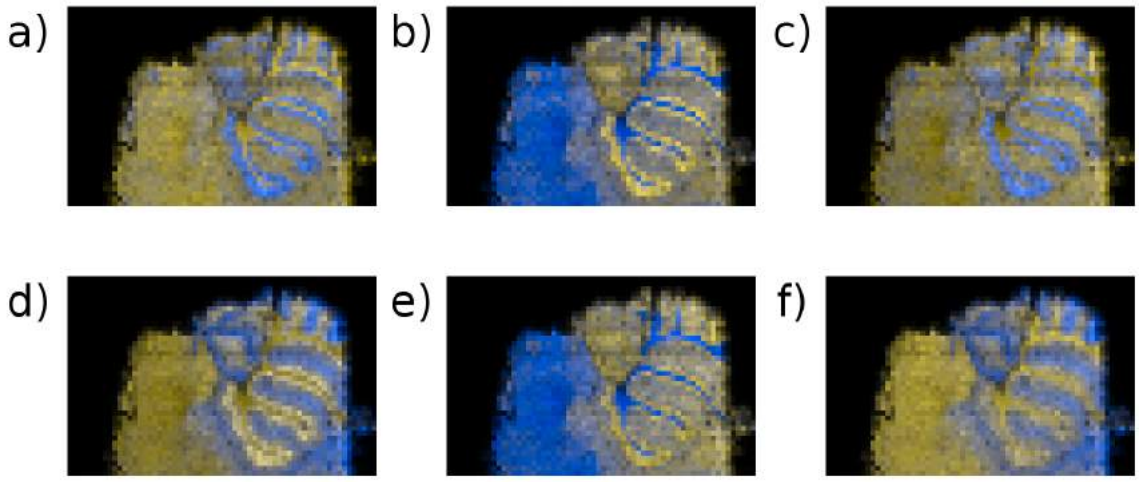


Figure 5.10: Deuteranope simulation of Figure 5.9.

5.3.3 Multivariate Analysis Images

In some cases the data to be displayed has both a positive and a negative component, such is the case in principal component analysis (PCA) score images. PCA is an unsupervised multivariate analysis technique which transforms a set of observations into a set of orthogonal variables, called principal components, as described in more detail in Section 3.5.

Once the principal components have been determined, the original data is projected into the new principal component space. These projected data are referred to as the scores. There are most commonly displayed by plotting one principal component against another to determine separation (or lack thereof) of different data categories. By retaining the original spatial position of each data point, it is possible to generate score images for individual principal components by assigning each pixel the score of the corresponding data point.

Only two of the reviewed articles contained PCA score images, one using the rainbow colour scheme from Figure 5.2r [180] and the other using the hot colour scheme shown in Figure 5.2k [222]. The drawback with using such schemes, apart from the reasons listed

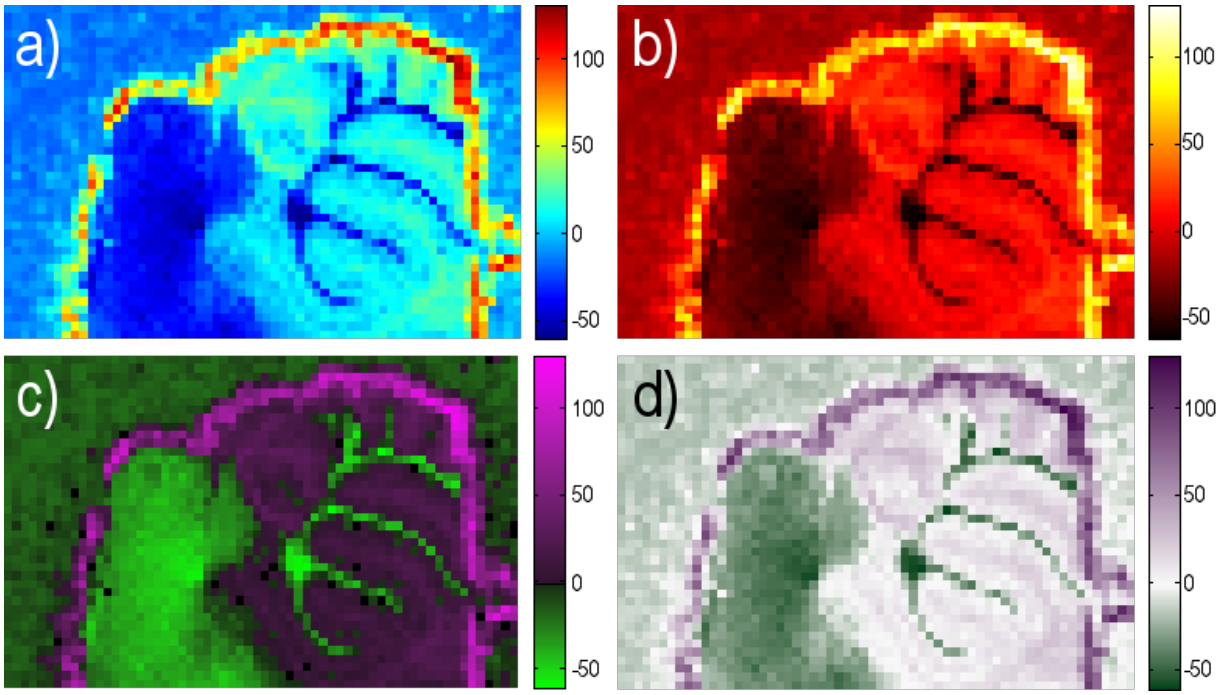


Figure 5.11: Principal component 2 score image displayed using a) rainbow colour scheme shown in Figure 5.2r, b) hot colour scheme shown in Figure 5.2k, c) diverging colour scheme with black at 0 and d) diverging colour scheme with white at 0.

previously, is that it becomes a colour matching exercise to determine which values are positive and which are negative. In such situations, the use of a diverging colour scheme centred around zero can provide a much more intuitive and informative representation of the data, as has been shown previously for SIMS data [273]. It is immediately evident which regions of the image have a positive value and which are negative in Figure 5.11c-d. These can either be centred around black as in Figure 5.11c, or around white as in Figure 5.11d (in which case it is useful to use the Msh colour space to avoid Mach band artefacts [147]).

This representation is also applicable for other multivariate analysis techniques that have both a positive and negative component (such as MAF). Other algorithms, such as NMF and pLSA, do not have a negative component and instead can be represented on a single colour scale in the same way ion images are.

5.3.4 Segmentation Images

Segmentation images are often employed as either region labels (based on histology) or clustering results. These are generally a form of nominal data as there is no order to the labels or clusters. Ideally objects should be distinguishably different, but without perceptual ordering [148]. Care should be taken when selecting colours for this task, as certain colour combinations can result in over or underestimating the comparative areas [274]. In general, more pastel colours result in more accurate perception of area coverage. A further consideration should be to ensure that the colours chosen are safe for people with CVD, otherwise multiple clusters or segments may be indistinguishable. This becomes increasingly difficult as the number of unique groups to display increases. A useful resource for selecting colours that take into account these effects is ColorBrewer [275].

5.4 Conclusions

The use of CVD simulation tools, such as Vischeck (http://www.vischeck.com/), is a simple and easy way to ensure that figures are suitable for a wider audience. Abandoning rainbow or rainbow based colour schemes and instead using perceptually linear colour schemes, which make use of as much of the lightness range as possible, for display of ion images allows intuitive interpretation of the underlying data. The use of the CIEDE2000 distance metric provides a simple way to determine the perceptual linearity of any given colour scheme.

In MS imaging the intensity indicates the relative measured concentration of a given analyte and this chapter has illustrated how apparent structure within an image differs according to the colour scale selected. For unambiguous understanding of analyte concentration additional experiments to achieve absolute quantitation should be performed alongside imaging experiments.

It is proposed that the community should strive to use perceptually linear colour

schemes, which denote the minimum and maximum values that the colour scheme spans, wherever possible when presenting MSI imaging data. In cases where a different colour scheme is used or where the colour scheme is adjusted for specific emphasis, the same data, but displayed with a perceptually linear scheme, should be included in the supplementary information to provide the reader with a better overview of the data as a whole and to ensure that the article is accessible to a wider audience, including people with CVD.

Instead of proposing a single colour scheme that should be used ubiquitously, this chapter presents a set of rules and metrics for designing and determining colour schemes that provide an accurate presentation of imaging data that takes into consideration human perception of colour. While the colour schemes that fit these rules may not be as visually appealing as some, they avoid a large number of artefacts which can result in misleading representations of data.

CHAPTER 6

APPLICATION TO THE STUDY OF TRAUMATIC BRAIN INJURY

6.1 Introduction

Traumatic brain injury (TBI) is a leading cause of disability and results in an estimated average life expectancy reduction of 4 years [276]. Hankin *et al.* have previously used MALDI MSI to investigate two rat brain injury models, ischemia/reperfusion and TBI [277]. In the TBI study an increase in ceramide (d18:0/18:1) ($[M+H\text{-}H_2O]^+$ m/z 548.5) and the sodium adduct of PC16:0/18:1 (m/z 782.6) as well as a decrease in the potassium adduct of PC16:0/18:1 (m/z 798.6) were observed at the site of injury.

In this chapter MALDI MSI data acquired from injured rat brains from two different mass spectrometers supplied by two different vendors is presented. Conversion and data analysis software presented in Chapter 2 is used to process the data and perform multivariate analysis. The memory efficient routines presented in Chapter 3 are required due to size of data produced when analysing four sections at once. The spatial and spectral hierarchy generation algorithms presented in Chapter 4 are employed to determine differences between the injured hemisphere and the control hemisphere of the injured brains.

6.2 Materials and Methods

A cortical stab injury (6 mm long x 4 mm deep) was inflicted on the left side of three rat brains 7 days prior to sacrifice. Full surgical procedure detail is provided in [278]. The sham control underwent the same surgical procedure, but with no cortical stab injury.

6.2.1 Sample Preparation

12 μm coronal sections were acquired at the site of injury and mounted onto an ITO coated glass slide for analysis on the ultrafleXtreme (Bruker Daltonics). 12 μm coronal sections were acquired at approximately the same anatomical region (midpoint of the brain) and mounted onto a glass slide for analysis on the Synapt G2S (Waters). Slides were each coated with 15 mL of 20 mg mL^{-1} CHCA (80% MeOH with 0.1% TFA), applied with an artist airbrush propelled by dry N_2.

6.2.2 Mass Spectrometry Imaging

Data were acquired using an ultrafleXtreme (Bruker Daltonics) with 100 μm pixels with a mass range of m/z 50-1000 (raw data size 15.7 GB with 48160 pixels) and a Synapt G2S (Waters) with 100 μm pixels with a mass range of m/z 50-1200 (raw data size 25.6 GB with 68471 pixels).

6.2.3 Data Processing

Data acquired using an ultrafleXtreme (Bruker Daltonics) were converted to mzML using CompassXport (Bruker Daltonics). mzML files were then converted to imzML using imzMLConverter as described in Section 2.2 (converted data size 26.5 GB). The total spectrum was generated using the memory efficient method presented in Chapter 3 where the preprocessing workflow method included TIC normalisation. Peak detection was performed using the gradient based method described in Section 1.6.5.

Data acquired using an Synapt G2S (Waters) were converted to mzML using msconvert as part of ProteoWizard [40]. mzML files were then converted to imzML using imzML-Converter as described in Section 2.2 (converted data size 51.5 GB). The total spectrum was generated using the memory efficient method presented in Chapter 3, where the preprocessing workflow consisted of 'set union of all m/z bins' (Algorithm 2.5), 'ensure constant number of bins per peak' (Section 2.5.1, with 30 bins per peak), Savitzky-Golay smoothing (window size = 15, polynomial order = 2) repeated twice and 'remove negatives' baseline correction. Peak detection was performed using the gradient based method described in Section 1.6.5.

Memory efficient PCA and datacube reduction were performed using the Algorithms 3.6 and 3.5 presented in Chapter 3 and the preprocessing workflows described above. NMF and PLSA were performed on the datacube in memory with $k = 25$.

6.2.4 Hierarchy generation

Spatial hierarchy of the data acquired using the ultrafleXtreme (Bruker Daltonics) (generated as described above) was generated using Algorithm 4.1. The spectral hierarchy was generated from the first brain in the ultrafleXtreme (Bruker Daltonics) datacube using Algorithm 4.2. Generated spectral hierarchies were applied to the remaining data acquired using the ultrafleXtreme (Bruker Daltonics) as well as the data acquired using the Synapt G2S (Waters) (after being reduced to a datacube, described above) as described in Section 4.3.1.

6.3 Results and Discussion

MSI data from four rat brains, three of which were inflicted with a cortical stab injury 7 days prior to sacrifice, were acquired using an ultrafleXtreme (Bruker). Acquisition was terminated approximately half way into the imaging experiment due to instrument overheating. The acquisition was resumed approximately 72 hours later without the

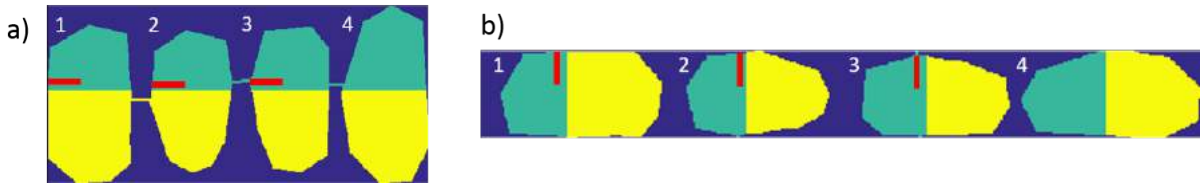


Figure 6.1: Diagram showing the split in acquisition for the ultrafleXtreme (Bruker Daltonics) dataset. The green region denotes the area surveyed in the first acquisition. The yellow region shows the area analysed 72 hours after the first acquisition. The red marker shows the site of injury. a) How the data were acquired. b) The orientation of the data presented for this chapter.

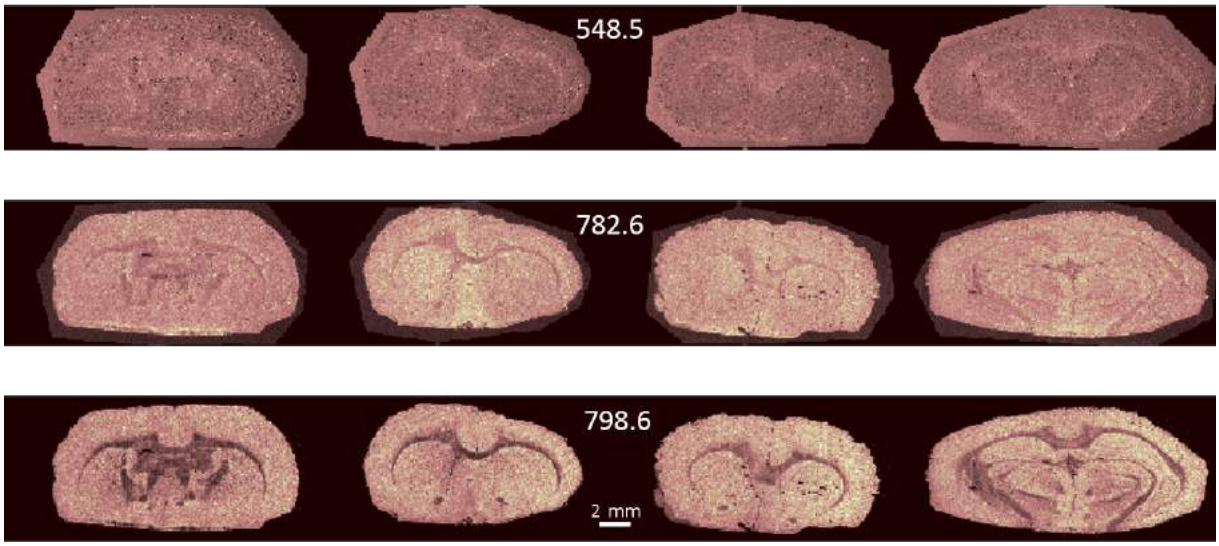


Figure 6.2: Ion images of m/z 548.5, 782.6 and 798.6 generated from the ultrafleXtreme (Bruker Daltonics) dataset. Scale bar 2 mm.

sample being removed from vacuum. The orientation of the brain sections as they were acquired (including the separation between the two acquisitions) and as they are presented in this chapter are presented in Figure 6.1.

Ion images of the m/z values of interest from the study by Hankin *et al.* [277] are shown in Figure 6.2. No observable increase or decrease in any of these ion images is present at the site of injury within this dataset.

A comparison of normalisation techniques applied to m/z 820.6 is shown in Figure 6.3. Median normalisation resulted in a number of pixels appearing as zero or low intensity (black) on the tissue edges caused by divide by 0, as discussed in Section 2.5.4. By instead calculating the median value of non-zero intensities in a given spectrum, the normalisa-

tion artefact is removed. A similar effect happens when normalising to the noise level, where the normalisation coefficient is calculated as the median absolute deviation of the difference between adjacent intensities. Removing zero intensity values from consideration again removes the normalisation artefact. Out of all normalisation methods applied, only TIC normalisation revealed the location of the inflicted injury.

An RGB composite (as described in Section 5.3.2) of m/z 820, 772 and 826 is shown in Figure 6.4. This provides a clearer view of the site of injury than a single ion image alone. However, as discussed in Section 5.3.2, the order in which the ion images are assigned to the red, green and blue channel affects the prominence of the injury. These ions were selected by the laborious and time consuming process of manual inspection of peaks that fell within the lipid region. Finding three ions, which when displayed as an RGB composite, appear to show the location of the injury was largely due to luck.

To provide more robust and unbiased interpretation of the data, the unsupervised methods included within SpectralAnalysis were employed to investigate whether the site of injury could be discerned and whether there was an observable molecular difference between the two hemispheres of the injured brains. The first five principal components are presented in Figure 6.5. As is common in MSI, the first principal component (and therefore the source of largest variance within the data) described the difference between the sample and substrate. The third principal component described a difference in the matrix region between the first and second acquisition. This indicates that different ions were detected at the different time points. The score image and the corresponding coefficients for PC 3 is shown in the orientation in which the data were acquired in Figure 6.6. From the coefficients it is possible to isolate ions which contribute significantly to the different acquisitions, with m/z 212 ($[M+Na]^+$), m/z 61 and m/z 401 ($[2M+Na]^+$) being predominant in the first acquisition and m/z 269 and m/z 165 in the second acquisition. Manual inspection of the ion images of all identified m/z values revealed that these were isolated to the matrix region, indicating that they are matrix related ions.

