Computational Methods for the Analysis of Mass Spectrometry Imaging Data

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat der Fakultät für Mathematik und Informatik

der Friedrich-Schiller-Universität Jena

von M.Sc. Purva Kulkarni

geboren am 10. Mai 1987 **in** Indore, Indien

Gutachter:

1. Prof. Dr. Sebastian Böcker, Friedrich-Schiller-Universität Jena
2. Dr. Aleš Svatoš, Max Planck Institute for Chemical Ecology Jena
3. Prof. Dr. Manja Marz, Friedrich-Schiller-Universität Jena

Tag der öffentlichen Verteidigung: 17. September 2018

Gedruckt auf alterungsbeständigem Papier nach DIN-ISO 9706

Abstract

Several thousand biochemical reactions occur in a living system. The low molecular weight organic compounds that take part in, or are formed by, these reactions are called *metabolites* and their study is known as *metabolomics*. System-wide comprehensive characterization of metabolites in an organism is crucial for understanding cellular function.

Metabolites cover a wide range of compound classes and exhibit large structural diversity. A vast majority of these metabolites are still unknown and identifying them is a bottleneck. This physicochemical diversity of metabolites necessitates the application of different complementary analytical methods for their detection and identification. Mass spectrometry (MS) is one such method that provides unmatched capabilities for the detection and identification of analytes. It is being extensively applied for metabolomics in the recent years. A powerful enhancement to MS-based detection is the addition of spatial information to the chemical data; an approach called mass spectrometry imaging (MSI). This label-free, non-targeted technique generates a series of localized mass spectra from discrete positions on the imaged sample, thereby providing comprehensive information on molecular composition as well as spatial distribution in a single experiment. MSI enables two- and three-dimensional overviews of hundreds of molecular species over a wide mass range in complex biological samples.

With constant advancements in spectral and spatial resolution and also depending on the biological sample analyzed, a single MSI experiment is capable to generate large amounts of multi-dimensional data. The localization of a specific ion species across the imaged sample can be visualized by creating an ion intensity map. The manual exploration and interpretation of such data, however, is far from simple. To harness the full potential of the acquired data, there is a strong need for robust computational approaches that can extract valuable chemical information captured in a single experiment. In this work, we present two computational methods and a workflow that address three different aspects of MSI data analysis: correction of mass shifts, unsupervised exploration of the data and importance of preprocessing and chemometrics to extract meaningful information from the data.

High mass accuracy of the acquired MS data is of prime importance for confident peak annotation. However, this can be compromised due to several factors, leading to mass errors. In the first part of this thesis, we introduce a new lock mass-free recalibration procedure that enables to significantly reduce these mass shift effects in MSI data. Our method exploits similarities amongst peaklist pairs and takes advantage of the spatial context in three different ways, to perform mass correction in an iterative manner. Evaluation of our method showed considerable reduction in the mass shifts for different datasets. As an extension of this work, we also present a Java-based tool, `MSICorrect`, that implements our recalibration approach and also allows data visualization.

In the next part, an unsupervised approach to rank ion intensity maps based on the abundance of their spatial pattern is presented. In current practice, the predominant approach to analyze MSI datasets is either looking at ion intensity maps to visualize the distribution of specific masses of interest or using data-mining methods. However, this conventional approach can prove extremely challenging and labor-intensive in getting an overview of the measured masses or to explore the distribution of unknown compounds. There is an acute need for methods that allow untargeted and exploratory analysis of MSI data. The method presented in this thesis is our first attempt in this direction. Our method provides a score to every ion intensity map based on the abundance of spatial pattern present in it and then ranks all the maps using it. To know which masses exhibit similar spatial distribution, our method uses spatial-similarity based grouping to provide lists of masses that exhibit similar distribution patterns. Results for ranking and spatial-similarity based grouping show a good agreement with visual observations.

In the last part, we demonstrate the application of a data preprocessing and multivariate analysis pipeline to a real-world biological dataset. An optimal data preprocessing strategy is the key to obtain reliable statistical interpretation of the data. It is critical to judge and select the right preprocessing and multivariate analysis algorithms based on the nature of the data. We demonstrate this by applying the full pipeline to a high-resolution MSI dataset acquired from the leaf surface of Black cottonwood (*Populus trichocarpa*). Application of the pipeline helped in highlighting and visualizing the chemical specificity on the leaf surface.

The computational solutions introduced in this thesis address different issues that are encountered while analyzing MSI data. The main advantages of our presented recalibration method are that mass correction is robust against outliers, is not dependent on the presence of a lock-mass (namely internal or external calibrant) and can be applied to datasets that have a varying degree of mass shift across the imaged sample. Our spatial pattern extraction and image ranking approach allows to quickly explore large MSI datasets, without any expert knowledge. Lastly, the data analysis workflow presented can guide researchers to select appropriate preprocessing and statistical methods for their MSI data. Thus, application of the presented methods will be helpful to improve mass accuracy, in turn enabling reliable compound annotation and they also hold the potential for use in untargeted studies, further aiding in the identification of novel biomarkers.

Zusammenfassung

In einem lebenden System treten tausende biochemische Reaktionen auf. Zum Verständnis der einzelnen zellulären Abläufe ist es entscheidend, die auftretenden Substanzen umfassend und systemweit zu charakterisieren. Die niedermolekularen organischen Verbindungen, die an diesen Prozessen beteiligt sind bzw. dabei entstehen, werden als Metabolite bezeichnet. Das zugehörige Forschungsgebiet heißt Metabolomics.

Die bisher bekannten Metabolite weisen eine große strukturelle Diversität auf, weswegen sie zu einer Vielzahl unterschiedlicher Substanzklassen gehören. Es muss beachtet werden, dass ein Großteil dieser Verbindungen noch immer unbekannt und die strukturelle Charakterisierung sehr schwierig sind. Aus diesem Grund ist die Anwendung von verschiedenen komplementären Analysemethoden zur Detektion und Identifizierung der Metabolite nötig.