Selected NMF factors and PLSA latent variables which show the differentiation of the

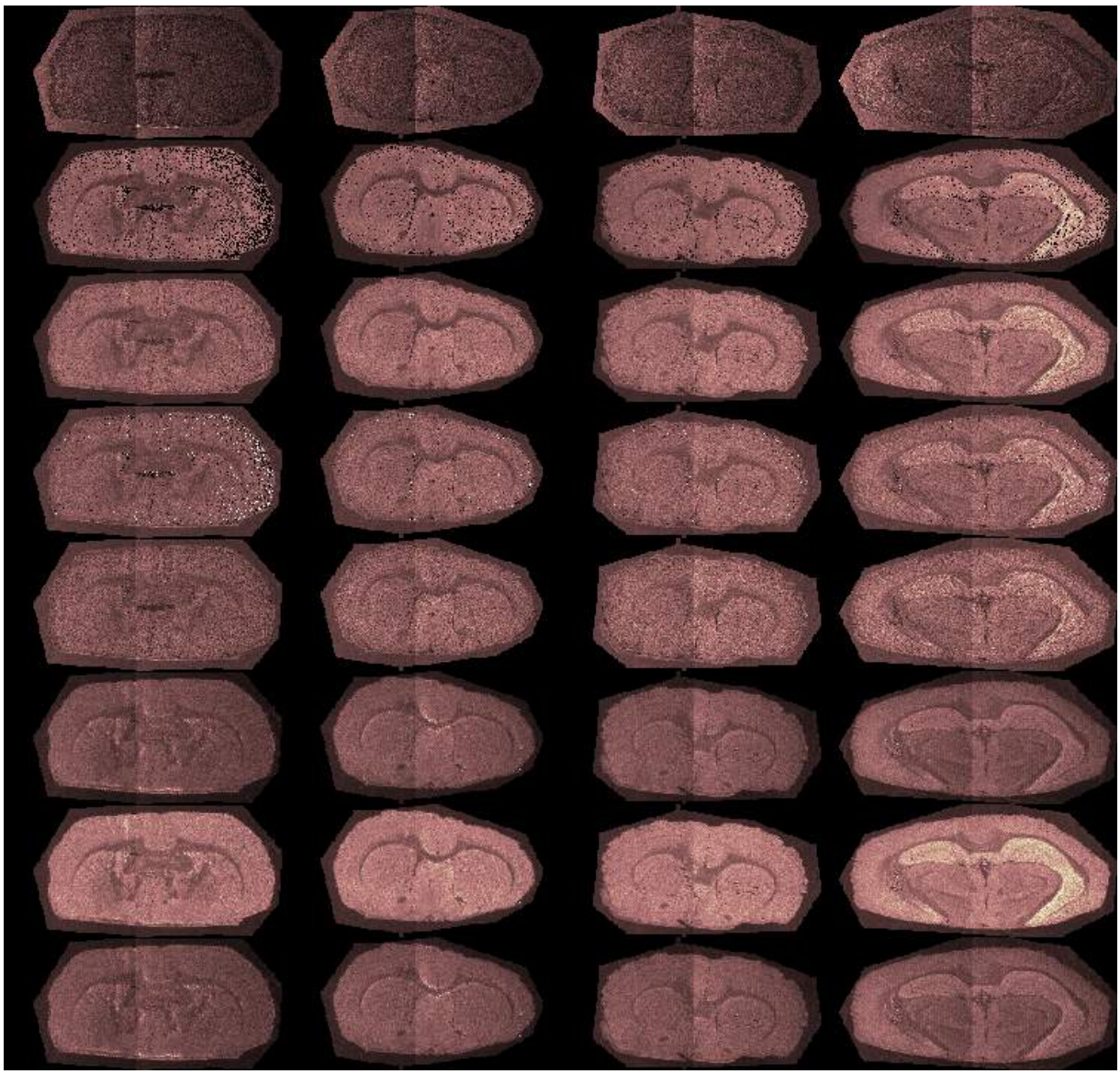


Figure 6.3: Comparison of normalisation techniques applied to m/z 820.6 from the ultrafleXtreme (Bruker Daltonics) dataset. From top to bottom: raw data, median, median with zero values excluded, noise level, noise level with zero values excluded, RMS, TIC and ℓ_2.

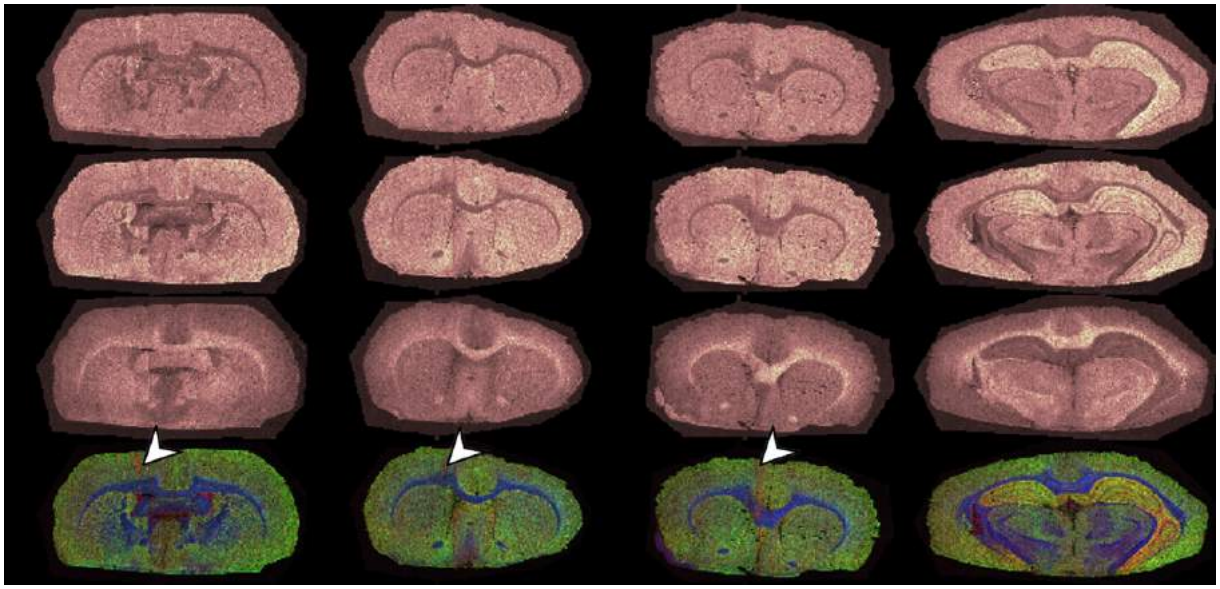


Figure 6.4: Co-localisation of m/z 820, 772 and 826 from the ultrafleXtreme (Bruker Daltonics) dataset as an RGB composite highlighting the site of injury.

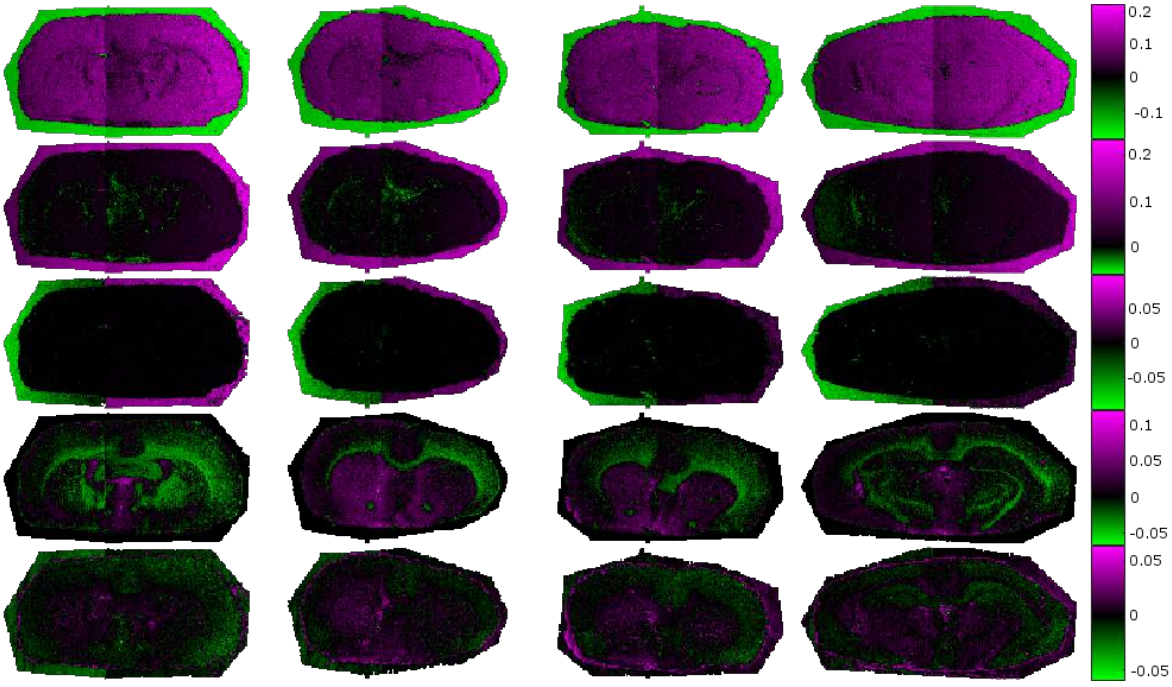


Figure 6.5: First five principal components from the Bruker dataset.

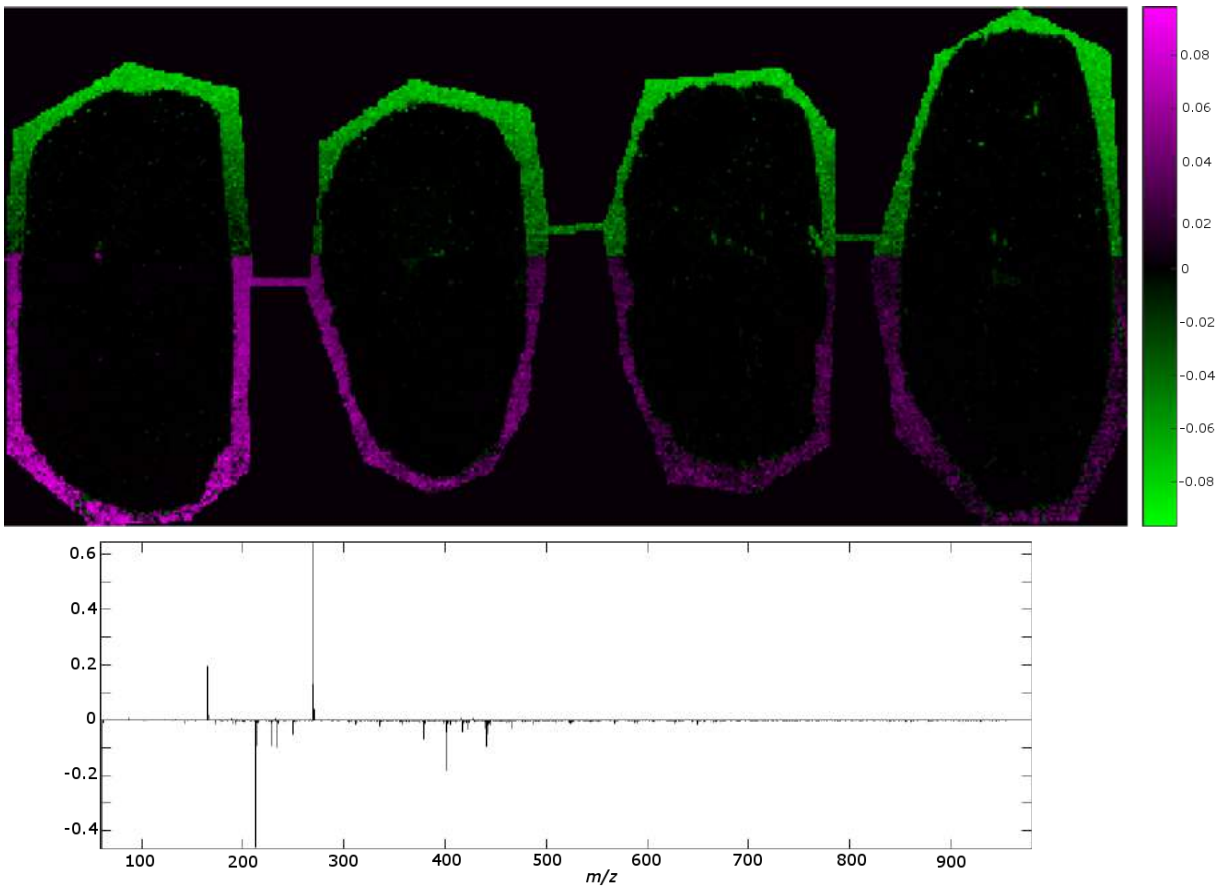


Figure 6.6: Third principal component from the Bruker dataset.

matrix region are presented in Figures 6.7 and 6.8 respectively. Predominant peaks in the NMF factor describing the first acquisition are m/z 61, 401 and 212. The same peaks, with the addition of m/z 165 were dominant in the PLSA latent variable describing the first acquisition. For the second acquisition dominant peaks in the NMF factors were m/z 269 and 165 and in PLSA were m/z 165, 269 and 212. All three multivariate analysis techniques highlighted similar peaks as being significant contributors to the difference in detected matrix ions between the two acquisitions.

Despite there being a discernible spectral difference in the ions detected in the matrix regions between the two acquisitions, no such difference was present in any PCA coefficients, NMF factors or PLSA latent variables describing on tissue features. Factors and latent variables associated with on tissue spatial distributions are presented within Figures 6.9 and 6.10 respectively. Both methods isolate distributions that describe grey and white matter regions of the brain without any obvious differences in the left and right hemispheres.

As SpectralAnalysis is suitable for the analysis of data from any instrument, it was also used for the visualisation, processing and comparison of a second MS imaging experiment acquired using a Synapt G2S (Waters). In this experiment the brains were sectioned at approximately the same anatomical location (the midpoint of the brain). The pixel size and matrix choice was constant between the two imaging experiments. A comparison of the spectral quality of each dataset is given in Figure 6.11. The higher mass resolving power of the Synapt G2S resulted in detection of almost ten times more resolvable peaks (18633 compared to 1935).

Factors (from NMF) and latent variables (from PLSA) associated with on tissue spatial distributions in the data acquired using the Synapt G2S (Waters) are presented within Figures 6.12 and 6.13 respectively. In comparison to the data acquired using the ultrafleXtreme (Bruker Daltonics), a wider variety of spatial distributions are present. In the results of PLSA, latent variable 1 describes the ventricles, 2 describes the thalamus, 3 and 4 the corpus callosum, 5 and 6 the cortex and amygdala and 7 the internal structure in

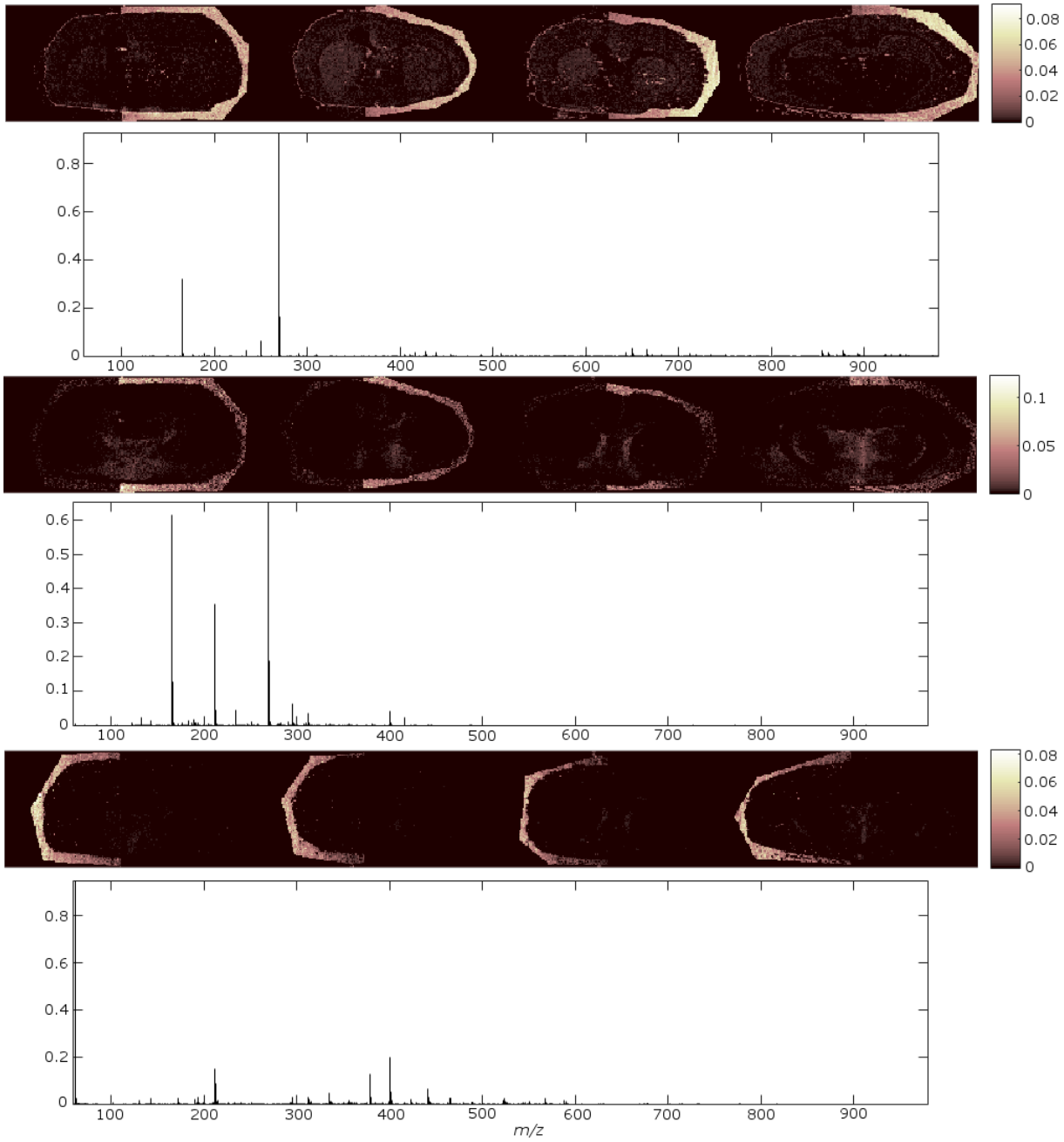


Figure 6.7: Selected NMF factors from the ultrafleXtreme (Bruker Daltonics) dataset showing the differentiation in the matrix region between the two acquisitions.

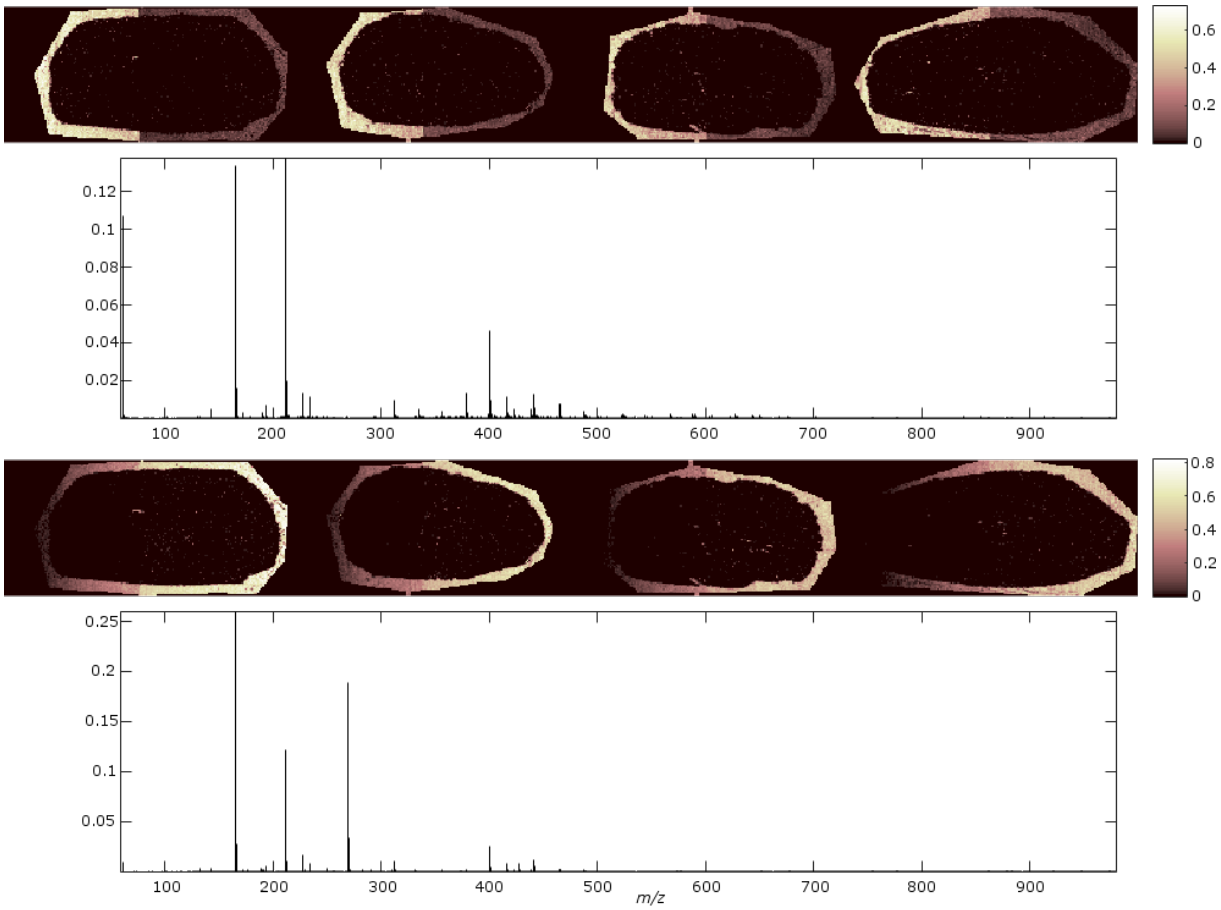


Figure 6.8: Selected PLSA latent variables from the ultrafleXtreme (Bruker Daltonics) dataset showing the differentiation in the matrix region between the two acquisitions.

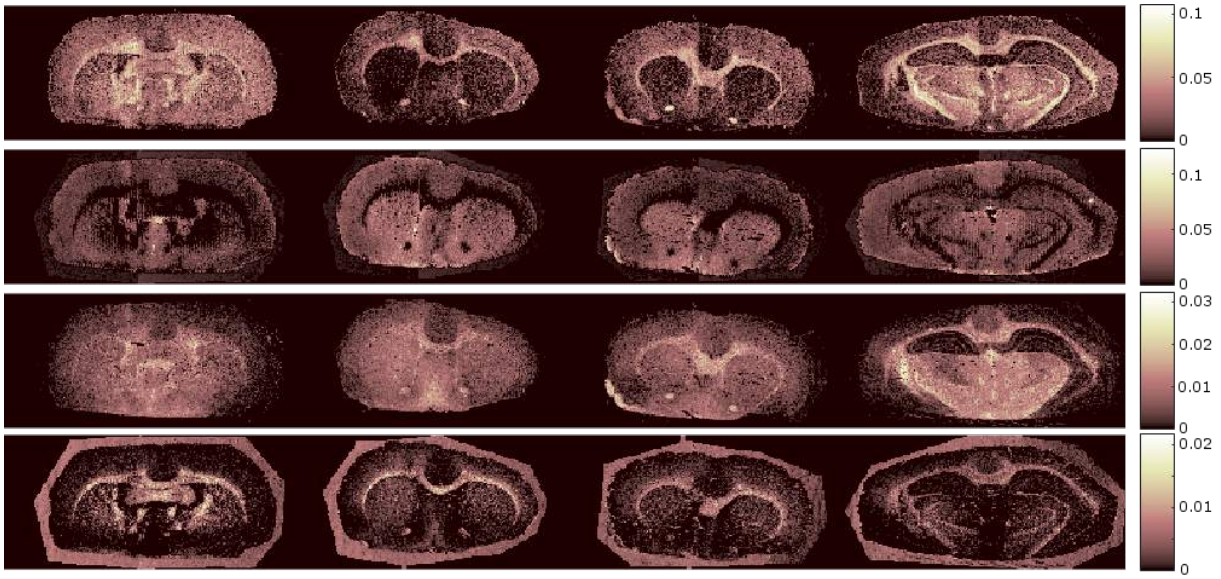


Figure 6.9: Selected NMF factors showing the differentiation in the grey and white matter of the brain in the ultrafleXtreme (Bruker Daltonics) dataset.

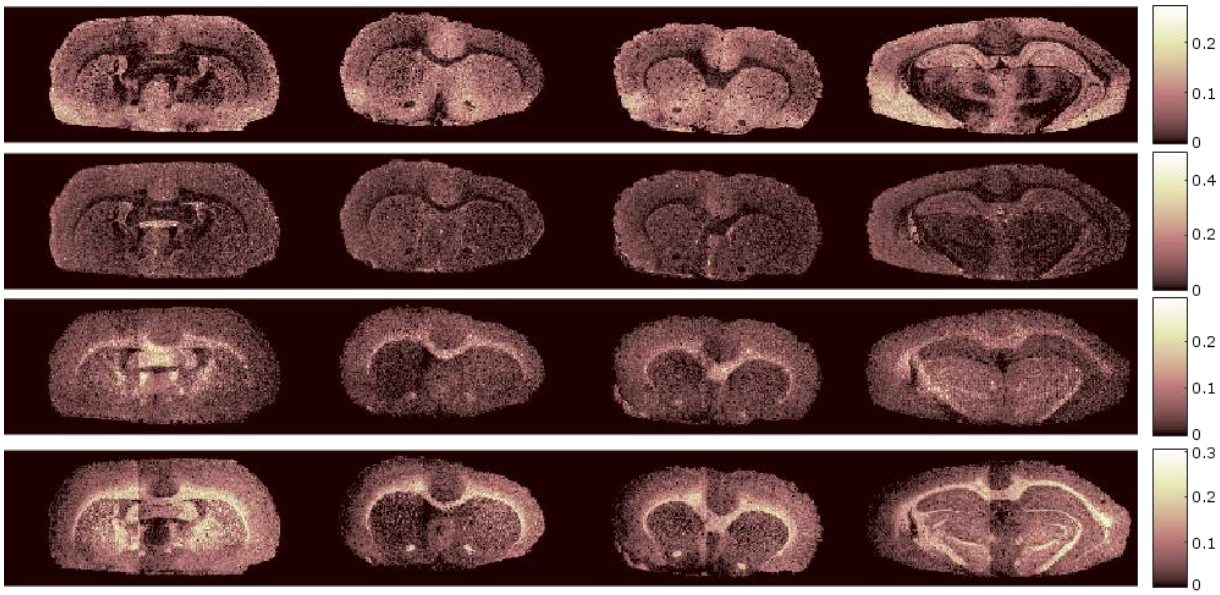


Figure 6.10: Selected PLSA latent variables showing the differentiation in the grey and white matter of the brain in the ultrafleXtreme (Bruker Daltonics) dataset.

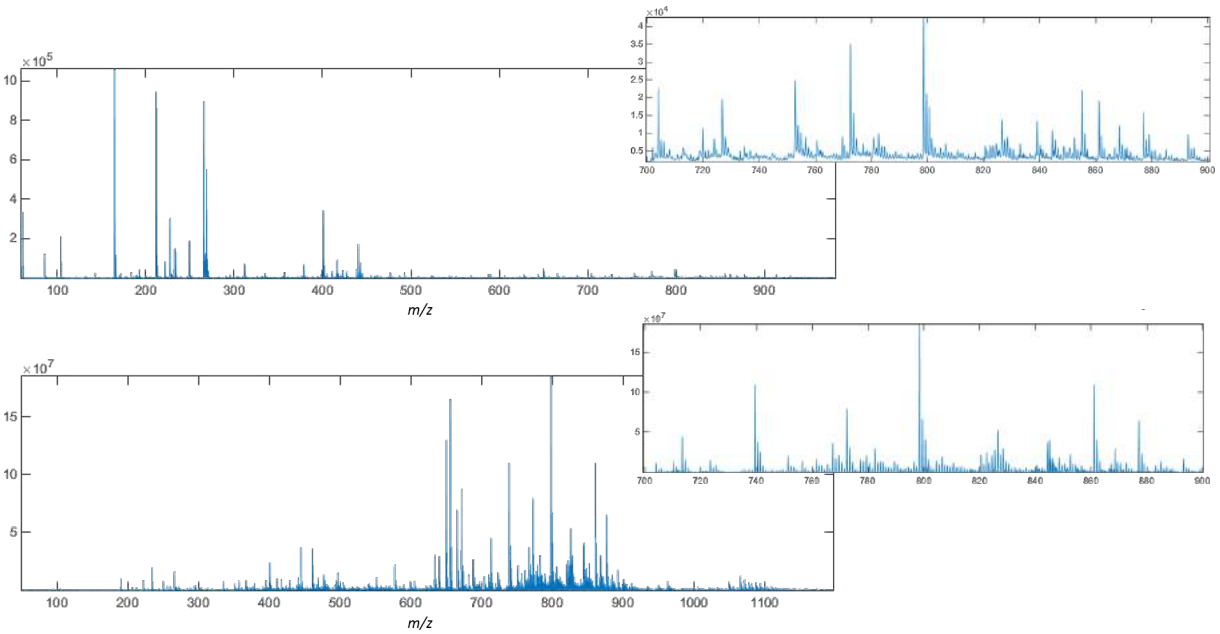


Figure 6.11: Comparison of the spectral quality of data acquired using an ultrafleXtreme (Bruker) shown top and a Synapt G2S (Waters) shown bottom.

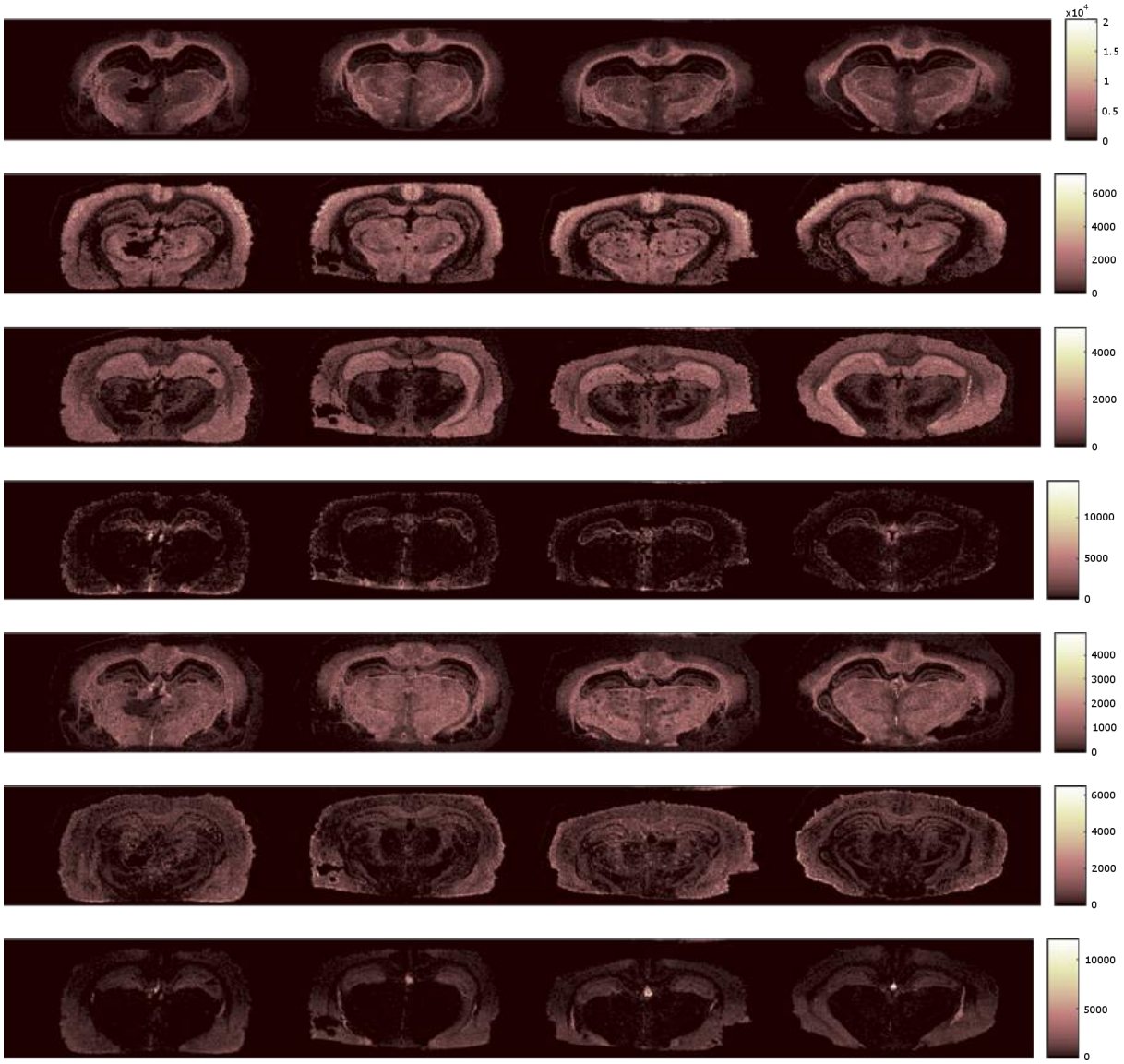


Figure 6.12: Selected NMF factors showing different spatial distributions identified from the data acquired using the Synapt G2S (Waters).

the hippocampus, as identified using Figure 6.14. This increased differentiation of internal structure can be attributed to both the location at which the sections were acquired and the increased number of resolvable peaks included in the multivariate analysis. As in the data acquired using the ultrafleXtreme (Bruker Daltonics), there was no observable difference between the injury hemisphere and the control hemisphere revealed by any of the applied techniques.

Application of the spatial hierarchy (described in Section 4.2) to the ultrafleXtreme (Bruker Daltonics) dataset generated 25 level 1 groups, the mean images of which are

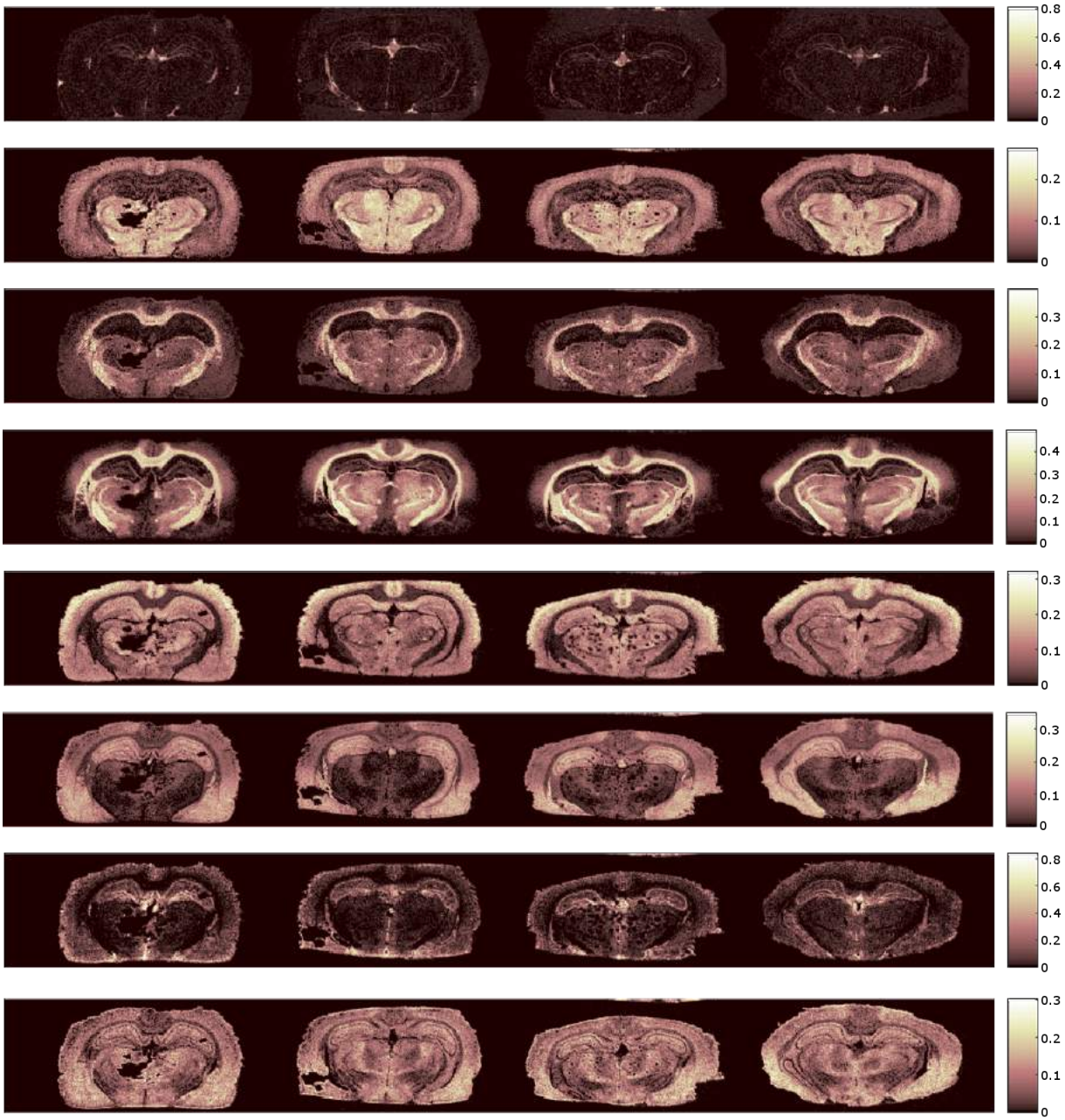


Figure 6.13: Selected PLSA latent variables showing different spatial distributions identified from the data acquired using the Synapt G2S (Waters).