Die Massenspektrometrie (MS) weist hierfür unübertroffene Fähigkeiten zur Analyse von niedermolekularen organischen Verbindungen auf und wird mittlerweile routinemäig im Bereich Metabolomics eingesetzt. Eine wichtige Erweiterung zur massenbasierten Detektion ist hierbei die sogenannte bildgebende Massenspektrometrie (MSI), welche neben den chemischen Daten auch Informationen zur räumlichen Verteilung liefert. Diese Non-Target-Technik kommt ohne chemische Markierungen aus und generiert innerhalb eines einzigen Experiments eine Serie von Massespektren für eine diskrete Position auf der zu untersuchenden Probe. MSI ermöglicht einen zwei- und dreidimensionalen Überblick zur räumlichen Verteilung von hunderten molekularen Spezies innerhalb eines großen Massenbereichs in komplexen biologischen Proben. Während eines einzigen MSI-Experiments wird in Abhängigkeit vom Gewebe eine riesige Menge multidimensionaler Daten generiert. Diese vergrößert sich auch durch die ständige Weiterentwicklung in der spektralen und räumlichen Auflösung. Die Lokalisierung der einzelnen Ionenspezies über die gesamte Probe lässt sich mithilfe einer sogenannten Ionenintensitätskarte (ion intensity map) visualisieren. Dennoch sind die manuelle Auswertung und Interpretation dieser Messdaten beraus aufwendig und kompliziert. Um das tatsächliche Potential der Analysen auszuschöpfen, werden robuste Berechnungsansätze benötigt, die die wertvolle chemische Information aus den Daten eines einzigen Experiments extrahieren können.

In der vorliegenden Arbeit präsentieren wir zwei Berechnungsmethoden und einen Workflow, die sich an drei verschiedene Probleme der MSI-Datenauswertung richten. Dazu gehören die Korrektur von Masseverschiebungen (mass shifts), die nicht-überwachte Datenauswertung sowie die essentielle Rolle von Datenvorbehandlung und Chemometrik zur Extraktion von sinnvollen Informationen aus den vorliegenden Daten.

Die hohe Massengenauigkeit der aufgenommenen Daten ist für die sichere Peakzuordnung besonders wichtig. Dennoch kann sie durch verschiedene Faktoren gestört werden, woraus Massefehler resultieren. Wir stellen im ersten Teil der Arbeit eine neue Methode zur Rekalibration vor, die ohne Lock-Masse (Sperrmasse) auskommt und eine signifikante Reduktion von Masseverschiebungseffekten in MSI-Daten ermöglicht. Diese Methode bezieht Ähnlichkeiten zwischen Paaren in den Peaklisten ein und nutzt den räumlichen Kontext auf drei verschiedenen Wegen, um die Massenkorrektur iterativ durchzuführen. Die Anwendung unserer Methode auf verschiedene Datenstze zeigte eine beträchtliche Reduktion von Masseverschiebungen. Weiterhin stellen wir ein Java-basiertes Tool `MSICorrect` vor, das einerseits unseren Rekalibrationsansatz beinhaltet und gleichzeitig Datenvisualisierung erlaubt.

Im zweiten Teil wird ein nicht-überwachter Ansatz vorgestellt, der die Ionenintensitätskarten nach der Häufigkeit ihrer räumlichen Muster ordnet. Gegenwärtig wird die Analyse

von MSI-Daten vorrangig mithilfe von Visualisierung der Massespeziesverteilung anhand der Ionenintensitätskarten oder mittels Methoden zum Data-Mining durchgeführt. Dennoch sind diese herkömmlichen Herangehensweisen sehr herausfordernd und arbeitsintensiv, wenn ein Überblick über die gemessenen Massen gewonnen oder die Verteilung von unbekannten Verbindungen untersucht werden sollen. Daher besteht eine akute Nachfrage an Methoden, die Non-Target- und explorative MSI-Datenanalysen ermöglichen. Die in der vorliegenden Arbeit vorgestellte Methode ist unser erster Vorstoß in dieses Forschungsgebiet. Sie ermittelt basierend auf der Häufigkeit der enthaltenen räumlichen Muster einen Wert für jede Ionenintensitätskarte und ordnet sie anhand dessen anschließend in eine Rangliste ein. Um festzustellen, welche Massen ähnliche Verteilungen aufweisen, nutzt unsere Methode auf räumlicher Ähnlichkeit basierende Gruppierung. So können Listen mit Massen ähnlicher Verteilungsmuster generiert werden. Die Ergebnisse aus diesen Vorgehensweisen zeigen eine gute Übereinstimmung mit visuellen Beobachtungen.

Im letzten Teil der Arbeit demonstrieren wir die Anwendung eines Workflows zur Datenvorbehandlung und zur multivariaten Analyse auf einen realen biologischen Datensatz. Eine optimale Datenvorbehandlung ist der Schlüssel, um eine zuverlässige statistische Interpretation zu ermöglichen. Es ist entscheidend, anhand der tatsächlich vorliegenden Daten zu beurteilen, welche Algorithmen zur Datenvorbehandlung und zur multivariaten Analyse ausgewählt werden sollten. Wir demonstrieren diesen Sachverhalt durch Anwendung des kompletten Workflows auf einen hochaufgelösten MSI-Datensatz einer Laubblattoberfläche der Westlichen Balsampappel (*Populus trichocarpa*). Es konnte gezeigt werden, dass dieser Workflow dabei hilft die chemische Spezifität auf der Laubblattoberfläche hervorzuheben und zu visualisieren.

Die in dieser Arbeit vorgestellten Berechnungsmethoden befassen sich mit verschiedenen Problemen, die bei der Analyse von MSI-Daten auftreten. Die wesentlichen Vorteile unserer Rekalibrierungsmethode bestehen darin, dass die Massekorrektur gegenüber Ausreißern robust und außerdem unabhängig von einer Lock-Masse (externer oder interner Kalibrant) ist sowie auch auf Datensätze angewendet werden kann, die einen variierenden Grad an Masseverschiebungen über die gesamte Probe aufweisen. Unsere räumliche Musterextraktion und unser Ansatz zum Image-Ranking erlauben es, große MSI-Datensätze ohne Expertenwissen zu untersuchen. Weiterhin unterstützt der vorgestellte Workflow zur Datenanalyse Wissenschaftlern darin, angemessene Datenvorbehandlungen und statistische Methoden für die jeweiligen MSI-Daten auszuwählen. Somit kann die zukünftige Anwendung der hier vorgestellten Methoden dazu beitragen, die Massengenauigkeit zu verbessern und gleichzeitig die zuverlässige Peakzuordnung zu ermöglichen. Weiterhin bergen unsere Methoden ein großes Potential für die Anwendung bei Non-Target-Analysen, insbesondere für die Identifizierung bisher unbekannter Biomarker.

To my grandparents

Malini and Madhav Sarwate

whose selfless mentoring and inspiration put me on the path
to where I am today

Acknowledgments

It takes a village to raise a PhD student!
It was Thursday, January 17th, 2013, when I arrived in snow-cladded Jena to start my doctoral studies at the Lehrstuhl für Bioinformatik, Friedrich Schiller University. Looking back over all these years, I can certainly say that I learned a lot; experienced a new culture that widened my perspective, gained scientific knowledge as well as worked and made friends with some amazing people around me, who still continue to inspire and support me in my journey ahead. This thesis appears in its current form due to the assistance and valuable guidance of several people, both on academic and personal front. I would therefore like to offer my sincere thanks to all of them.

I would primarily like to thank my supervisor Prof. Dr. Sebastian Böcker for accepting me as a PhD student within his group and for all his invaluable mentoring, insightful discussions and critical comments, during my PhD pursuit. I would also like to offer my sincere thanks to him for providing me manifold opportunities for interdisciplinary collaborations as well as financial aid to attend various workshops and conferences. I would also like to express my deepest gratitude to my second supervisor Dr. Aleš Svatoš for his intellectual input, continuous encouragement throughout the course of my PhD as well as for a few months of financial aid during the final stages of my PhD work.

Thanks also to Dr. Filip Kaftan (MPI-CE, Jena) for a successful collaboration and for providing me multiple datasets from time to time. Further thanks goes to former diploma student Phillip Kynast, who laid the initial foundation for the recalibration algorithm that I have worked on during my PhD. I would also like to thank my other collaborators - Prof. Wilhelm Boland and Mina Dost (MPI-CE Jena, Germany) for sharing their scientific vision and considering me capable to design the analysis pipeline for their imaging data.

I have also greatly benefited from many fruitful and inspiring discussions with my present and former colleagues at the Lehrstuhl für Bioinformatik, specially Franziska Hufsky, Martin Engler, Markus Fleischauer, Marcus Ludwig, Kai Dührkop, Tim White and Sascha Winter. I would especially like to thank my office mates Markus Fleischauer and Marcus Ludwig for a fun and wonderful camaraderie since the time I started my PhD. I am also thankful to them for being ever-ready to help me with technical issues, non-scientific issues and with German translations. A special thanks also goes to Kathrin Schowtka for taking care of all the bureaucratic and administrative formalities.

Moreover, I am very grateful for the generous financial support received from the International Max Planck Research School, the Jena School for Microbial Ecology, the Max Planck Society and the University of Jena. I would also like to gratefully acknowledge the training sessions provided by the International Max Planck Research School and the Jena Graduate Academy that helped me gain additional skill sets.

I warmly appreciate the time and effort of Kumar Saurabh Singh, Anne-Christin Warskulat, Franziska Hufsky, Tim White and Riya Christina Menezes for proofreading parts of this thesis and giving advices to improve legibility. A special thanks to Christoph Zimmer and Anne-Christin Warskulat for helping me translate the abstract to German.

I am immensely grateful to my friends here in Jena: Karen, Anne-Christin, Jelena, Verena, Garima, Dinesh, Rajarajan, Arpita, Kamlesh, Priya, Rakesh, Govind and Sneha, for keeping me sane and for being there to listen and share all the highs and lows of this journey. A big thank you for all the supportive hugs during the trying times.

Finally, I owe my deepest gratitude to my grandparents, my mentors, without whom this would have not started in the first place. My parents Jyoti and Bhalchandra Kulkarni deserve a particular note of thanks for their endless support and love, for never questioning me no matter how implausible my decisions seemed and for strongly believing in my dreams and letting me achieve them. I feel immensely lucky to have such a backing! I am also very thankful to my in-laws as well as my extended family for being very supportive.

Last, but not the least, I'm truly grateful to my *partner in science* and in life, Saurabh. For just everything. We made it possible despite the distance!

Contents

Preface xiii

1 Introduction 1
1.1 Emerging technologies to visualize and explore biological systems 1
1.2 Contribution and scope of this thesis 2

2 An overview of analytical and technological concepts 5
2.1 Biomolecules 5
2.2 Mass spectrometry 6
2.2.1 Sample inlet 7
2.2.2 Ion source 8
2.2.3 Mass analyzer 8
2.2.4 Detector 11
2.2.5 Mass spectrum 11
2.3 Mass spectrometry imaging 11
2.3.1 Techniques 12
2.3.2 Data features, computational aspects and visualization 16
2.3.3 Applications of mass spectrometry imaging 17

3 Computational processing and analysis of MSI data 19
3.1 Overview of the data analysis workflow 19
3.2 Data import and conversion 20
3.3 Spectral data preprocessing 20
3.3.1 Smoothing 21
3.3.2 Baseline correction 22
3.3.3 Normalization 23
3.3.4 Peak picking 25
3.3.5 Spectral recalibration 26
3.4 Dimensionality reduction and unsupervised data mining 26
3.4.1 Component analysis 27
3.4.2 Cluster analysis 29
3.5 Biological interpretation and further studies 32
3.5.1 Molecular identification 32
3.5.2 Supervised classification 33
3.5.3 Combination with other imaging modalities 33