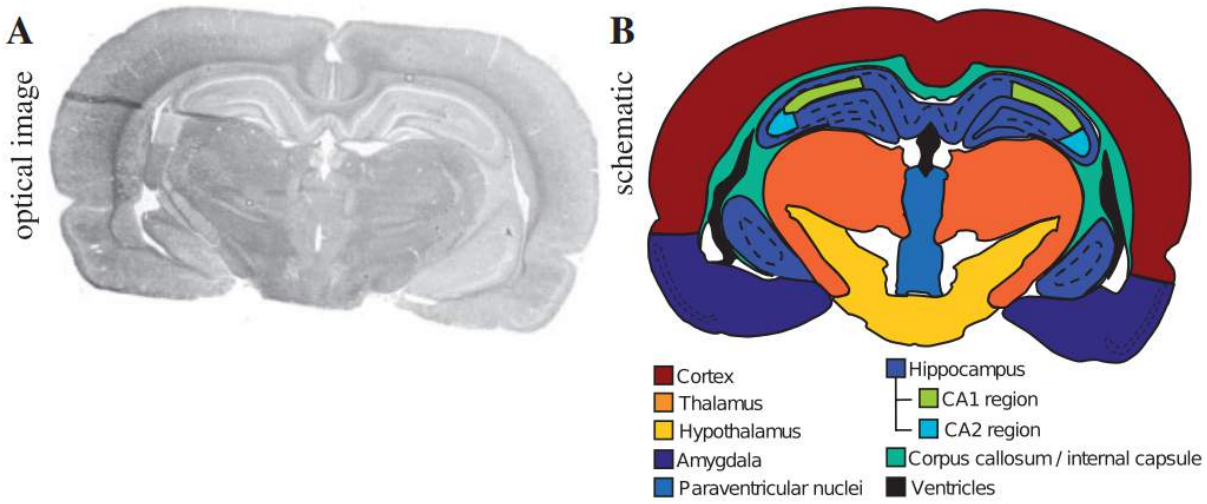


Figure 6.14: Labelled coronal section of rat brain. Modified from [106].

shown in Figure 6.15. The mean ion image for group 15 appeared to show an intensity difference between regions of the left and right hemisphere of the injured brains. Further investigation of this group (shown in Figure 6.16) revealed ions of interest which were detected at a higher intensity in the injured hemisphere, close to the site of injury, in the three injured brains.

Spectral hierarchies were generated using Algorithm 4.2 from the first brain in the data acquired using the ultrafleXtreme (Bruker Daltonics) dataset and subsequently applied to the remaining three brains. An example hierarchy including the ion image m/z 826.6 is given in Figure 6.17. Each of the ions in the hierarchy (m/z 826.6, 827.6 and 866.7) appear to localise primarily in the corpus callosum and the resulting correlation values imply high correlation between the ion images in each image segment. Application of this hierarchy to the data acquired using the Synapt G2S (Waters) dataset as a whole is shown Figure 6.18 where the correlation between m/z 826.6 and 827.6 is 0.9907 but the correlation between all members of the hierarchy is 0.3961. Visual inspection of the ion images shows a similar spatial distribution between all members of the hierarchy, which is comparable to that of the ultrafleXtreme (Bruker Daltonics) dataset. However, the low correlation value for the hierarchy as a whole indicates that this is not the case. This low correlation is likely caused due to the matrix region visible in m/z 866.7, possibly caused

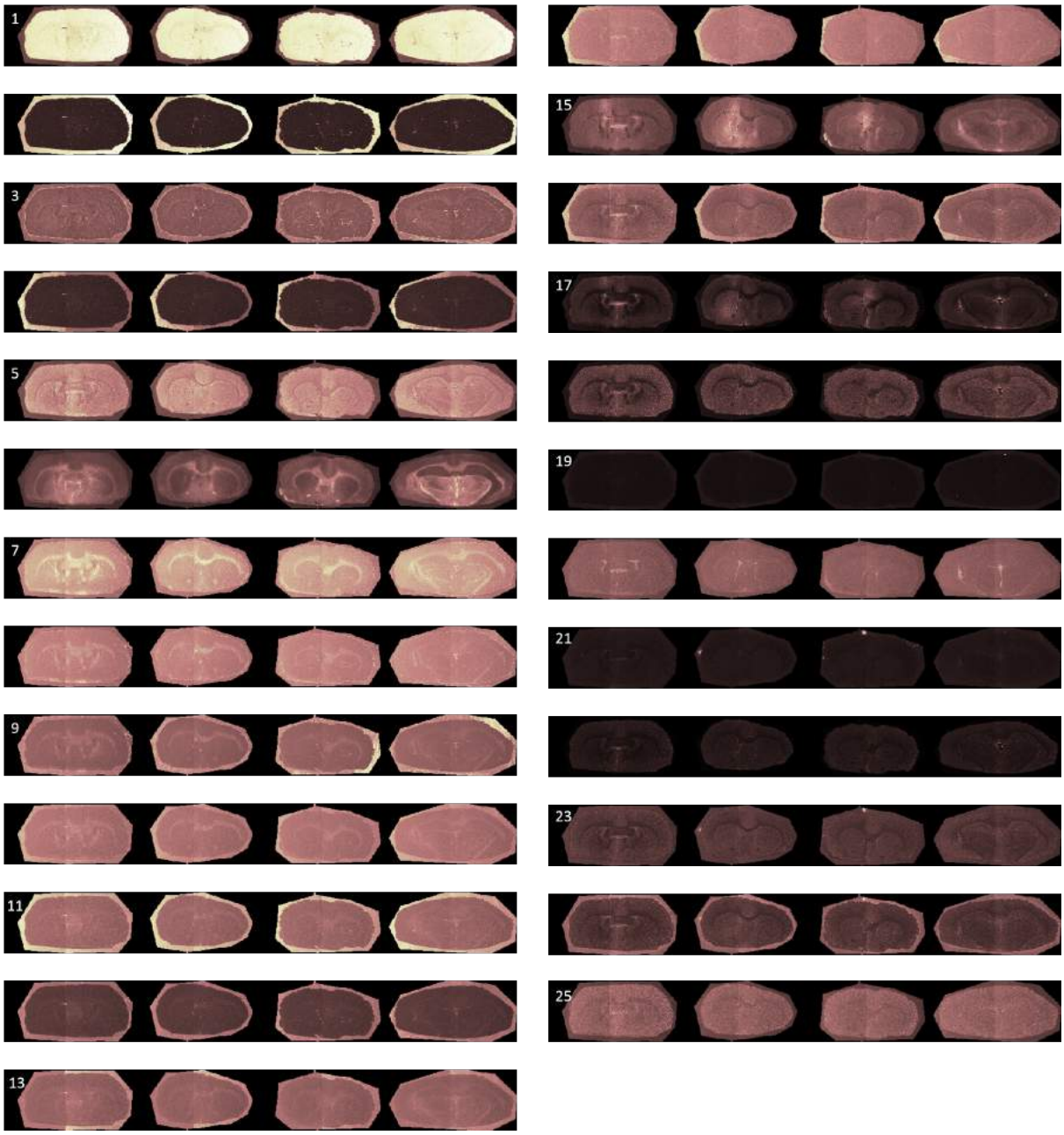


Figure 6.15: Level 1 hierarchy calculated from the data acquired from the ultrafleXtreme (Bruker Daltonics) using the spatial hierarchy, Algorithm 4.1.

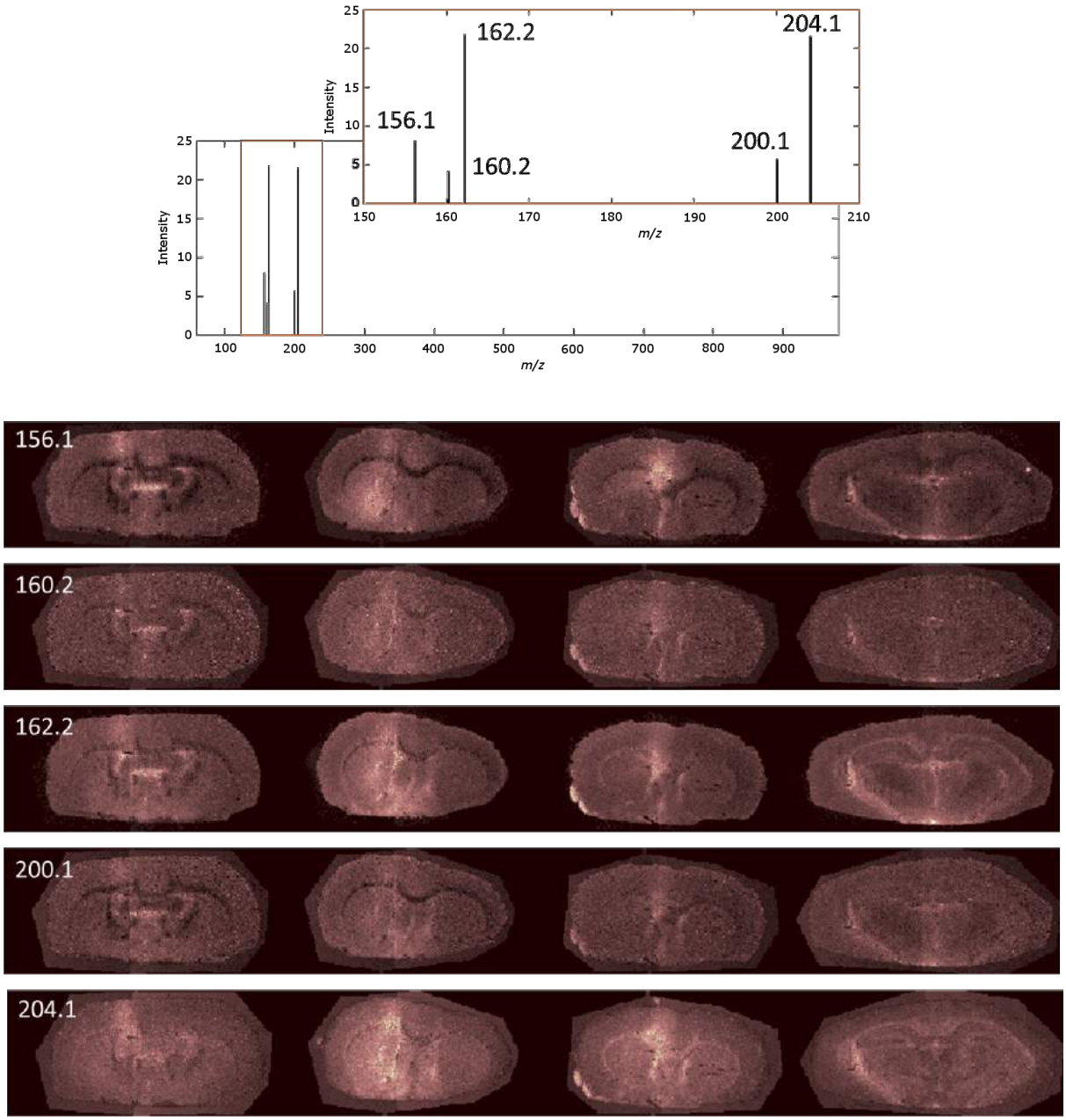


Figure 6.16: Group 15 of the level 1 hierarchy calculated from the data acquired from the ultrafleXtreme (Bruker Daltonics) using the spatial hierarchy, Algorithm 4.1.

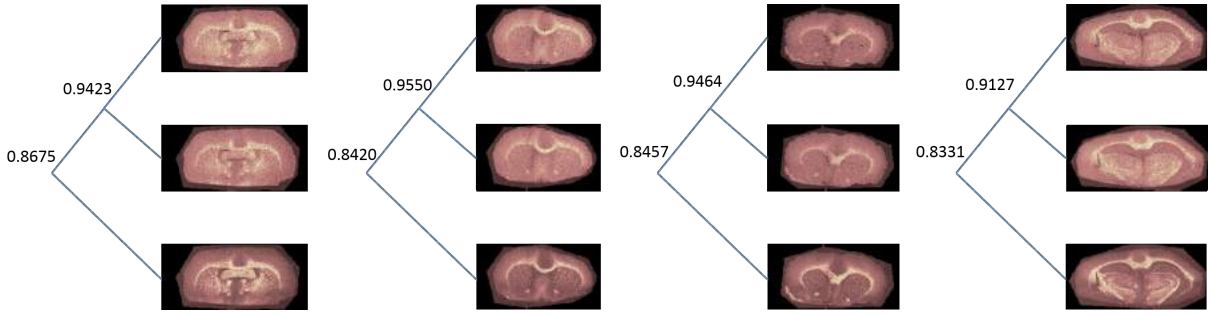


Figure 6.17: Hierarchy containing m/z 826.6 generated from spectral hierarchy, Algorithm 4.2, of the injury control (left) and subsequently applied to the other brain images in the ultrafleXtreme (Bruker Daltonics) dataset.

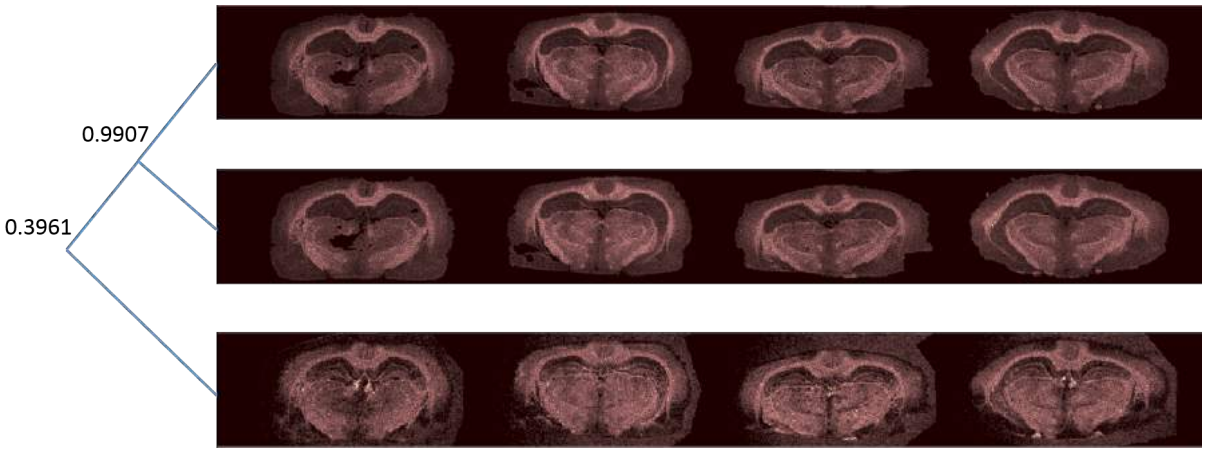


Figure 6.18: Hierarchy from Figure 6.17 applied to the data acquired using the Synapt G2S (Waters).

by an unresolved matrix cluster.

A hierarchy which included m/z 204.1 (identified previously to have increased intensity in the injured hemisphere) is shown in Figure 6.19. The hierarchy consisted of the ion images of m/z 204.1, 162.1 and 258.0. Application of this hierarchy to the data acquired using the Synapt G2S (Waters) is shown in Figure 6.20, with a correlation value of 0.0192 indicating that the member ion images do not have similar distributions. Visual inspection of these show that the ion images of m/z 204.1 and 162.1 are predominantly noise and no obvious structure is present.

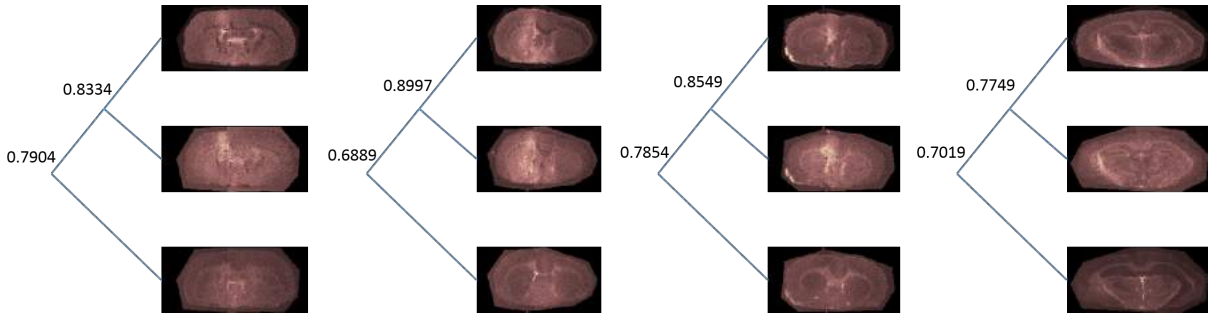


Figure 6.19: Hierarchy containing m/z 204.1 generated from spectral hierarchy, Algorithm 4.2, of the injury control (left) and subsequently applied to the other brain images in the ultrafleXtreme (Bruker Daltonics) dataset.

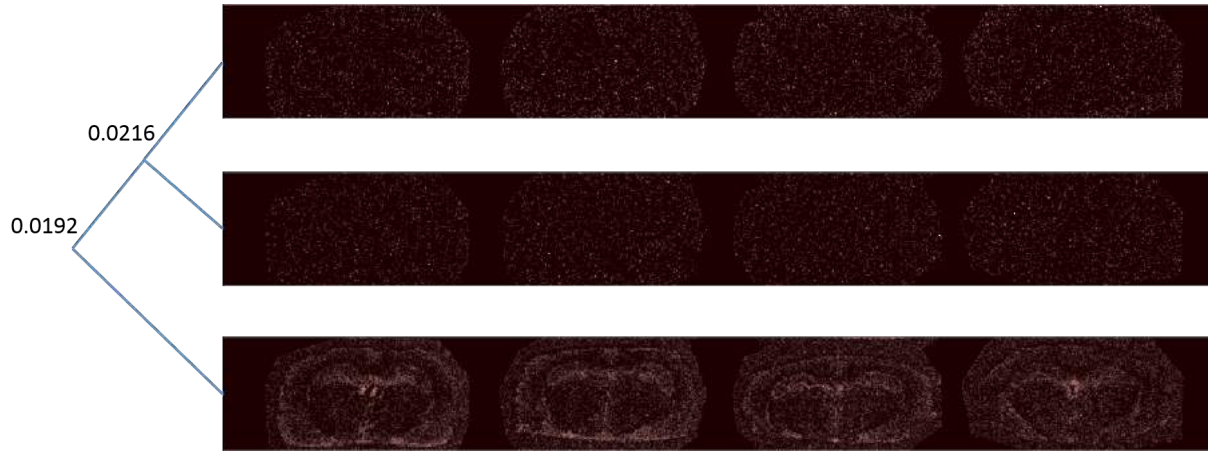


Figure 6.20: Hierarchy from Figure 6.19 applied to the data acquired using the Synapt G2S (Waters).

6.4 Conclusions

SpectralAnalysis has successfully been used to both preprocess and perform multivariate analysis on data acquired on two different instruments supplied by different mass spectrometer vendors. The use of memory efficient PCA, NMF and PLSA identified a difference in the detected matrix ions between the two partial acquisitions of the same sample, started approximately 72 hours apart. These methods did not identify the site of injury or any difference between the two brain hemispheres. Ion images which show elevated intensities in the injured hemisphere of the brain were identified through the use of the spatial hierarchy algorithm presented in Chapter 4.

Application of a spectral hierarchy generated from data acquired on an ultrafleXtreme (Bruker Daltonics) was applied to data acquired using a Synapt G2S, revealing ions with comparable spatial distributions. The ions with elevated intensity in the injured hemisphere in the data from the ultrafleXtreme (Bruker Daltonics) were not detected in the data acquired using the Synapt G2S (Waters), with the application of the spectral hierarchy showing no correlation. Further work is required to identify these ions as well as repeat experiments acquiring more data at the site of injury to confirm that the elevated intensity is not an artefact of the original imaging experiment.

CHAPTER 7

CONCLUSIONS AND FURTHER WORK

The research presented in this thesis has described new software for converting, processing and visualising MSI data, the effects of preprocessing, methods for handling extremely large datasets in a memory efficient manner, a novel method for comparing data and learning hierarchies of parts and optimal ways of displaying the data generated by each of these methods.

In Chapter 2 software for the conversion to the open imaging format imzML was presented. As the size of the metadata portion of the file is proportional to the number of spectra, this file format will become less suitable for datasets containing a large number of pixels, such as large area SIMS images, where the metadata becomes larger than the actual data. More efficient and customisable formats such as HDF5 are becoming increasingly commonly used for storing large data. A combination of HDF5 and mzML, retaining the storage efficiency of HDF5 and the metadata and capability to store MIAPE data of mzML, is mz5 [279]. In the same way, imzML and HDF5 could be combined to develop a new, more efficient open format for MSI data.

Also in Chapter 2, SpectralAnalysis, an extensible platform for processing, analysing and visualising large spectral imaging data was presented. There still exist a vast number of preprocessing methods that have been presented in the literature but are not currently implemented. The most promising of these are wavelet based techniques such as the continuous wavelet transform (CWT). The CWT applies a wavelet function at various scales

to transform the data into the wavelet space. This essentially performs the smoothing and baseline correction preprocessing steps in a single process. It is also possible to detect peaks in the wavelet space by matching ridge lines.

Only a single preprocessing method, TopHat baseline correction, was sped up using GPGPU, however the major speed improvement was realised through the use of a more efficient algorithm. The development of more GPGPU algorithms for preprocessing methods will mitigate the cost of the transfer time as more steps of the preprocessing worklow could be performed on the GPU. If this is extended to the MVA and clustering algorithms then further speed increases can be achieved as no additional data transfer is required.

Currently only three dimensional data (two spatial dimensions and a spectral dimension) is supported within SpectralAnalysis. This could be extended to a include a third spatial dimension as would be necessary for processing multi-section MALDI data or SIMS depth profiles. This would require the development of a means to visualise not only ion images, but also MVA results and cluster results in 3D. This would fit easily into SpectralAnalysis due to the extensible nature, by extending the 'Display' class as 'ImageDisplay' and 'SpectrumDisplay' currently do. A further dimension is that of ion mobility, where multiple spectra per spatial location are acquired in a time resolved manner following the manipulation of the ion plume by some means, such as a travelling wave. The processing of such data requires additional algorithm development, as the peaks are no longer detected on a one dimensional spectrum, but on a two dimensional mobility (e.g. drift time), m/z plot. As multiple spectra are present per pixel within an image, this also results in a significant increase in the data size.

In Chapter 3 memory efficient methods for processing and reducing data to a datacube and performing PCA were presented. As demonstrated in Chapter 2, preprocessing of each spectrum can be applied in parallel and so multithreading can be employed to reduce the time cost of this method. GPGPU can also be used in combination with multithreading to provide a further speed increase. Once the covariance matrix has been generated, it is also possible to perform SVD on the GPU, provided the number of peaks retained is

sufficiently small [280]. The suite of memory efficient algorithms could also be extended to include memory efficient clustering [281] and random projection dimensionality reduction and compression [282, 283].

In Chapter 4, two algorithms for generating hierarchical representations of MSI data were presented. Currently all hierarchies were generated from a single data set and then applied to additional data sets. An extension of this would be to learn additional hierarchies from datasets that do not conform to the current models, for example the formalin fixed data presented in this chapter. Then, each of these generated hierarchies could be annotated with the experimental conditions so subsequent application to new data would rapidly determine likely conditions or analyte features. If these did not match expectations, this would indicate there was something different about the data than was expected, forming a quality control procedure.

In Chapter 5 a means to evaluate the perceptual linearity of existing colour schemes used to present MSI data was presented. All data presented in this chapter were 2D datasets. The consideration of perceptual linearity is equally important in 3D datasets, where there are additional complicating factors. Visualisation in 3D typically involves the use of lighting and shading of the presented object(s) to give the illusion of depth, however as these effects alter the colour, and the colour maps directly to an intensity value on the colour scale, they are also changing the interpreted intensity.

In Chapter 6 the software and algorithms developed throughout this thesis were applied to the study of traumatic brain injury models. Manual exploration and use of multivariate analysis methods such as PCA did not reveal any differences between the injured hemisphere of the brain and the control hemisphere, however the hierarchical composition algorithms identified multiple ion images which appear elevated in the injured hemisphere. The elevated ions were not detected in subsequent experiment performed on a Synapt G2S. Further work is required to repeat these imaging experiments to confirm the localisation of these ions in the injured hemisphere. MS/MS experiments could then be performed to identify the unknown ions. As the brains were sectioned at the site of

injury, the exact anatomical location of these sections is unknown. The method for automatic registration to the Allen brain atlas developed by Abdelmoula *et al.* [252] could be used to more accurately determine the location at which the sections were acquired, providing a means to assess the comparability of each brain section.

It is becoming an increasingly common desire to incorporate complementary data alongside mass spectrometry imaging, either through the use of different matrices and mass ranges to target different classes of molecule on the same sample using MALDI MSI [284], using multiple ionisation techniques [129, 285] or incorporating multiple spectral imaging modalities [286, 130]. The primary aim of incorporating complementary data is to gain additional insight that was not available from either modality in isolation. One technique for achieving this that has recently been applied to mass spectrometry imaging is data fusion [287, 288]. As SpectralAnalysis is capable of handling data from multiple MSI instruments as well as from different imaging modalities, it is a well suited platform to cater to this trend. One prerequisite for advancing support for this is the development and integration of multi-modality data registration. This could be achieved by modifying a recently published automatic optical-MSI data registration workflow, making use of other work described within this thesis and which are already integrated features in SpectralAnalysis such as the memory efficient handling and reduction of data [252].

Even when a single mass spectrometer is used within a study, the number of datasets being analysed together is increasing. Recently, 9 tissue sections per time point resulting in a total of 27 sections were analysed to investigate protein digestion [289], 63 sections from gastric cancer and 32 from breast cancer were analysed to investigate intratumour heterogeneity [290] and 96 sections taken from 32 mice were analysed to investigate the consequences of cortical spreading depression [291]. With the trend towards larger and larger data, it is important that the software tools used to analyse such studies are capable of supporting such volumes of data. The work presented within this thesis provides routines for visualising, preprocessing, reducing and performing multivariate analysis on data of such sizes, and supports data of even greater sizes still. The integrated multi-

variate analysis and the hierarchical composition provide complementary, powerful tools for investigating and aiding interpretation of complex datasets such as those generated within large cohort studies.

LIST OF REFERENCES

[1] Peter Arrowsmith. Laser ablation of solids for elemental analysis by inductively coupled plasma mass spectrometry. *Analytical Chemistry*, 59(10):1437–1444, 1987.

[2] M. Stoeckli, D. Staab, and A. Schweitzer. Compound and metabolite distribution measured by maldi mass spectrometric imaging in whole-body tissue sections. *International Journal of Mass Spectrometry*, 260(2-3):195–202, 2007.

[3] Dale S Cornett, Sara L Frappier, and Richard M Caprioli. Maldi-fticr imaging mass spectrometry of drugs and metabolites in tissue. *Analytical chemistry*, 80(14):5648–5653, 2008.

[4] 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.

[5] Andrew D. Palmer, Rian Griffiths, Iain Styles, Ela Claridge, Antonio Calcagni, and Josephine Bunch. Sucrose cryo-protection facilitates imaging of whole eye sections by maldi mass spectrometry. *Journal of Mass Spectrometry*, 47(2):237–241, 2012.

[6] Ioana M Taban, AF Altelaar, Yuri EM van der Burgt, Liam A McDonnell, Ron Heeren, Jens Fuchser, and Gökhan Baykut. Imaging of peptides in the rat brain using maldi-fticr mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 18(1):145–151, 2007.

[7] Sheerin Khatib-Shahidi, Malin Andersson, Jennifer L Herman, Todd A Gillespie, and Richard M Caprioli. Direct molecular analysis of whole-body animal tissue sections by imaging maldi mass spectrometry. *Analytical chemistry*, 78(18):6448–6456, 2006.

[8] Alfred Benninghoven. Chemical analysis of inorganic and organic surfaces and thin films by static time-of-flight secondary ion mass spectrometry (tof-sims). *Angewandte Chemie International Edition in English*, 33(10):1023–1043, 1994.

[9] Alina Zamfir and Jasna Peter-Katalinić. Capillary electrophoresis-mass spectrometry for glycoscreening in biomedical research. *Electrophoresis*, 25(13):1949–1963, 2004.

[10] Julie A Vrana, Jeffrey D Gamez, Benjamin J Madden, Jason D Theis, H Robert Bergen III, and Ahmet Dogan. Classification of amyloidosis by laser microdissection and mass spectrometry–based proteomic analysis in clinical biopsy specimens. *Blood*, 114(24):4957–4959, 2009.

[11] Hong Mei, Yunsheng Hsieh, Cymbylene Nardo, Xiaoying Xu, Shiyong Wang, Kwokei Ng, and Walter A Korfmacher. Investigation of matrix effects in bioanalytical high-performance liquid chromatography/tandem mass spectrometric assays: application to drug discovery. *Rapid Communications in Mass Spectrometry*, 17(1):97–103, 2003.

[12] John R Yates. Mass spectrometry: from genomics to proteomics. *Trends in Genetics*, 16(1):5–8, 2000.

[13] F Couderc, JM Berjeaud, JC Promé, R Graham Cooks, and Alan L Rockwood. Letters to the editor. *Rapid Communications in Mass Spectrometry*, 5(2):92–93, 1991.

[14] A. D. McNaught and A. Wilkinson. *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book").* Blackwell Scientific Publications, Oxford, 1997.

[15] CS Ho, CWK Lam, MHM Chan, RCK Cheung, LK Law, LCW Lit, KF Ng, MWM Suen, and HL Tai. Electrospray ionisation mass spectrometry: principles and clinical applications. *The Clinical Biochemist Reviews*, 24(1):3, 2003.

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

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

[18] Michael Karas, Ute Bahr, Kerstin Strupat, Franz Hillenkamp, Anthony Tsarbopoulos, and Birendra N Pramanik. Matrix dependence of metastable fragmentation of glycoproteins in maldi tof mass spectrometry. *Analytical Chemistry*, 67(3):675–679, 1995.

[19] Julien Franck, Karim Arafah, Mohamed Elayed, David Bonnel, Daniele Vergara, Amélie Jacquet, Denis Vinatier, Maxence Wisztorski, Robert Day, Isabelle Fournier, and Michel Salzet. Maldi imaging mass spectrometry state of the art technology in clinical proteomics. *Molecular & Cellular Proteomics*, 8(9):2023–2033, 2009.

[20] Rian L Griffiths, Joscelyn Sarsby, Emily J Guggenheim, Alan M Race, Rory T Steven, Janine Fear, Patricia F Lalor, and Josephine Bunch. Formal lithium fixation improves direct analysis of lipids in tissue by mass spectrometry. *Analytical chemistry*, 85(15):7146–7153, 2013.

[21] Rory T Steven, Alan M Race, and Josephine Bunch. para-nitroaniline is a promising matrix for maldi-ms imaging on intermediate pressure ms systems. *Journal of The American Society for Mass Spectrometry*, 24(5):801–804, 2013.

[22] Yanfeng Chen, Ying Liu, Jeremy Allegood, Elaine Wang, Begoña Cachón-González, Timothy M Cox, Alfred H Merrill Jr, and M Cameron Sullards. Imaging maldi mass spectrometry of sphingolipids using an oscillating capillary nebulizer matrix application system. In *Mass Spectrometry Imaging*, pages 131–146. Springer, 2010.

[23] Dodge L Baluya, Timothy J Garrett, and Richard A Yost. Automated maldi matrix deposition method with inkjet printing for imaging mass spectrometry. *Analytical chemistry*, 79(17):6862–6867, 2007.

[24] Joseph A Hankin, Robert M Barkley, and Robert C Murphy. Sublimation as a method of matrix application for mass spectrometric imaging. *Journal of the American Society for Mass Spectrometry*, 18(9):1646–1652, 2007.

[25] DESI Imaging. `http://www.prosolia.com/desi-imaging`. Accessed: 12/08/2015.

[26] R.L. Griffiths and J. Bunch. A survey of useful salt additives in matrix-assisted laser desorption/ionization mass spectrometry and tandem mass spectrometry of lipids: introducing nitrates for improved analysis. *Rapid Communications in Mass Spectrometry*, 26(13):1557–1566, 2012.

[27] John C Jurchen, Stanislav S Rubakhin, and Jonathan V Sweedler. Maldi-ms imaging of features smaller than the size of the laser beam. *Journal of the American Society for Mass Spectrometry*, 16(10):1654–1659, 2005.

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

[29] JH Beynon. Iupac recommendations on symbolism and nomenclature for mass-spectroscopy, 1977.

[30] John FJ Todd. Recommendations for nomenclature and symbolism for mass spectroscopy (including an appendix of terms used in vacuum technology).(recommendations 1991). *Pure and applied chemistry*, 63(10):1541–1566, 1991.

[31] Kermit K Murray, Robert K Boyd, Marcos N Eberlin, G John Langley, Liang Li, and Yasuhide Naito. Definitions of terms relating to mass spectrometry (iupac recommendations 2013). *Pure and Applied Chemistry*, 85(7):1515–1609, 2013.

[32] Applied Biosystems. *QSTAR XL LC/MS/TOF System*, June 2005.

[33] Chris F Taylor, Norman W Paton, Kathryn S Lilley, Pierre-Alain Binz, Randall K Julian, Andrew R Jones, Weimin Zhu, Rolf Apweiler, Ruedi Aebersold, Eric W Deutsch, et al. The minimum information about a proteomics experiment (miape). *Nature biotechnology*, 25(8):887–893, 2007.

[34] Patrick GA Pedrioli, Jimmy K Eng, Robert Hubley, Mathijs Vogelzang, Eric W Deutsch, Brian Raught, Brian Pratt, Erik Nilsson, Ruth H Angeletti, Rolf Apweiler, Kei Cheung, Catherine E Costello, Henning Hermjakob, Sequin Huang, Randall K Julian Jr, Eugene Kapp, Mark E McComb, Stephen G Oliver, Gilbert Omenn, Norman W Paton, Richard Simpson, Richard Smith, Chris F Taylor, Weimin Zhu, and Ruedi Aebersold. A common open representation of mass spectrometry data and its application to proteomics research. *Nature biotechnology*, 22(11):1459–1466, 2004.

[35] Peter Lampen, Heinrich Hillig, Antony N Davies, and Michael Linscheid. Jcamp-dx for mass spectrometry. *Applied spectroscopy*, 48(12):1545–1552, 1994.

[36] Lennart Martens, Matthew Chambers, Marc Sturm, Darren Kessner, Fredrik Levander, Jim Shofstahl, Wilfred H Tang, Andreas Römpp, Steffen Neumann, Angel D Pizarro, Luisa Montecchi-Palazzi, Natalie Tasman, Mike Coleman, Florian Reisinger, Puneet Souda, Henning Hermjakob, Pierre-Alain Binz, and Eric W Deutsch. mzmla community standard for mass spectrometry data. *Molecular & Cellular Proteomics*, 10(1):R110–000133, 2011.

[37] Sandra Orchard, Henning Hermjakob, Randall K Julian, Kai Runte, David Sherman, Jérôme Wojcik, Weimin Zhu, and Rolf Apweiler. Common interchange standards for proteomics data: Public availability of tools and schema. report on the proteomic standards initiative workshop, 2nd annual hupo congress, montreal, canada, 8–11th october 2003. *Proteomics*, 4(2):490–491, 2004.

[38] Sandra Orchard, Henning Hermjakob, Chris F Taylor, Frank Potthast, Phil Jones,

Weimin Zhu, Randall K Julian, and Rolf Apweiler. Further steps in standardisation report of the second annual proteomics standards initiative spring workshop (siena, italy 17–20th april 2005). *Proteomics*, 5(14):3552–3555, 2005.

[39] Eric Deutsch. mzml: a single, unifying data format for mass spectrometer output. *Proteomics*, 8(14):2776–2777, 2008.

[40] Matthew C Chambers, Brendan Maclean, Robert Burke, Dario Amodei, Daniel L Ruderman, Steffen Neumann, Laurent Gatto, Bernd Fischer, Brian Pratt, Jarrett Egertson, Katherine Hoff, Darren Kessner, Natalie Tasman, Nicholas Shulman, Barbara Frewen, Tahmina A Baker, Mi-Youn Brusniak, Christopher Paulse, David Creasy, Lisa Flashner, Kian Kani, Chris Moulding, Sean L Seymour, Lydia M Nuwaysir, Brent Lefebvre, Frank Kuhlmann, Joe Roark, Paape Rainer, Suckau Detlev, Tina Hemenway, Andreas Huhmer, James Langridge, Brian Connolly, Trey Chadick, Krisztina Holly, Josh Eckels, Eric W Deutsch, Robert L Moritz, Jonathan E Katz, David B Agus, Michael MacCoss, David L Tabb, and Parag Mallick. A cross-platform toolkit for mass spectrometry and proteomics. *Nature biotechnology*, 30(10):918–920, 2012.

[41] Burkhard A Schäfer, Dominik Poetz, and Gary W Kramer. Documenting laboratory workflows using the analytical information markup language. *Journal of the Association for Laboratory Automation*, 9(6):375–381, 2004.

[42] Naofumi Hosokawa, Yuki Sugiura, and Mitsutoshi Setou. Ion image reconstruction using biomap software. In *Imaging Mass Spectrometry*, pages 113–126. Springer, 2010.

[43] ANALYZE^{TM} 7.5 File Format. `http://eeg.sourceforge.net/ANALYZE75.pdf`. Accessed: 06/08/2015.

[44] Andreas Römpp, Thorsten Schramm, Alfons Hester, Ivo Klinkert, Jean-Pierre Both, Ron MA Heeren, Markus Stöckli, and Bernhard Spengler. imzml: imaging mass

spectrometry markup language: a common data format for mass spectrometry imaging. In *Data Mining in Proteomics*, pages 205–224. Springer, 2011.