4 Recalibration of mass spectrometry imaging data 35
4.1 Mass shifts and their correction 35
4.2 Previous contributions 36
4.3 Lock mass-free recalibration 37
4.3.1 Spectral similarity measure 38
4.3.2 Spectral ordering approaches 38
4.3.2.1 Minimum spanning tree ordering 38
4.3.2.2 Topological greedy ordering 39

4.3.2.3 Crystal growth ordering 40
4.3.3 Recalibrating a pair of peaklists 41
4.3.4 Generation of consensus spectrum 42
4.3.5 Final correction 43
4.4 Evaluation of mass shift correction 43
4.4.1 Experimental datasets 43
4.4.2 Data acquisition and processing 43
4.4.3 Recalibration accuracy and performance 44
4.4.4 Evaluation of running times 47

5 MSICorrect: Mass spectrometry imaging data recalibration tool 49
5.1 Graphical user interface 49
5.2 Data import 51
5.3 Lock mass-free recalibration 51
5.4 Visualization of recalibration results 52

6 Towards an automated method to characterize biologically relevant spatial patterns in MSI data 53
6.1 Image processing and pattern extraction 53
6.2 Previous contributions 54
6.3 Our pattern extraction and image ranking approach 54
6.3.1 Image Content score for an ion intensity map 55
6.3.2 Selection of an image threshold 56
6.3.3 Spatial similarity-based grouping of ion intensity maps 57
6.4 Evaluation of our pattern extraction and image-ranking approach 58
6.4.1 Experimental datasets 58
6.4.2 Method performance 59
6.4.3 Evaluation of running times 62

7 Preprocessing and multivariate analysis of TOF-SIMS imaging data 65
7.1 Introduction 65
7.2 TOF-SIMS imaging 66
7.3 Data processing and analysis 66
7.3.1 Data preprocessing 68
7.3.2 Multivariate analysis and clustering 68
7.4 Results 69

8 Conclusion 77

A Supplementary information for Chapter 4 97

B Supplementary information for Chapter 6 105

C Supplementary information for Chapter 7 109

Preface

This dissertation covers large parts of my research in method development and analysis of mass spectrometry imaging data, for the last four years. During this time, I was working at the Bioinformatics Group of Professor Sebastian Böcker at the Friedrich-Schiller Universität, Jena and for a short period at the Mass Spectrometry Group of Dr. Aleš Svatoš at the Max Planck for Chemical Ecology, Jena. My research was financed by a scholarship from the International Max Planck Research School, Jena and later by the Jena School of Microbial Ecology, the Max Planck Society and university's basic funding.

Most of the results presented in this work are published [1–3] and have been achieved in cooperation with my supervisor Sebastian Böcker and our main collaborators Aleš Svatoš, Filip Kaftan, Wilhelm Boland, Mina Dost and former diploma student Philipp Kynast. I also participated in developing a method to read and create ion intensity maps from the data acquired using laser ablation electrospray ionization mass spectrometry imaging of non-flat biological samples [4]. This work was performed in collaboration with Aleš Svatoš and Benjamin Bartels. I was also involved in analyzing bacterial metabolomics data generated from liquid extraction surface analysis combined with high resolution mass spectrometry. This work was performed in collaboration with Aleš Svatoš, Paolina Garbeva and Riya Christina Menezes.

This thesis consists of seven chapters. The main results of this thesis are presented in Chapters 4,5, 6 and 7.

Chapter 4 presents a new recalibration method for mass spectrometry imaging data and has been published [1]. Sebastian Böcker had the basic idea for the method and I participated in developing it further. The method encompasses three mass spectral ordering approaches. One of the ordering approaches was initially developed by Philipp Kynast for his diploma thesis [5], but later was improved and re-implemented by me, along with developing and implementing the other two approaches. The approach for recalibrating two mass spectra is based on the Maximum Line-Pair stabbing algorithm developed by Böcker and Mäkinen [6]. Evaluation of the recalibration method was performed by me on multiple imaging datasets provided by Aleš Svatoš and Filip Kaftan. The recalibration method was presented by me at the *3rd Mass Spectrometry Imaging Conference* (OurCon 2015).

Chapter 5 presents a tool, named `MSICorrect`, for visualization and recalibration of mass spectrometry imaging data. For recalibration, the tool implements the method described in Chapter 4. Implementation of the graphical user interface and different data visualization functions was performed by me. Furthermore, I also performed the evaluation of `MSICorrect` using the datasets provided by Aleš Svatoš and Filip Kaftan.

Chapter 6 presents a simplistic method to computationally identify spatial patterns visible in ion intensity maps generated using mass spectrometry imaging data in a unsupervised manner. I designed the basic concept of the approach with valuable inputs from Sebastian Böcker and my colleague Tim White, which I also implemented. The complete evaluation of the approach was performed by me using publicly available mass spectrometry imaging datasets.

In Chapter 7, a use-case of a typical mass spectrometry imaging data analysis work flow [3] is described. This work was performed in collaboration with Aleš Svatoš, Wilhelm Boland and Mina Dost, wherein the data was provided by Mina Dost. I designed the complete analysis pipeline, which utilizes existing data processing and chemometric approaches. I also analyzed the acquired data using the designed pipeline to achieve biological conclusions.

For the remainder of this thesis, I will use “we” as the first person pronoun, as it is common in scientific literature. This may be interpreted as “the reader and I” or as “my collaborators and I”, whichever suits best in the situation.

Publications resulting from this work

Large parts of the research presented in this thesis can also be found in the following publications:

Published/Accepted

- **P. Kulkarni**, M. Dost, O.D. Bulut, A. Welle; S. Böcker, W. Boland, A. Svatoš. Secondary ion mass spectrometry imaging and multivariate data analysis reveal co-aggregation patterns of *Populus trichocarpa* leaf surface compounds on a micrometer scale *The Plant Journal*, *Accepted for publication* Oct 2017.
- **P. Kulkarni**, F. Kaftan, P. Kynast, A. Svatoš, S. Böcker. Correcting mass shifts: A lock-mass-free recalibration procedure for mass spectrometry imaging data *Anal Bioanal Chem*, 405, 7603-7613, 2015.
- F. Kaftan, V. Vrkoslav, P. Kynast, **P. Kulkarni**, S. Böcker, J. Cvačka, M. Knaden, and A. Svatoš. Mass Spectrometry Imaging of Surface Lipids on Intact *Drosophila melanogaster* flies *J Mass Spectrom*, 49, 223-232, 2014.