[45] Thorsten Schramm, Alfons Hester, Ivo Klinkert, Jean-Pierre Both, Ron Heeren, Alain Brunelle, Olivier Laprévote, Nicolas Desbenoit, Marie-France Robbe, Markus Stoeckli, Bernhard Spengler, and Andreas Römpp. imzmla common data format for the flexible exchange and processing of mass spectrometry imaging data. *Journal of proteomics*, 75(16):5106–5110, 2012.

[46] Liam A McDonnell, Andreas Römpp, Benjamin Balluff, Ron MA Heeren, Juan Pablo Albar, Per E Andrén, Garry L Corthals, Axel Walch, and Markus Stoeckli. Discussion point: reporting guidelines for mass spectrometry imaging. *Analytical and bioanalytical chemistry*, 407(8):2035–2045, 2014.

[47] Ivo Klinkert, Kamila Chughtai, Shane R Ellis, and Ron Heeren. Methods for full resolution data exploration and visualization for large 2d and 3d mass spectrometry imaging datasets. *International Journal of Mass Spectrometry*, 362(0):40–47, 2014.

[48] Janina Oetjen, Kirill Veselkov, Jeramie Watrous, James S. McKenzie, Michael Becker, Lena Hauberg-Lotte, Jan Hendrik Kobarg, Nicole Strittmatter, Anna K. Mróz, Franziska Hoffmann, Dennis Trede, Andrew Palmer, Stefan Schiffler, Klaus Steinhorst, Michaela Aichler, Robert Goldin, Orlando Guntinas-Lichius, Ferdinand Eggeling, Herbert Thiele, Kathrin Maedler, Axel Walch, Peter Maass, Pieter C. Dorrestein, Zoltan Takats, and Theodore Alexandrov. Benchmark datasets for 3d maldi-and desi-imaging mass spectrometry. *GigaScience*, 4(1):1–8, 2015.

[49] Dhaka Ram Bhandari, Qing Wang, Wolfgang Friedt, Bernhard Spengler, Sven Gottwald, and Andreas Römpp. High resolution mass spectrometry imaging of plant tissues: towards a plant metabolite atlas. *Analyst*, 140(22):7696–7709, 2015.

[50] 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.

[51] Guillaume Robichaud, Kenneth P Garrard, Jeremy A Barry, and David C Muddiman. Msireader: An open-source interface to view and analyze high resolving power ms imaging files on matlab platform. *Journal of The American Society for Mass Spectrometry*, 24(5):718–721, 2013.

[52] 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.

[53] R Mitchell Parry, Asiri S Galhena, Chaminda M Gamage, Rachel V Bennett, May D Wang, and Facundo M Fernández. Omnispect: An open matlab-based tool for visualization and analysis of matrix-assisted laser desorption/ionization and desorption electrospray ionization mass spectrometry images. *Journal of The American Society for Mass Spectrometry*, 24(4):646–649, 2013.

[54] Oliver Rübel, Annette Greiner, Shreyas Cholia, Katherine Louie, E Wes Bethel, Trent R Northen, and Benjamin P Bowen. Openmsi: A high-performance web-based platform for mass spectrometry imaging. *Analytical chemistry*, 85(21):10354–10361, 2013.

[55] C Paschke, A Leisner, A Hester, K Maass, S Guenther, W Bouschen, and B Spengler. Mirion - a software package for automatic processing of mass spectrometric images. *Journal of The American Society for Mass Spectrometry*, 24(8):1296–1306, 2013.

[56] SCiLS / SCiLS Lab - The statistical analysis software. `http://scils.de/software/`, 2014. Accessed: 21/05/2014.

[57] PREMIER Biosoft. Maldivision. `http://www.premierbiosoft.com/maldi-tissue-imaging/index.html`. Accessed: 20/03/2016.

[58] imabiotech. Quantinetix. `https://www.imabiotech.com/Benefits`. Accessed: 20/03/2016.

[59] Applied Biosystems. *Analyst QS Administrator's Guide*, July 2004.

[60] Applied Biosystems. *TissueView*, June 2010.

[61] Bruker Daltonics. *fleximaging 3.0 User Manual*, June 2011.

[62] Bruker Daltonics. *flexanalysis 3.3 User Manual*, June 2009.

[63] Thermo Fisher Scientific. *Thermo Xcalibur Qualitative Analysis User Guide*, September 2010.

[64] Thermo Fisher Scientific. *Thermo ImageQuest Version 1.0.1 User Guide*, May 2009.

[65] Thermo Fisher Scientific. *Thermo Xcalibur Acquisition and Processing User Guide*, September 2010.

[66] Waters Corporation. *MassLynx 4.1 Getting Started Guide*, 2005.

[67] Novartis. *BioMAP 3x*, August 2005.

[68] Kyle D. Bemis and April Harry. *Cardinal: Analytic tools for mass spectrometry imaging*, April 2016.

[69] Dante Mantini, Francesca Petrucci, Damiana Pieragostino, Piero Del Boccio, Marta Di Nicola, Carmine Di Ilio, Giorgio Federici, Paolo Sacchetta, Silvia Comani, and Andrea Urbani. Limpic: a computational method for the separation of protein maldi-tof-ms signals from noise. *BMC bioinformatics*, 8(1):101, 2007.

[70] FOM Institute AMOLF. *Datacube Explorer User Manual*, May 2014.

[71] NC State University. *MSiReader User's Manual*, December 2012.

[72] Kevin R Coombes, Keith A Baggerly, and Jeffrey S Morris. Pre-processing mass spectrometry data. In *Fundamentals of Data Mining in Genomics and Proteomics*, pages 79–102. Springer, 2007.

[73] J Harms. Automatic dead-time correction for multichannel pulse-height analyzers at variable counting rates. *Nuclear Instruments and Methods*, 53:192–196, 1967.

[74] T Stephan, J Zehnpfenning, and A Benninghoven. Correction of dead time effects in time-of-flight mass spectrometry. *Journal of Vacuum Science & Technology A*, 12(2):405–410, 1994.

[75] Bonnie J Tyler and Richard E Peterson. Dead-time correction for time-of-flight secondary-ion mass spectral images: a critical issue in multivariate image analysis. *Surface and Interface Analysis*, 45(1):475–478, 2013.

[76] 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.

[77] Chao Yang, Zengyou He, and Weichuan Yu. Comparison of public peak detection algorithms for maldi mass spectrometry data analysis. *BMC bioinformatics*, 10(1):4, 2009.

[78] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[79] James Riordon, Elizabeth Zubritsky, and Alan Newman. Top 10 articles. *Analytical chemistry*, 72(9):324–A, 2000.

[80] Ronald W Schafer. What is a savitzky-golay filter?[lecture notes]. *Signal Processing Magazine, IEEE*, 28(4):111–117, 2011.

[81] 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.

[82] Arsalan S Haqqani, John F Kelly, and Danica B Stanimirovic. Quantitative protein profiling by mass spectrometry using label-free proteomics. In *Genomics Protocols*, pages 241–256. Springer, 2008.

[83] MM Burrell, CJ Earnshaw, and MR Clench. Imaging matrix assisted laser desorption ionization mass spectrometry: a technique to map plant metabolites within tissues at high spatial resolution. *Journal of experimental botany*, 58(4):757–763, 2007.

[84] David A Pirman, Richard F Reich, Andras Kiss, Ron MA Heeren, and Richard A Yost. Quantitative maldi tandem mass spectrometric imaging of cocaine from brain tissue with a deuterated internal standard. *Analytical chemistry*, 85(2):1081–1089, 2012.

[85] Patrik Källback, Mohammadreza Shariatgorji, Anna Nilsson, and Per E Andrén. Novel mass spectrometry imaging software assisting labeled normalization and quantitation of drugs and neuropeptides directly in tissue sections. *Journal of proteomics*, 75(16):4941–4951, 2012.

[86] Liam A McDonnell, Alexandra Van Remoortere, Nico De Velde, Rene JM Van Zeijl, and André M Deelder. Imaging mass spectrometry data reduction: automated feature identification and extraction. *Journal of the American Society for Mass Spectrometry*, 21(12):1969–1978, 2010.

[87] A. Römpp, S. Guenther, Z. Takats, and B. Spengler. Mass spectrometry imaging with high resolution in mass and space (hr 2 msi) for reliable investigation of drug compound distributions on the cellular level. *Analytical and bioanalytical chemistry*, 401(1):65–73, 2011.

[88] M. Stoeckli, D. Staab, A. Schweitzer, J. Gardiner, and D. Seebach. Imaging of a [beta]-peptide distribution in whole-body mice sections by maldi mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 18(11):1921–1924, 2007.

[89] J.S. Fletcher, S. Rabbani, A. Henderson, N.P. Lockyer, and J.C. Vickerman. Three-dimensional mass spectral imaging of hela-m cells–sample preparation, data interpretation and visualisation. *Rapid Communications in Mass Spectrometry*, 25(7):925–932, 2011.

[90] S. Ghosal, S.J. Fallon, T.J. Leighton, K.E. Wheeler, M.J. Kristo, I.D. Hutcheon, and P.K. Weber. Imaging and 3d elemental characterization of intact bacterial spores by high-resolution secondary ion mass spectrometry. *Analytical chemistry*, 80(15):5986–5992, 2008.

[91] D. Breitenstein, C.E. Rommel, R. Möllers, J. Wegener, and B. Hagenhoff. The chemical composition of animal cells and their intracellular compartments reconstructed from 3d mass spectrometry. *Angewandte Chemie International Edition*, 46(28):5332–5335, 2007.

[92] J.S. Fletcher, N.P. Lockyer, and J.C. Vickerman. Developments in molecular sims depth profiling and 3d imaging of biological systems using polyatomic primary ions. *Mass spectrometry reviews*, 30(1):142–174, 2011.

[93] T.K. Sinha, S. Khatib-Shahidi, T.E. Yankeelov, K. Mapara, M. Ehtesham, D.S. Cornett, B.M. Dawant, R.M. Caprioli, and J.C. Gore. Integrating spatially resolved three-dimensional maldi ims with in vivo magnetic resonance imaging. *Nature methods*, 5(1):57–59, 2007.

[94] Xingchuang Xiong, Wei Xu, LiviaS. Eberlin, JustinM. Wiseman, Xiang Fang, You Jiang, Zejian Huang, Yukui Zhang, R.Graham Cooks, and Zheng Ouyang. Data processing for 3d mass spectrometry imaging. *Journal of The American Society for Mass Spectrometry*, 23:1147–1156, 2012.

[95] Anna C Crecelius, D Shannon Cornett, Richard M Caprioli, Betsy Williams, Benoit M Dawant, and Bobby Bodenheimer. Three-dimensional visualization of protein expression in mouse brain structures using imaging mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 16(7):1093–1099, 2005.

[96] E.H. Seeley and R.M. Caprioli. 3d imaging by mass spectrometry: A new frontier. *Analytical chemistry*, 84(5):2105–2110, 2012.

[97] Dennis Trede, Stefan Schiffler, Michael Becker, Stefan Wirtz, Klaus Steinhorst, Jan Strehlow, Michaela Aichler, Jan Hendrik Kobarg, Janina Oetjen, Andrey Dyatlov, Stefan Heldmann, Axel Walch, Herbert Thiele, Peter Maass, and Theodore Alexandrov. 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.

[98] R.J.A. Goodwin, A.R. Pitt, D. Harrison, S.K. Weidt, P.R.R. Langridge-Smith, M.P. Barrett, and C. Logan Mackay. Matrix-free mass spectrometric imaging using laser desorption ionisation fourier transform ion cyclotron resonance mass spectrometry. *Rapid Communications in Mass Spectrometry*, 25(7):969–972, 2011.

[99] D.F. Smith, K. Aizikov, M.C. Duursma, F. Giskes, D.J. Spaanderman, L.A. McDonnell, P.B. O'Connor, and R.M.A. Heeren. An external matrix-assisted laser desorption ionization source for flexible ft-icr mass spectrometry imaging with internal calibration on adjacent samples. *Journal of the American Society for Mass Spectrometry*, 22(1):130–137, 2011.

[100] S.E. Reichenbach, A. Henderson, R. Lindquist, and Q. Tao. Efficient encoding and rapid decoding for interactive visualization of large three-dimensional hyperspectral chemical images. *Rapid Communications in Mass Spectrometry*, 23(9):1229–1233, 2009.

[101] S.E. Reichenbach, X. Tian, R. Lindquist, Q. Tao, A. Henderson, and J.C. Vicker-

man. Visualization and analysis of large three-dimensional hyperspectral images. In *SPIE Defense, Security, and Sensing*, pages 734108–734108. International Society for Optics and Photonics, 2009.

[102] S.O. Deininger, M.P. Ebert, A. Fütterer, M. Gerhard, and C. Röcken. 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.

[103] L.A. Klerk, A. Broersen, I.W. Fletcher, R. van Liere, and R. Heeren. Extended data analysis strategies for high resolution imaging ms: New methods to deal with extremely large image hyperspectral datasets. *International Journal of Mass Spectrometry*, 260(2):222–236, 2007.

[104] Judith M Fonville, Claire Carter, Olivier Cloarec, Jeremy K Nicholson, John C Lindon, Josephine Bunch, and Elaine Holmes. Robust data processing and normalization strategy for maldi mass spectrometric imaging. *Analytical chemistry*, 84(3):1310–1319, 2012.

[105] Emrys A Jones, Sören-Oliver Deininger, Pancras CW Hogendoorn, André M Deelder, and Liam A McDonnell. Imaging mass spectrometry statistical analysis. *Journal of proteomics*, 75(16):4962–4989, 2012.

[106] 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.

[107] G. McCombie, D. Staab, M. Stoeckli, and R. Knochenmuss. Spatial and spectral correlations in maldi mass spectrometry images by clustering and multivariate analysis. *Analytical chemistry*, 77(19):6118–6124, 2005.

[108] M.C. Biesinger, P.Y. Paepegaey, N.S. McIntyre, R.R. Harbottle, and N.O. Petersen. Principal component analysis of tof-sims images of organic monolayers. *Analytical chemistry*, 74(22):5711–5716, 2002.

[109] Paul J Trim, Sally J Atkinson, Alessandra P Princivalle, Peter S Marshall, Andrew West, and Malcolm R Clench. Matrix-assisted laser desorption/ionisation mass spectrometry imaging of lipids in rat brain tissue with integrated unsupervised and supervised multivariant statistical analysis. *Rapid Communications in Mass Spectrometry*, 22(10):1503–1509, 2008.

[110] A. Broersen and R. Van Liere. Transfer functions for imaging spectroscopy data using principal component analysis. In *Proc. Eurographics/IEEE VGTC Symposium on Visualization*. Citeseer, 2005.

[111] A.F.M. Altelaar, S.L. Luxembourg, L.A. McDonnell, S.R. Piersma, and R.M.A. Heeren. Imaging mass spectrometry at cellular length scales. *Nature protocols*, 2(5):1185–1196, 2007.

[112] P. Sjövall, J. Lausmaa, and B. Johansson. Mass spectrometric imaging of lipids in brain tissue. *Analytical chemistry*, 76(15):4271–4278, 2004.

[113] R. Van de Plas, F. Ojeda, M. Dewil, L. Van Den Bosch, B. De Moor, and E. Waelkens. Prospective exploration of biochemical tissue composition via imaging mass spectrometry guided by principal component analysis. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 12, pages 3–7. Citeseer, 2007.

[114] Victor W Lee, Changkyu Kim, Jatin Chhugani, Michael Deisher, Daehyun Kim, Anthony D Nguyen, Nadathur Satish, Mikhail Smelyanskiy, Srinivas Chennupaty, Per Hammarlund, Ronak Singhal, and Pradeep Dubey. Debunking the 100x gpu vs. cpu myth: an evaluation of throughput computing on cpu and gpu. In *ACM SIGARCH Computer Architecture News*, volume 38, pages 451–460. ACM, 2010.

[115] Marc Baboulin, Alfredo Buttari, Jack Dongarra, Jakub Kurzak, Julie Langou, Julien Langou, Piotr Luszczek, and Stanimire Tomov. Accelerating scientific computations with mixed precision algorithms. *Computer Physics Communications*, 180(12):2526–2533, 2009.

[116] Lydia Ashleigh Baumgardner, Avinash Kumar Shanmugam, Henry Lam, Jimmy K Eng, and Daniel B Martin. Fast parallel tandem mass spectral library searching using gpu hardware acceleration. *Journal of proteome research*, 10(6):2882–2888, 2011.

[117] Thomas Gröger and Ralf Zimmermann. Application of parallel computing to speed up chemometrics for gc× gc–tofms based metabolic fingerprinting. *Talanta*, 83(4):1289–1294, 2011.

[118] You Li, Hao Chi, Leihao Xia, and Xiaowen Chu. Accelerating the scoring module of mass spectrometry-based peptide identification using gpus. *BMC Bioinformatics*, 15(1):121, 2014.

[119] Rene Hussong, Barbara Gregorius, Andreas Tholey, and Andreas Hildebrandt. Highly accelerated feature detection in proteomics data sets using modern graphics processing units. *Bioinformatics*, 25(15):1937–1943, 2009.

[120] Jan Hendrik Kobarg, Peter Maass, Janina Oetjen, Oren Tropp, Eyal Hirsch, Chen Sagiv, Mohammad Golbabaee, and Pierre Vandergheynst. Numerical experiments with maldi imaging data. *Advances in Computational Mathematics*, pages 1–16, 2013.

[121] Emrys A Jones, René JM van Zeijl, Per E Andrén, André M Deelder, Lex Wolters, and Liam A McDonnell. High speed data processing for imaging ms-based molecular histology using graphical processing units. *Journal of the American Society for Mass Spectrometry*, 23(4):745–752, 2012.

[122] Sebastian Gibb and Korbinian Strimmer. Maldiquant: a versatile r package for the analysis of mass spectrometry data. *Bioinformatics*, 28(17):2270–2271, 2012.

[123] Jeffrey S Morris, Kevin R Coombes, John Koomen, Keith A Baggerly, and Ryuji Kobayashi. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics*, 21(9):1764–1775, 2005.

[124] Kevin R Coombes, Spiridon Tsavachidis, Jeffrey S Morris, Keith A Baggerly, Mien-Chie Hung, and Henry M Kuerer. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics*, 5(16):4107–4117, 2005.

[125] Joseph Gil and Ron Kimmel. Efficient dilation, erosion, opening, and closing algorithms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(12):1606–1617, 2002.

[126] Emrys A Jones, Alexandra van Remoortere, René JM van Zeijl, Pancras CW Hogendoorn, JVMG Bovee, 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.

[127] Allan Aasbjerg Nielsen. Kernel maximum autocorrelation factor and minimum noise fraction transformations. *Image Processing, IEEE Transactions on*, 20(3):612–624, 2011.

[128] 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.

[129] Livia S Eberlin, Xiaohui Liu, Christina R Ferreira, Sandro Santagata, Nathalie YR Agar, and R Graham Cooks. Desorption electrospray ionization then maldi mass spectrometry imaging of lipid and protein distributions in single tissue sections. *Analytical chemistry*, 83(22):8366–8371, 2011.

[130] Dorothy R Ahlf, Rachel N Masyuko, Amanda B Hummon, and Paul W Bohn. Correlated mass spectrometry imaging and confocal raman microscopy for studies of three-dimensional cell culture sections. *Analyst*, 139(18):4578–4585, 2014.

[131] C. Ordonez. Statistical model computation with udfs. *IEEE Transactions on Knowledge and Data Engineering*, 22(12):1752–1765, 2010.

[132] Erin Shammel Baker, Eric A. Livesay, Daniel J. Orton, Ronald J. Moore, William F. Danielson, David C. Prior, Yehia M. Ibrahim, Brian L. LaMarche, Anoop M. Mayampurath, Athena A. Schepmoes, Derek F. Hopkins, Keqi Tang, Richard D. Smith, and Mikhail E. Belov. An lc-ims-ms platform providing increased dynamic range for high-throughput proteomic studies. *Journal of proteome research*, 9(2):997–1006, 2010.

[133] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999.

[134] Paul Geladi and Bruce R Kowalski. Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17, 1986.

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

[136] Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, et al. Fiji: an open-source platform for biological-image analysis. *Nature methods*, 9(7):676–682, 2012.

[137] O. Woodford. vol3d v2. `http://www.mathworks.com/matlabcentral/fileexchange/22940-vol3d-v2`. Accessed: 13/07/2015.

[138] G. Stone, D. Clifford, J.O.R. Gustafsson, S.R. McColl, and P. Hoffmann. Visualisation in imaging mass spectrometry using the minimum noise fraction transform. *BMC Research Notes*, 5(1):419, 2012.

[139] MS Wagner and D.G. Castner. Characterization of adsorbed protein films by time-of-flight secondary ion mass spectrometry with principal component analysis. *Langmuir*, 17(15):4649–4660, 2001.

[140] L.A. McDonnell, A. van Remoortere, R.J.M. van Zeijl, and A.M. Deelder. Mass spectrometry image correlation: quantifying colocalization. *Journal of proteome research*, 7(8):3619–3627, 2008.

[141] Bonnie J Tyler, Gaurav Rayal, and David G Castner. Multivariate analysis strategies for processing tof-sims images of biomaterials. *Biomaterials*, 28(15):2412–2423, 2007.

[142] Sanja Fidler and Aleš Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[143] Michael W Senko, Steven C Beu, and Fred W McLaffertycor. 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.

[144] 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.

[145] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[146] Kenneth O McGraw and Seok P Wong. Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1):30, 1996.

[147] Kenneth Moreland. Diverging color maps for scientific visualization. In *Advances in Visual Computing*, pages 92–103. Springer, 2009.

[148] Bernice E Rogowitz, Lloyd A Treinish, and Steve Bryson. How not to lie with visualization. *Computers in Physics*, 10(3):268–273, 1996.

[149] David Borland and Russell M Taylor II. Rainbow color map (still) considered harmful. *IEEE computer graphics and applications*, 27(2):14–17, 2007.

[150] Adam Light and Patrick J Bartlein. The end of the rainbow? color schemes for improved data graphics. *Eos, Transactions American Geophysical Union*, 85(40):385–391, 2004.

[151] Reto Stauffer, Georg J Mayr, Markus Dabernig, and Achim Zeileis. *Somewhere over the rainbow: How to make effective use of colors in meteorological visualizations.* University of Innsbruck. Department of Public Finance, 2013.

[152] Marko Tkalcic and Jurij F Tasic. Colour spaces: perceptual, historical and applicational background. In *Eurocon*, 2003.

[153] Vischeck. Vischeck. `http://www.vischeck.com/`, 2014. [Online; accessed 24-July-2014].

[154] AF Maarten Altelaar, Ivo Klinkert, Kees Jalink, Robert PJ de Lange, Roger AH Adan, Ron MA Heeren, and Sander R Piersma. Gold-enhanced biomolecular surface imaging of cells and tissue by sims and maldi mass spectrometry. *Analytical chemistry*, 78(3):734–742, 2006.

[155] Sally J Atkinson, Paul M Loadman, Chris Sutton, Laurence H Patterson, and Malcolm R Clench. Examination of the distribution of the bioreductive drug aq4n and its active metabolite aq4 in solid tumours by imaging matrix-assisted laser desorption/ionisation mass spectrometry. *Rapid communications in mass spectrometry*, 21(7):1271–1276, 2007.

[156] Farida Benabdellah, David Touboul, Alain Brunelle, and Olivier Laprévote. In situ

primary metabolites localization on a rat brain section by chemical mass spectrometry imaging. *Analytical chemistry*, 81(13):5557–5560, 2009.

[157] 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.

[158] Josephine Bunch, Malcolm R Clench, and Don S Richards. Determination of pharmaceutical compounds in skin by imaging matrix-assisted laser desorption/ionisation mass spectrometry. *Rapid communications in mass spectrometry*, 18(24):3051–3060, 2004.

[159] Subhash Chandra. Quantitative imaging of chemical composition in single cells by secondary ion mass spectrometry: cisplatin affects calcium stores in renal epithelial cells. In *Mass spectrometry imaging*, pages 113–130. Springer, 2010.

[160] Pierre Chaurand and Richard M Caprioli. Direct profiling and imaging of peptides and proteins from mammalian cells and tissue sections by mass spectrometry. *Electrophoresis*, 23(18):3125–3135, 2002.

[161] Pierre Chaurand, Sarah A Schwartz, Dean Billheimer, Baogang J Xu, Anna Crecelius, and Richard M Caprioli. Integrating histology and imaging mass spectrometry. *Analytical chemistry*, 76(4):1145–1155, 2004.

[162] Pierre Chaurand, Sarah A Schwartz, Michelle L Reyzer, and Richard M Caprioli. Imaging mass spectrometry: principles and potentials. *Toxicologic pathology*, 33(1):92–101, 2005.

[163] Pierre Chaurand, D Shannon Cornett, and Richard M Caprioli. Molecular imaging of thin mammalian tissue sections by mass spectrometry. *Current opinion in biotechnology*, 17(4):431–436, 2006.

[164] Pierre Chaurand, Jeremy L Norris, D Shannon Cornett, James A Mobley, and Richard M Caprioli. New developments in profiling and imaging of proteins from tissue sections by maldi mass spectrometry. *Journal of proteome research*, 5(11):2889–2900, 2006.

[165] Marie-Claude Djidja, Emmanuelle Claude, Marten F Snel, Peter Scriven, Simona Francese, Vikki Carolan, and Malcolm R Clench. Maldi-ion mobility separation-mass spectrometry imaging of glucose-regulated protein 78 kda (grp78) in human formalin-fixed, paraffin-embedded pancreatic adenocarcinoma tissue sections. *Journal of proteome research*, 8(10):4876–4884, 2009.

[166] 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.

[167] Ron Heeren, Donald F Smith, Jonathan Stauber, Basak Kükrer-Kaletas, and Luke MacAleese. Imaging mass spectrometry: hype or hope? *Journal of the American Society for Mass Spectrometry*, 20(6):1006–1014, 2009.

[168] DR Ifa, LM Gumaelius, LS Eberlin, NE Manicke, and RG Cooks. Forensic analysis of inks by imaging desorption electrospray ionization (desi) mass spectrometry. *Analyst*, 132(5):461–467, 2007.

[169] Julia H Jungmann, Luke MacAleese, Ronald Buijs, Frans Giskes, Ad De Snaijer, Jan Visser, Jan Visschers, Marc JJ Vrakking, and Ron Heeren. Fast, high resolution mass spectrometry imaging using a medipix pixelated detector. *Journal of the American Society for Mass Spectrometry*, 21(12):2023–2030, 2010.

[170] 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.

[171] Dennis B Lazof, Jack G Goldsmith, Thomas W Rufty, and Richard W Linton. The early entry of al into cells of intact soybean roots (a comparison of three developmental root regions using secondary ion mass spectrometry imaging). *Plant physiology*, 112(3):1289–1300, 1996.

[172] Qiang Liu, Zhong Guo, and Lin He. Mass spectrometry imaging of small molecules using desorption/ionization on silicon. *Analytical chemistry*, 79(10):3535–3541, 2007.

[173] John A McLean, Whitney B Ridenour, and Richard M Caprioli. Profiling and imaging of tissues by imaging ion mobility-mass spectrometry. *Journal of mass spectrometry*, 42(8):1099–1105, 2007.

[174] Sara G Ostrowski, Craig T Van Bell, Nicholas Winograd, and Andrew G Ewing. Mass spectrometric imaging of highly curved membranes during tetrahymena mating. *Science*, 305(5680):71–73, 2004.

[175] Thomas P Roddy, Donald M Cannon, Sara G Ostrowski, Nicholas Winograd, and Andrew G Ewing. Identification of cellular sections with imaging mass spectrometry following freeze fracture. *Analytical chemistry*, 74(16):4020–4026, 2002.

[176] Thomas P Roddy, Donald M Cannon, Chad A Meserole, Nicholas Winograd, and Andrew G Ewing. Imaging of freeze-fractured cells with in situ fluorescence and time-of-flight secondary ion mass spectrometry. *Analytical chemistry*, 74(16):4011–4019, 2002.

[177] 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.

[178] Markus Stoeckli, Terry B Farmer, and Richard M Caprioli. Automated mass spectrometry imaging with a matrix-assisted laser desorption ionization time-of-flight

instrument. *Journal of the American Society for Mass Spectrometry*, 10(1):67–71, 1999.

[179] Yuki Sugiura, Shuichi Shimma, and Mitsutoshi Setou. Two-step matrix application technique to improve ionization efficiency for matrix-assisted laser desorption/ionization in imaging mass spectrometry. *Analytical chemistry*, 78(24):8227–8235, 2006.

[180] 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.

[181] Yuki Sugiura, Shuichi Shimma, Yoshiyuki Konishi, Maki K Yamada, and Mitsutoshi Setou. Imaging mass spectrometry technology and application on ganglioside study; visualization of age-dependent accumulation of c20-ganglioside molecular species in the mouse hippocampus. *PLoS One*, 3(9):e3232, 2008.

[182] Yuki Sugiura and Mitsutoshi Setou. Imaging mass spectrometry for visualization of drug and endogenous metabolite distribution: toward in situ pharmacometabolomes. *Journal of Neuroimmune Pharmacology*, 5(1):31–43, 2010.

[183] Peter J Todd, T Gregory Schaaff, Pierre Chaurand, and Richard M Caprioli. Organic ion imaging of biological tissue with secondary ion mass spectrometry and matrix-assisted laser desorption/ionization. *Journal of Mass Spectrometry*, 36(4):355–369, 2001.

[184] Paul J Trim, Sally J Atkinson, Alessandra P Princivalle, Peter S Marshall, Andrew West, and Malcolm R Clench. Matrix-assisted laser desorption/ionisation mass spectrometry imaging of lipids in rat brain tissue with integrated unsupervised and supervised multivariant statistical analysis. *Rapid Communications in Mass Spectrometry*, 22(10):1503–1509, 2008.

[185] Maxence Wisztorski, Julien Franck, Michel Salzet, and Isabelle Fournier. Maldi direct analysis and imaging of frozen versus ffpe tissues: what strategy for which sample? In *Mass Spectrometry Imaging*, pages 303–322. Springer, 2010.

[186] Rui Hong, Jan True, and Christopher Bieniarz. Enzymatically amplified mass tags for tissue mass spectrometry imaging. *Analytical chemistry*, 86(3):1459–1467, 2014.

[187] Patrick J Horn and Kent D Chapman. Metabolite imager: customized spatial analysis of metabolite distributions in mass spectrometry imaging. *Metabolomics*, 10(2):337–348, 2014.

[188] András Kiss, Donald F Smith, Brent R Reschke, Matthew J Powell, and Ron Heeren. Top-down mass spectrometry imaging of intact proteins by laser ablation esi ft-icr ms. *Proteomics*, 14(10):1283–1289, 2014.

[189] Michelle L Reyzer, Yunsheng Hsieh, Kwokei Ng, Walter A Korfmacher, and Richard M Caprioli. Direct analysis of drug candidates in tissue by matrix-assisted laser desorption/ionization mass spectrometry. *Journal of Mass Spectrometry*, 38(10):1081–1092, 2003.

[190] 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.

[191] Demian R Ifa, Nicholas E Manicke, Allison L Dill, and R Graham Cooks. Latent fingerprint chemical imaging by mass spectrometry. *Science*, 321(5890):805–805, 2008.

[192] Mary L Kraft, Peter K Weber, Marjorie L Longo, Ian D Hutcheon, and Steven G Boxer. Phase separation of lipid membranes analyzed with high-resolution secondary ion mass spectrometry. *Science*, 313(5795):1948–1951, 2006.

[193] Stefan L Luxembourg, Todd H Mize, Liam A McDonnell, and Ron MA Heeren. High-spatial resolution mass spectrometric imaging of peptide and protein distributions on a surface. *Analytical chemistry*, 76(18):5339–5344, 2004.

[194] Trent R Northen, Oscar Yanes, Michael T Northen, Dena Marrinucci, Winnie Uritboonthai, Junefredo Apon, Stephen L Golledge, Anders Nordström, and Gary Siuzdak. Clathrate nanostructures for mass spectrometry. *Nature*, 449(7165):1033–1036, 2007.

[195] Sandra Rauser, Claudio Marquardt, Benjamin Balluff, Sören-Oliver Deininger, Christian Albers, Eckhard Belau, Ralf Hartmer, Detlev Suckau, Katja Specht, Matthias Philip Ebert, Manfred Schmitt, Michaela Aubele, Heinz Höfler, and Axel Walch. Classification of her2 receptor status in breast cancer tissues by maldi imaging mass spectrometry. *Journal of proteome research*, 9(4):1854–1863, 2010.

[196] Yvonne Schober, Sabine Guenther, Bernhard Spengler, and Andreas Römpp. Single cell matrix-assisted laser desorption/ionization mass spectrometry imaging. *Analytical chemistry*, 84(15):6293–6297, 2012.

[197] Kristina Schwamborn and Richard M Caprioli. Molecular imaging by mass spectrometrylooking beyond classical histology. *Nature Reviews Cancer*, 10(9):639–646, 2010.

[198] Jonathan Stauber, Luke MacAleese, Julien Franck, Emmanuelle Claude, Marten Snel, Basak Kükrer Kaletas, Ingrid MVD Wiel, Maxence Wisztorski, Isabelle Fournier, and Ron 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.

[199] Veronika Vidová, Petr Novák, Martin Strohalm, Jaroslav Pól, Vladimír Havlíček, and Michael Volný. Laser desorption-ionization of lipid transfers: tissue mass spec-

trometry imaging without maldi matrix. *Analytical chemistry*, 82(12):4994–4997, 2010.

[200] Axel Walch, Sandra Rauser, Sören-Oliver Deininger, and Heinz Höfler. Maldi imaging mass spectrometry for direct tissue analysis: a new frontier for molecular histology. *Histochemistry and cell biology*, 130(3):421–434, 2008.

[201] John G Swales, James W Tucker, Nicole Strittmatter, Anna Nilsson, Diego Cobice, Malcolm R Clench, C Logan Mackay, Per E Andren, Zoltán Takáts, Peter JH Webborn, et al. Mass spectrometry imaging of cassette-dosed drugs for higher throughput pharmacokinetic and biodistribution analysis. *Analytical chemistry*, 86(16):8473–8480, 2014.

[202] Dušan Veličković, David Ropartz, Fabienne Guillon, Luc Saulnier, and Hélène Rogniaux. New insights into the structural and spatial variability of cell-wall polysaccharides during wheat grain development, as revealed through maldi mass spectrometry imaging. *Journal of experimental botany*, 65(8):2079–2091, 2014.

[203] AF Maarten Altelaar, Jan van Minnen, Connie R Jiménez, Ron MA Heeren, and Sander R Piersma. Direct molecular imaging of lymnaea stagnalis nervous tissue at subcellular spatial resolution by mass spectrometry. *Analytical chemistry*, 77(3):735–741, 2005.

[204] Pierre Chaurand, Sophie Fouchécourt, Beverly B DaGue, Baogang J Xu, Michelle L Reyzer, Marie-Claire Orgebin-Crist, and Richard M Caprioli. Profiling and imaging proteins in the mouse epididymis by imaging mass spectrometry. *Proteomics*, 3(11):2221–2239, 2003.

[205] 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.

[206] Michaeleen L Pacholski, Donald M Cannon, Andrew G Ewing, and Nicholas Winograd. Static time-of-flight secondary ion mass spectrometry imaging of freeze-fractured, frozen-hydrated biological membranes. *Rapid communications in mass spectrometry*, 12(18):1232–1235, 1998.

[207] Erin H Seeley and Richard M Caprioli. Molecular imaging of proteins in tissues by mass spectrometry. *Proceedings of the National Academy of Sciences*, 105(47):18126–18131, 2008.

[208] Diogo N de Oliveira, Mônica S Ferreira, and Rodrigo R Catharino. Rapid and simultaneous in situ assessment of aflatoxins and stilbenes using silica plate imprinting mass spectrometry imaging. *PloS one*, 9(3):e90901, 2014.

[209] 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.

[210] Markus Stoeckli, Pierre Chaurand, Dennis E Hallahan, and Richard M Caprioli. Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. *Nature medicine*, 7(4):493–496, 2001.

[211] Yanfeng Chen, Jeremy Allegood, Ying Liu, Elaine Wang, Begoña Cachón-González, Timothy M Cox, Alfred H Merrill, and M Cameron Sullards. Imaging maldi mass spectrometry using an oscillating capillary nebulizer matrix coating system and its application to analysis of lipids in brain from a mouse model of tay-sachs/sandhoff disease. *Analytical chemistry*, 80(8):2780–2788, 2008.

[212] Helene Meistermann, Jeremy L Norris, Hans-Rudolf Aerni, Dale S Cornett, Arno Friedlein, Annette R Erskine, Angelique Augustin, Maria Cristina De Vera Mudry,

Stefan Ruepp, Laura Suter, Hanno Langen, Richard M Caprioli, and Axel Ducret. Biomarker discovery by imaging mass spectrometry transthyretin is a biomarker for gentamicin-induced nephrotoxicity in rat. *Molecular & Cellular Proteomics*, 5(10):1876–1886, 2006.

[213] Yunsheng Hsieh, Roger Casale, Elaine Fukuda, Jiwen Chen, Ian Knemeyer, Julia Wingate, Richard Morrison, and Walter Korfmacher. Matrix-assisted laser desorption/ionization imaging mass spectrometry for direct measurement of clozapine in rat brain tissue. *Rapid Communications in Mass Spectrometry*, 20(6):965–972, 2006.

[214] Rosalind Wolstenholme, Robert Bradshaw, Malcolm R Clench, and Simona Francese. Study of latent fingermarks by matrix-assisted laser desorption/ionisation mass spectrometry imaging of endogenous lipids. *Rapid communications in mass spectrometry*, 23(19):3031–3039, 2009.