Submitted

- F. Kaftan, **P. Kulkarni**, M. Knaden, S. Böcker, S, A. Svatoš. *Drosophila melanogaster* chemical ecology revisited: 2-D distribution maps of sex pheromones on whole virgin and copulated flies by mass spectrometry imaging *Submitted to BMC Biology*, 2017.

Other publications not part of this thesis

- B. Bartels, **P. Kulkarni**, S. Böcker, A. Svatoš. Mapping metabolites from rough terrain: laser ablation electrospray ionization on non-flat samples *RSC Advances*, 7, 9045-9050, 2016.

Chapter 1

Introduction

> *Science is to see what everyone else has seen, but think what no one else has thought.*
>
> – Albert Szent-Gyorgyi

1.1 Emerging technologies to visualize and explore biological systems

In the early 1950s, when Watson and Crick identified the structure of DNA, molecular-level understanding of biology started emerging. Genome-based studies became the main focus of life sciences post this. Exactly 50 years after the discovery of DNA, with the completion of the Human Genome Project [7] in April 2003, it however became clear that complex cellular processes are regulated more on other levels than the DNA sequence alone. This realization triggered the rapid growth of numerous fields in molecular biology that together are now termed as "*omics*"technologies. This omics cascade [8] encompasses within itself: genomics, focusing on the systematic study of an organism's genome; transcriptomics which studies the global change of gene expression at mRNA level [9]; proteomics that involves systematic identification of all proteins expressed in a cell or tissue [10] and lastly metabolomics that involves characterization of the metabolome of an organism.

Metabolome is the final downstream product of the genome and is considered closest to the phenotype [11]. Analogous to the terms 'genome', 'transcriptome' and 'proteome', the metabolome refers to the complete set of small molecules, termed as *metabolites*, produced by a cell, as a result of its metabolism. These include, but are not limited to lipids, sugars, metabolic intermediates, products of biochemical reactions as well as building blocks including proteins, nucleic acids and cell membranes [12]. The metabolome is a highly complex and dynamic system where the concentration of each metabolite depends strictly on changes in the physiological conditions of the cell, induced in response to environmental or developmental stimuli and many such factors. Studying the metabolome can help in providing comprehensive and simultaneous understanding of the changes occurring in a system at the cellular level, in turn aiding in the characterization of attractive candidates to understand disease phenotypes [13]. However, the heterogeneity of the metabolome of an organism is much more complex, when compared to combinations of four nucleotide bases in the DNA sequence and 20 amino acids in a protein. This complexity is further enhanced when a wide variety of small molecules are produced through metabolism, having diverse atomic configurations.

With inception of the omics era, there has been an ever increasing demand to identify and characterize variation within biological systems. This has led to the advent of several high-throughput screening techniques with essential attributes like high sensitivity, throughput, robustness, linear range of quantification and cost efficiency. Genomics and transcriptomics studies routinely use well-established microarray and next-generation sequencing technologies whereas core proteomics technologies include mass spectrometry(MS)-based methods to define the protein sequence, study protein:protein interactions and post-translational modifications [14]. As for metabolomics, owing to its complexity, it is essential to use strategies that have a wider coverage in terms of the type and number of metabolites analyzed. MS-based analysis following chromatographic separation and Nuclear Magnetic Resonance (NMR) spectroscopy have played a central role in metabolomics [15], owing to their high sensitivity and resolution, enabling fast identification of a wide range of species. Although NMR allows absolute quantification and precise structural determination in a non-destructive manner, it has shown to have low sensitivity when compared to MS-based techniques [16]. This has allowed MS-based approaches to emerge as the foremost technology of the time, for metabolomics [17].

MS-based metabolomics can provide highly valuable information for a wide range of metabolites, aiding in novel compound discovery to clinical application. However, one needs to remember that in order to understand a biological system in exquisite detail, it is also important to study the molecular entities in their spatial context, owing to their dynamic nature as well as high temporal and spatial variability. And, as the old adage goes; "a picture is worth a thousand words", is no less true here. In regards to metabolomics, a picture can act as an apt means to mirror the intrinsic complexity and convey detailed and immediate information about the distribution and localization of metabolites on a biological sample. Imaging techniques like fluorescence microscopy, atomic force microscopy, positron tomography, infrared and Raman imaging have been used as fundamental tools in molecular imaging to investigate molecular distributions both *in vivo* and *in vitro*. These techniques are capable of generating images with spatial resolution sometimes approaching the atomic scale. However, most of these techniques require analyte pre-selection, labeling or specific markers and are not suitable to detect a wide range of compounds, thereby limiting their utility for molecular discovery [18]. Mass spectrometry-based imaging (MSI) techniques offer distinct advantages over these methods.

Application of the MSI technique commenced in 1997, with the pioneering work performed in the laboratory of Professor Richard M. Caprioli. Here they applied MSI to simultaneously study the localization of hundreds of peptides in a biological tissue sample, without prior knowledge and without labeling [19]. The allure of MSI is its ability to collect molecular 'snapshots' of a given biological sample, in the form of mass spectra from every single coordinate position on the sample and superimpose this molecular information onto optical, fluorescence or histological images. Developments in instrumentation have further improved this technique that has led to the current widespread application of MSI in different areas of biological sciences. Since MSI generates molecular images showing the distribution of ion species on the imaged sample, it enables to connect observations at the phenotypic level with specific changes at the chemical level [20].

Routine metabolomics analysis using MS itself generates avalanches of biological data, often up to multiple gigabytes, presenting immense problems in terms of its use. When using MSI, this complexity is further enhanced in terms dimensionality, rendering a direct manual analysis extremely challenging and practically infeasible. This makes the analysis of such data highly dependent on modern bioinformatics approaches.

1.2 Contribution and scope of this thesis

Converting multi-dimensional MSI data into information requires efficient computational methods. As the popularity of MSI is increasing, the need for computational methods tailored to the special needs of the field, is surging as well. Apart from commercial software like SCiLSLab (SCiLS, Bremen, Germany) or the ones bundled with mass spectrometers, open source packages like MALDIquant by Gibb *et al.*, CARDINAL by Bemis *et al.* and many more are also available. These tools mainly incorporate methods for preprocessing, visualization and multivariate statistical analysis so as to identify important features in the MSI data. Since this field is still relatively young, there is substantial scope to develop novel methods that can address additional challenges that arise when working with MSI data. Several groups are developing efficient approaches to aid in data compression [21], spatial segmentation [22], normalization [23] and also improving standard methods like principal component analysis to accommodate the spatial nature of MSI data [24]. These new methods can be integrated within existing preprocessing pipelines to help in precise identification, feature selection and in improving the overall analysis performance.

In this thesis, we focus on one such issue: mass accuracy. To aid in accurate compound annotation, it is essential to have high mass accuracy in MSI data. But, this can be compromised due to several factors leading to mass errors. In this direction, we develop and deploy a computational approach that is helpful in reducing mass shifts in MSI data.

Further, taking our first few steps towards automated feature identification, we present a basic approach that performs unsupervised feature identification in high-dimensional MSI data. We then discuss the application of a data analysis workflow to a real world MSI dataset, highlighting the importance of method selection, based on the nature of the data.

To present the methods and the analysis pipeline, this thesis is structured as follows:

In Chapter 2, we start with introducing the main analytical and technological concepts of MS-based approaches. Chapter 3 covers the computational aspects, data preprocessing workflows and chemometric methods that are commonly employed in analyzing and interpreting MSI datasets.

In Chapter 4, we describe our computational recalibration method, that helps in reducing mass shifts in the data. Our method performs recalibration on the MSI spectra without the use of any reference spectrum or presence of any external or internal calibrant, making the method completely independent of these factors. We then evaluate the performance of our method by applying it to reduce mass shifts present in three MSI datasets and further discuss the performance of our method in terms of amount of mass shift correction and processing time.

In Chapter 5, we present our MSI data recalibration and visualization tool, `MSICorrect`. It is an easy-to-use, Java-based application, which provides an organized interface to visualize MSI data for user-selected mass values in the form of pseudo-colored ion intensity maps. Apart from this, the tool offers the functionality to reduce mass shifts in the data, visualize the amount of mass shift correction performed and allows to easily export the corrected mass spectra. `MSICorrect` implements the recalibration method described in Chapter 4. In this chapter, we describe the tool architecture and provide technical details on the implementation.

In Chapter 6, we present our very first attempt and the principle concept to develop a computational method for characterizing spatial patterns in large MSI datasets, in an unsupervised manner. We present our simplistic approach in a step-wise manner to report significant spatial patterns in ion intensity maps and rank the maps based on a score. Our method further allows grouping mass values based on their similar spatial distribution patterns. We then evaluate our approach using two publicly available MSI datasets.

Chapter 7 demonstrates the application of a standard MSI data analysis workflow to a high-resolution biological MSI dataset. This use-case provides a detailed insight into the critical selection and application of specific data preprocessing and multivariate analysis methods to understand the chemistry of small structures on the adaxial leaf surface of Black cottonwood (*Populus trichocarpa*). The applied approach revealed a set of unique crystal formation patterns on the leaf surface.

Finally, in Chapter 8, we conclude this thesis with a summary of the findings of our research work and some suggested possible future directions of the study.

Chapter 2

An overview of analytical and technological concepts

In this chapter we provide an introduction to the theoretical, analytical and biological aspects which are a prerequisite to understand the work presented in this thesis. We first provide a brief introduction to biomolecules (Section 2.1), since data used and the analysis methods presented in this thesis revolve around this biological unit. We then introduce mass spectrometry (Section 2.2) which is a unique analytical technology with the capability to measure individual molecular species in complex samples. We then go on further to introduce mass spectrometry imaging (Section 2.3). It is a powerful molecular imaging technology, that provides an extension to the existing capabilities of mass spectrometry by making region-specific molecular measurements directly from the biological sample. Along with explaining the basic principles and work flow of these two techniques, we outline the common technical setups and explain the nature of the acquired imaging data (Section 2.3.2). We conclude this chapter by providing a broad overview of the current applications of mass spectrometry imaging (Section 2.3.3).

2.1 Biomolecules

It is important to characterize the complex structure, behavior and diversity found in living forms. To study the functional processes, it is crucial to have an understanding of the biology at a molecular level.

Atoms are the smallest particles that are responsible for the characteristic properties of solids, liquids and gases. Atoms consist of a *nucleus* and *electrons.* The electrons revolve around the nucleus and bear a negative charge. The nucleus is composed of *protons*, that have a positive charge equal in magnitude to the negative charge of an electron and *neutrons*, that have no charge but have the same mass as protons. The total number of protons in an atom of an element indicates the *atomic number* of that element, whereas the *mass number* is the total number of neutrons and protons present in the atom. The number of neutrons determines the *isotope.* Most elements are composed of more than one naturally occurring isotope. These have the same atomic number but a different mass number due to difference in the number of neutrons.

When two or more atoms are held together by chemical bonds, they form a larger entity known as a *molecule.* Molecules are electrically neutral as they contain equal number of protons and electrons. When an atom or a molecule has a lower or higher number of electrons than the number of protons, then such a particle is known as an *ion.* An ion may be negatively or positively charged. The *molecular formula* indicates the total number of atoms of each element in a molecule of the sample.

Molecular mass can be calculated from its molecular formula using three different ways [25, 26]:

1. *Average mass*: It is the sum of weighted average of masses of all the naturally occurring stable isotopes for each element in the compound.

2. *Monoisotopic mass*: It is calculated using the exact mass (the theoretical or true mass) of the most abundant isotope of each element in the compound.
3. *Nominal mass*: It is the sum of integer masses of the most abundant isotope of each element in the compound. Nominal mass and the mass number of an element have the same value.