[215] Yue Li, Bindesh Shrestha, and Akos Vertes. Atmospheric pressure molecular imaging by infrared maldi mass spectrometry. *Analytical Chemistry*, 79(2):523–532, 2007.

[216] Yue Li, Bindesh Shrestha, and Akos Vertes. Atmospheric pressure infrared maldi imaging mass spectrometry for plant metabolomics. *Analytical Chemistry*, 80(2):407–420, 2008.

[217] Pierre Chaurand, Sarah A Schwartz, and Richard M Caprioli. Assessing protein patterns in disease using imaging mass spectrometry. *Journal of proteome research*, 3(2):245–252, 2004.

[218] Guillaume Robichaud, Jeremy A Barry, and David C Muddiman. Ir-maldesi mass spectrometry imaging of biological tissue sections using ice as a matrix. *Journal of The American Society for Mass Spectrometry*, 25(3):319–328, 2014.

[219] Peter Sjövall, Tanja M Greve, Susanne K Clausen, Kristian Moller, Stefan Eirefelt, Björn Johansson, and Kim T Nielsen. Imaging of distribution of topically applied

drug molecules in mouse skin by combination of time-of-flight secondary ion mass spectrometry and scanning electron microscopy. *Analytical chemistry*, 86(7):3443–3452, 2014.

[220] Young-Pil Kim, Eunkeu Oh, Mi-Young Hong, Dohoon Lee, Min-Kyu Han, Hyun Kyong Shon, Dae Won Moon, Hak-Sung Kim, and Tae Geol Lee. Gold nanoparticle-enhanced secondary ion mass spectrometry imaging of peptides on self-assembled monolayers. *Analytical chemistry*, 78(6):1913–1920, 2006.

[221] Peter Sjövall, Jukka Lausmaa, Håkan Nygren, Lennart Carlsson, and Per Malmberg. Imaging of membrane lipids in single cells by imprint-imaging time-of-flight secondary ion mass spectrometry. *Analytical chemistry*, 75(14):3429–3434, 2003.

[222] Rana NS Sodhi. Time-of-flight secondary ion mass spectrometry (tof-sims):versatility in chemical and imaging surface analysis. *Analyst*, 129(6):483–487, 2004.

[223] Liam A McDonnell, Sander R Piersma, AF Altelaar, Todd H Mize, Stefan L Luxembourg, Peter DEM Verhaert, Jan van Minnen, and Ron Heeren. Subcellular imaging mass spectrometry of brain tissue. *Journal of mass spectrometry*, 40(2):160–168, 2005.

[224] Peter Nemes, Alexis A Barton, Yue Li, and Akos Vertes. Ambient molecular imaging and depth profiling of live tissue by infrared laser ablation electrospray ionization mass spectrometry. *Analytical Chemistry*, 80(12):4575–4582, 2008.

[225] Peter Nemes, Alexis A Barton, and Akos Vertes. Three-dimensional imaging of metabolites in tissues under ambient conditions by laser ablation electrospray ionization mass spectrometry. *Analytical Chemistry*, 81(16):6668–6675, 2009.

[226] Peter Nemes, Amina S Woods, and Akos Vertes. Simultaneous imaging of small metabolites and lipids in rat brain tissues at atmospheric pressure by laser ablation

electrospray ionization mass spectrometry. *Analytical chemistry*, 82(3):982–988, 2010.

[227] 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.

[228] Alain Brunelle and Olivier Laprévote. Lipid imaging with cluster time-of-flight secondary ion mass spectrometry. *Analytical and bioanalytical chemistry*, 393(1):31–35, 2009.

[229] Nora Tahallah, Alain Brunelle, Sabine De La Porte, and Olivier Laprévote. Lipid mapping in human dystrophic muscle by cluster-time-of-flight secondary ion mass spectrometry imaging. *Journal of lipid research*, 49(2):438–454, 2008.

[230] David Touboul, Felix Kollmer, Ewald Niehuis, Alain Brunelle, and Olivier Laprévote. Improvement of biological time-of-flight-secondary ion mass spectrometry imaging with a bismuth cluster ion source. *Journal of the American Society for Mass Spectrometry*, 16(10):1608–1618, 2005.

[231] David Touboul, Alain Brunelle, Frédéric Halgand, Sabine De La Porte, and Olivier Laprévote. Lipid imaging by gold cluster time-of-flight secondary ion mass spectrometry: application to duchenne muscular dystrophy. *Journal of lipid research*, 46(7):1388–1395, 2005.

[232] 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.

[233] Marie-Claude Djidja, Emmanuelle Claude, Marten F Snel, Simona Francese, Peter Scriven, Vikki Carolan, and Malcolm R Clench. Novel molecular tumour classification using maldi–mass spectrometry imaging of tissue micro-array. *Analytical and bioanalytical chemistry*, 397(2):587–601, 2010.

[234] 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.

[235] N Abbassi-Ghadi, K Veselkov, S Kumar, J Huang, E Jones, N Strittmatter, H Kudo, R Goldin, Z Takáts, and GB Hanna. Discrimination of lymph node metastases using desorption electrospray ionisation-mass spectrometry imaging. *Chemical Communications*, 50(28):3661–3664, 2014.

[236] 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.

[237] Stephanie S DeKeyser, Kimberly K Kutz-Naber, Joshua J Schmidt, Gregory A Barrett-Wilt, and Lingjun Li. Imaging mass spectrometry of neuropeptides in decapod crustacean neuronal tissues. *Journal of proteome research*, 6(5):1782–1791, 2007.

[238] Stormy L Koeniger, Nari Talaty, Yanping Luo, Damien Ready, Martin Voorbach, Terese Seifert, Steve Cepa, Jane A Fagerland, Jennifer Bouska, Wayne Buck, Robert W Johnson1, and Stephen Spanton. A quantitation method for mass spectrometry imaging. *Rapid Communications in Mass Spectrometry*, 25(4):503–510, 2011.

[239] Peter DEM Verhaert, Martijn WH Pinkse, Kerstin Strupat, and Maria C Prieto Conaway. Imaging of similar mass neuropeptides in neuronal tissue by enhanced

resolution maldi ms with an ion trap–orbitraptm hybrid instrument. In *Mass spectrometry imaging*, pages 433–449. Springer, 2010.

[240] 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.

[241] Stéphanie Marie Boudon, Grégory Morandi, Brendan Prideaux, Dieter Staab, Ursula Junker, Alex Odermatt, Markus Stoeckli, and Daniel Bauer. Evaluation of sparfloxacin distribution by mass spectrometry imaging in a phototoxicity model. *Journal of The American Society for Mass Spectrometry*, 25(10):1803–1809, 2014.

[242] Marie-Claude Djidja, Joan Chang, Andreas Hadjiprocopis, Fabian Schmich, John Sinclair, Martina Mrǒnik, Erwin M Schoof, Holly E Barker, Rune Linding, Claus Jørgensen, and Janine T Erler. Identification of hypoxia-regulated proteins using maldi-mass spectrometry imaging combined with quantitative proteomics. *Journal of proteome research*, 13(5):2297–2313, 2014.

[243] Hanane Kadar, Gael Le Douaron, Majid Amar, Laurent Ferrié, Bruno Figadère, David Touboul, Alain Brunelle, and Rita Raisman-Vozari. Maldi mass spectrometry imaging of 1-methyl-4-phenylpyridinium (mpp+) in mouse brain. *Neurotoxicity research*, 25(1):135–145, 2014.

[244] Filip Kaftan, Vladimír Vrkoslav, Philipp Kynast, Purva Kulkarni, Sebastian Böcker, Josef Cvačka, Markus Knaden, and Aleš Svatoš. Mass spectrometry imaging of surface lipids on intact drosophila melanogaster flies. *Journal of Mass Spectrometry*, 49(3):223–232, 2014.

[245] Liang Qiao, Elena Tobolkina, Andreas Lesch, Alexandra Bondarenko, Xiaoqin Zhong, Baohong Liu, Horst Pick, Horst Vogel, and Hubert H Girault. Electrostatic spray ionization mass spectrometry imaging. *Analytical chemistry*, 86(4):2033–2041, 2014.

[246] J Sa Becker, MV Zoriy, C Pickhardt, N Palomero-Gallagher, and K Zilles. Imaging of copper, zinc, and other elements in thin section of human brain samples (hippocampus) by laser ablation inductively coupled plasma mass spectrometry. *Analytical chemistry*, 77(10):3208–3216, 2005.

[247] J Sabine Becker, Miroslav Zoriy, J Susanne Becker, Justina Dobrowolska, and Andreas Matusch. Laser ablation inductively coupled plasma mass spectrometry (la-icp-ms) in elemental imaging of biological tissues and in proteomics. *Journal of Analytical Atomic Spectrometry*, 22(7):736–744, 2007.

[248] J Sa Becker, A Matusch, C Depboylu, J Dobrowolska, and MV Zoriy. Quantitative imaging of selenium, copper, and zinc in thin sections of biological tissues (slugs-genus arion) measured by laser ablation inductively coupled plasma mass spectrometry. *Analytical chemistry*, 79(16):6074–6080, 2007.

[249] Sangwon Cha and Edward S Yeung. Colloidal graphite-assisted laser desorption/ionization mass spectrometry and ms n of small molecules. 1. imaging of cerebrosides directly from rat brain tissue. *Analytical chemistry*, 79(6):2373–2385, 2007.

[250] Asiri S Galhena, Glenn A Harris, Leonard Nyadong, Kermit K Murray, and Facundo M Fernandez. Small molecule ambient mass spectrometry imaging by infrared laser ablation metastable-induced chemical ionization. *Analytical chemistry*, 82(6):2178–2181, 2010.

[251] Michael E Kurczy, Paul D Piehowski, Craig T Van Bell, Michael L Heien, Nicolas Winograd, and Andrew G Ewing. Mass spectrometry imaging of mating tetrahymena show that changes in cell morphology regulate lipid domain formation. *Proceedings of the National Academy of Sciences*, 107(7):2751–2756, 2010.

[252] 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.

[253] Rachel V Bennett, Chaminda M Gamage, Asiri S Galhena, and Facundo M Fernández. Contrast-enhanced differential mobility-desorption electrospray ionization-mass spectrometry imaging of biological tissues. *Analytical chemistry*, 86(8):3756–3763, 2014.

[254] 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.

[255] Delphine Debois, Virginie Bertrand, Loic Quinton, Marie-Claire De Pauw-Gillet, and Edwin De Pauw. Maldi-in source decay applied to mass spectrometry imaging: a new tool for protein identification. *Analytical chemistry*, 82(10):4036–4045, 2010.

[256] Julien Franck, Karim Arafah, Alan Barnes, Maxence Wisztorski, Michel Salzet, and Isabelle Fournier. Improving tissue preparation for matrix-assisted laser desorption ionization mass spectrometry imaging. part 1: using microspotting. *Analytical chemistry*, 81(19):8193–8202, 2009.

[257] Julien Franck, Rémi Longuespée, Maxence Wisztorski, Alexandra Van Remoortere, Rene Van Zeijl, Andre Deelder, Michel Salzet, Liam McDonnell, and Isabelle Fournier. Maldi mass spectrometry imaging of proteins exceeding 30,000 daltons. *Med Sci Monit*, 16(9):299, 2010.

[258] Stephanie Kaspar, Manuela Peukert, Ales Svatos, Andrea Matros, and Hans-Peter Mock. Maldi-imaging mass spectrometry–an emerging technique in plant biology. *Proteomics*, 11(9):1840–1850, 2011.

[259] Karim Arafah, Rémi Longuespée, Annie Desmons, Olivier Kerdraon, Isabelle Fournier, and Michel Salzet. Lipidomics for clinical diagnosis: dye-assisted laser

desorption/ionization (daldi) method for lipids detection in maldi mass spectrometry imaging. *Omics: a journal of integrative biology*, 18(8):487–498, 2014.

[260] Loïc Becker, Vincent Carré, Anne Poutaraud, Didier Merdinoglu, and Patrick Chaimbault. Maldi mass spectrometry imaging for the simultaneous location of resveratrol, pterostilbene and viniferins on grapevine leaves. *Molecules*, 19(7):10587–10600, 2014.

[261] Stefan Vergeiner, Lukas Schafferer, Hubertus Haas, and Thomas Müller. Improved maldi-tof microbial mass spectrometry imaging by application of a dispersed solid matrix. *Journal of The American Society for Mass Spectrometry*, 25(8):1498–1501, 2014.

[262] Hui Ye, Rakesh Mandal, Adam Catherman, Paul M Thomas, Neil L Kelleher, Chrysanthy Ikonomidou, and Lingjun Li. Top-down proteomics with mass spectrometry imaging: A pilot study towards discovery of biomarkers for neurodevelopmental disorders. *PloS one*, 9(4):e92831, 2014.

[263] Robert W Hutchinson, Alan G Cox, Cameron W McLeod, Peter S Marshall, Alex Harper, Emma L Dawson, and David R Howlett. Imaging and spatial distribution of β-amyloid peptide and metal ions in alzheimers plaques by laser ablation–inductively coupled plasma–mass spectrometry. *Analytical biochemistry*, 346(2):225–233, 2005.

[264] Camilla Ricci, Leonard Nyadong, Facundo M Fernandez, Paul N Newton, and Sergei G Kazarian. Combined fourier-transform infrared imaging and desorption electrospray-ionization linear ion-trap mass spectrometry for analysis of counterfeit antimalarial tablets. *Analytical and bioanalytical chemistry*, 387(2):551–559, 2007.

[265] Edward H Adelson. Perceptual organization and the judgment of brightness. *Science-AAAS-Weekly Paper Edition-including Guide to Scientific Information*, 262(5142):2042–2044, 1993.

[266] Tom Kimpe and Tom Tuytschaever. Increasing the number of gray shades in medical display systemshow much is enough? *Journal of digital imaging*, 20(4):422–432, 2007.

[267] MP Simunovic. Colour vision deficiency. *Eye*, 24(5):747–755, 2009.

[268] Gabriele Jordan, Samir S Deeb, Jenny M Bosten, and JD Mollon. The dimensionality of color vision in carriers of anomalous trichromacy. *Journal of vision*, 10(8):1–19, 2010.

[269] M Ronnier Luo, Guihua Cui, and B Rigg. The development of the cie 2000 colour-difference formula: Ciede2000. *Color Research & Application*, 26(5):340–350, 2001.

[270] Gaurav Sharma, Wencheng Wu, and Edul N Dalal. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research & Application*, 30(1):21–30, 2005.

[271] Martin Habekost. Which color differencing equation should be used? *International Circular of Graphic Education and Research*, 6:20–33, 2013.

[272] Colin Ware. *Information visualization: perception for design*. Elsevier, 2004.

[273] Steven J Pachuta. Enhancing and automating tof-sims data interpretation using principal component analysis. *Applied surface science*, 231:217–223, 2004.

[274] William S Cleveland and William S Cleveland. A color-caused optical illusion on a statistical graph. *The American Statistician*, 37(2):101–105, 1983.

[275] Cynthia A. Brewer. Colorbrewer 2.0, August 2014.

[276] Cynthia L Harrison-Felix, Gale G Whiteneck, Amitabh Jha, Michael J DeVivo, Flora M Hammond, and Denise M Hart. Mortality over four decades after traumatic brain injury rehabilitation: a retrospective cohort study. *Archives of physical medicine and rehabilitation*, 90(9):1506–1513, 2009.

[277] Joseph A Hankin, Santiago E Farias, Robert M Barkley, Kim Heidenreich, Lauren C Frey, Kei Hamazaki, Hee-Yong Kim, and Robert C Murphy. Maldi mass spectrometric imaging of lipids in rat brain injury models. *Journal of the American Society for Mass Spectrometry*, 22(6):1014–1021, 2011.

[278] Ann Logan, Martin Berry, Ana Maria Gonzalez, Sally A Frautschy, Michael B Sporn, and Andrew Baird. Effects of transforming growth factor β1, on scar production in the injured central nervous system of the rat. *European Journal of Neuroscience*, 6(3):355–363, 1994.

[279] Mathias Wilhelm, Marc Kirchner, Judith AJ Steen, and Hanno Steen. mz5: space- and time-efficient storage of mass spectrometry data sets. *Molecular & Cellular Proteomics*, 11(1):O111–011379, 2012.

[280] Sheetal Lahabar and PJ Narayanan. Singular value decomposition on gpu using cuda. In *Parallel & Distributed Processing, 2009. IPDPS 2009. IEEE International Symposium on*, pages 1–10. IEEE, 2009.

[281] 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.

[282] 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.

[283] 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.

[284] Rory T Steven and Josephine Bunch. Repeat maldi ms imaging of a single tissue section using multiple matrices and tissue washes. *Analytical and bioanalytical chemistry*, 405(14):4719–4728, 2013.

[285] GB Eijkel, B Kükrer Kaletaş, IM Van der Wiel, JM Kros, TM Luider, and RMA Heeren. Correlating maldi and sims imaging mass spectrometric datasets of biological tissue surfaces. *Surface and Interface Analysis*, 41(8):675–685, 2009.

[286] TW Bocklitz, AC Crecelius, C Matthaus, N Tarcea, F Von Eggeling, M Schmitt, US Schubert, and J Popp. Deeper understanding of biological tissue: quantitative correlation of maldi-tof and raman imaging. *Analytical chemistry*, 85(22):10829–10834, 2013.

[287] Jay G Tarolli, Lauren M Jackson, and Nicholas Winograd. Improving secondary ion mass spectrometry image quality with image fusion. *Journal of The American Society for Mass Spectrometry*, 25(12):2154–2162, 2014.

[288] Raf Van de Plas, Junhai Yang, Jeffrey Spraggins, and Richard M Caprioli. Image fusion of mass spectrometry and microscopy: a multimodality paradigm for molecular tissue mapping. *Nature methods*, 12(4):366–372, 2015.

[289] Bram Heijs, Else A Tolner, Judith VMG Bovée, Arn MJM van den Maagdenberg, and Liam A McDonnell. Brain region-specific dynamics of on-tissue protein digestion using maldi mass spectrometry imaging. *Journal of proteome research*, 14(12):5348–5354, 2015.

[290] Benjamin Balluff, Christian K Frese, Stefan K Maier, Cédrik Schöne, Bernhard Kuster, Manfred Schmitt, Michaela Aubele, Heinz Höfler, André M Deelder, Albert JR Heck, Pancras CW Hogendoorn, Johannes Morreau, AF Maarten Altelaar, Axel Walch, and Liam A McDonnell. De novo discovery of phenotypic intratumour heterogeneity using imaging mass spectrometry. *The Journal of pathology*, 235(1):3–13, 2015.

[291] Ricardo J Carreira, Reinald Shyti, Benjamin Balluff, Walid M Abdelmoula, Sandra H van Heiningen, Rene J van Zeijl, Jouke Dijkstra, Michel D Ferrari, Else A Tolner, Liam A McDonnell, and Arn M. J. M. Maagdenberg. Large-scale mass

spectrometry imaging investigation of consequences of cortical spreading depression in a transgenic mouse model of migraine. *Journal of the American Society for Mass Spectrometry*, 26(6):853–861, 2015.