The unit of mass is *Dalton* (Da), also known earlier as *unified atomic mass* (u). It is defined as one-twelfth the mass of a neutral carbon atom (^{12}C), in its ground state.

Molecules are the building blocks of nature. While the smallest molecule consists of two atoms, the large molecules (also known as macromolecules) are complex and are composed of several thousands of atoms. There are certain molecules that are the building blocks of living organisms and are essential for their survival, growth and maintenance. These are known as *biomolecules.* These include macromolecules like proteins, carbohydrates, lipids and nucleic acids as well as small molecules like primary and secondary metabolites and natural products.

With advancements in technology, different methods have been developed to study this molecular machinery. Based on the research goal, the types of analysis can be broadly classified as follows:

1. *Qualitative analysis*: Deals with the identification of analytes present in the sample.
2. *Quantitative analysis*: Aims to determine the amount of analytes present in the sample.
3. *Structural analysis*: Aims to elucidate chemical structures.

Mass spectrometry is an analytical technique that plays an important role in the different types of analyses defined above. Section 2.2 provides insights into this technique.

2.2 Mass spectrometry

Mass spectrometry (MS) is a powerful analytical technique that separates gas phase ions extracted from a sample, based on their *mass-to-charge ratio* (m/z) and detects them qualitatively and quantitatively using their respective m/z and abundance. The m/z represents mass of the ion divided by its charge number. The number of electrons added or removed is the charge number of the ion. Mass spectrometry should not be confused with spectroscopy which deals with absorption of electromagnetic radiations. The basic variable parameter in spectroscopy is wavelength whereas in mass spectrometry it is m/z [26, 27]. With advancements in technology, this technique has grown further to offer unequaled sensitivity, speed and detection limits, as compared to other analytical methods. The important role of MS as an analytical tool is well recognized due to which this technique is routinely used in the qualitative, quantitative and structural analysis of peptides and proteins [28], carbohydrates [29], nucleic acids [30], lipids [31], drugs [32] and metabolites [8].

A standard MS experiment is performed using an instrument known as a mass spectrometer and usually involves the following steps - Sample introduction into the mass spectrometer, analyte ionization, mass analysis, ion detection and recording, data processing and interpretation of results [27, 33]. The mass spectrometer is usually operated

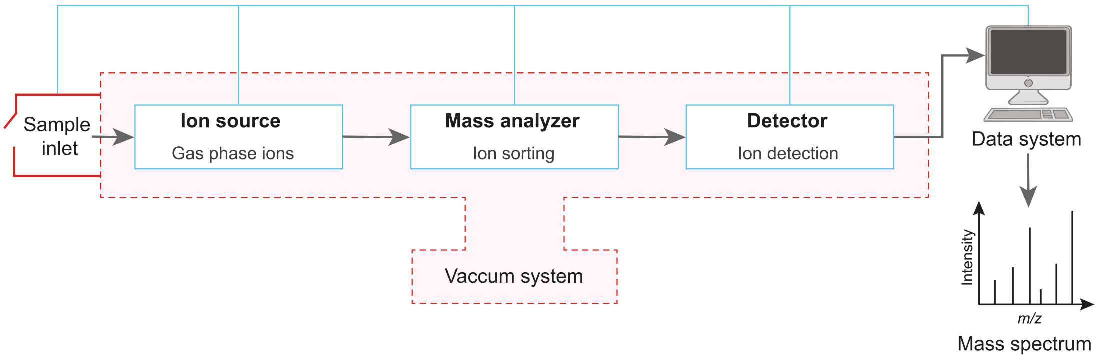


Figure 2.1: Schematic representation of a mass spectrometer. Standard mass spectrometer consists of five components: The sample inlet using which the sample is introduced into the instrument, an ion source that generates the gas phase ions, a mass analyzer which separates these ions based on their m/z, an ion detector which detects the arriving ions and a data system which records this data and generates the mass spectrum. The data system is also responsible to control the components of the mass spectrometer, represented as blue lines within the diagram.

under high vacuum so as to minimize ion-molecule reactions as well as scattering and neutralization of the ions. In the following section, we look into the basic setup of a mass spectrometer.

A mass spectrometer consists of the following essential components:

1. *Sample inlet* using which the sample is introduced into the mass spectrometer.
2. *Ion source* which generates the gas phase ions.
3. *Mass analyzer* which receives the ions from the ion source and separates them according to their m/z.
4. *Ion detector* which registers and counts the arriving ions.
5. *Data system* which is a computational setup that controls the various components of the mass spectrometer and acts as a data recorder/processor that stores the data once it is generated.

A schematic diagram of a mass spectrometer can be found in Figure 2.1. Each component comes in a variety of forms. The components mainly differ based on their mode of operation and by their advantages and limitations for a particular analytical application. By linking individual components together, different MS systems can be configured to be specific for different analytical goals. In the coming sections, we will describe only those technologies that are relevant to this thesis.

2.2.1 Sample inlet

The sample is introduced through an inlet to the ion source. The selection of the sample inlet to be used mainly depends on the state of the sample i.e. gas, solid, liquid or solution and the means by which ionization is induced. Gases and samples with high vapor pressure are introduced directly into the ion source region. Liquids and solids are usually heated to increase the vapor pressure for analysis. If the analyte is thermally labile i.e. it decomposes at high temperatures or if it does not have a sufficient vapor pressure, then the sample is directly ionized from the condensed phase. Some of the routinely used sample introduction techniques are direct vapor inlet, gas chromatography, liquid chromatography, direct insertion probe and direct ionization of the sample [34].

2.2.2 Ion source

The purpose of an ion source is to generate charged gas phase ions from the sample injected. This charge can be either positive or negative in polarity. This is done by evaporation of solid samples, vaporization of liquids, atomization of gaseous compounds and ionization of atoms and molecules so as to generate ions that can be sent to the mass analyzer. In all types of ion sources, during the ionization process singly or multiply charged atomic ions are formed. The most important considerations while using an ionization technique are the internal energy transferred during the ionization process and the physico-chemical properties of the analyte that has to be ionized. Depending on the amount of energy that is added to the molecules during ionization, the ionization technique can either induce little or no fragmentation that only produces ions of the molecular species, also called *soft ionization*, or result in extensive fragmentation, which is known as *hard ionization*.

2.2.3 Mass analyzer

Once ions are formed in the ion source, they are accelerated into the mass analyzer by an electric field. The main function of a mass analyzer is to separate these ions based on their m/z values in order to be detected. Mass analyzers can either be continuous or pulsed type. Continuous mass analyzers include quadrupole filters and magnetic sectors whereas pulsed analyzers include time-of-flight, ion cyclotron resonance and quadrupole ion trap mass spectrometers. Each mass analyzer has its own advantages and limitations and there is no single mass analyzer optimal for all situations.

The choice of which analyzer to use is critical and usually involves consideration of the mass range, scan speed, mass accuracy, mass resolving power and abundance sensitivity [35, 36]. Before we get into the details of the commonly used mass analyzers, it is important to get acquainted with the following terminologies:

Mass range determines the limit of m/z over which the mass analyzer can measure ions. It is the difference between the highest and the lowest measurable m/z .

Scan speed describes how fast a mass analyzer can record mass measurements over a particular mass range. It is expressed in mass units per second (u s^{-1}) or in mass units per millisecond (u ms^{-1}). A fast scan speed is desirable when a rapidly changing system is being analyzed whereas a slow scan speed can be used to obtain precise mass measurements.

Mass accuracy Δm_{ac} is a measure of the closeness of the observed mass to the true mass of the analyte. It is defined as the difference between the *exact mass*[1] (m_a) of an ion and its observed peak centroid m/z (m_{exp}) and is usually quoted in *parts per million* (*ppm*) (see Figure 2.2(a)). This can be expressed in the form of Equation 2.1.

$$\Delta m_{ac} = \mid m_a - m_{exp} \mid \tag{2.1}$$

---

[1] *Exact mass* is the calculated mass of a molecular ion or molecule whose elemental formula, composition and charge state are known.

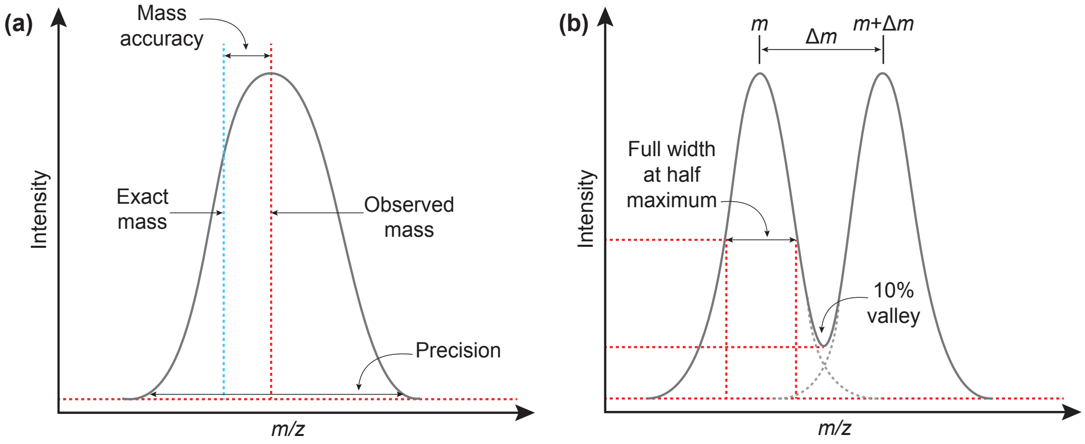


Figure 2.2: Illustration of the measurement of mass accuracy and mass resolution. **(a)** Plot representing measurement of mass accuracy which is the difference between the exact mass of an ion (in *blue*) and the observed mass (experimentally measured peak centroid, in *red*). **(b)** Plot displaying the most common measures to calculate mass resolution are full width at half maximum (FWHM) and 10% valley.

It should be noted that m_{exp} or the experimentally determined mass is also known as the accurate mass[2].

The accuracy of a single reading is described using the *mass measurement error*. It is always desirable to have a high mass accuracy since this increases the likelihood of uniquely identifying the elemental compositions of the measured ions. Another term related to mass accuracy is *precision*. It is the repeatability of the measurement reflecting random error. A set of measurements is considered to be precise if these errors are low.

Mass resolving power is the ability of the mass analyzer to separate one mass from an adjacent mass differing by a small increment. *Mass resolution* (R) is defined as the observed m/z divided by the smallest difference $\Delta(m/z)$ for two ions that can be separated. It is important that the procedure using which $\Delta(m/z)$ was measured and the m/z value at which the measurement was made is reported [37]. The mass resolution can be calculated using Equation 2.2.

$$R = \frac{m/z}{\Delta(m/z)} \tag{2.2}$$

where $\Delta(m/z)$ is either the peak width which is a specified fraction of the maximum peak height or the spacing between two equal-intensity peaks that are separated by a valley which at its lowest point is 10% of the height of either of the peaks. Mostly three values for peak width are used 50%, 5% or 0.5%. A common standard is the definition of resolution based upon $\Delta(m/z)$ defined as the Full Width of the peak at Half its Maximum height (FWHM). A poor mass resolution leads to the inability to determine the peak position accurately in the presence of nearby peaks. A diagrammatic representation of the 10% valley and FWHM to calculate mass resolution is shown in Figure 2.2(b).

Abundance sensitivity is the degree to which the signal arising from a mass peak contributes to the adjacent masses. It is defined as the signal contribution of the tail of a peak at one mass lower and one mass higher than the actual analyte peak [38]. If the

---

[2] *Accurate mass* is the mass of a molecular ion or molecule determined experimentally measured to a significant degree of accuracy.

Table 2.1: Comparison of some of the popularly used mass analyzers. The listing of figures of merit of all the mass analyzers are taken from MucLuckey *et al.* [39] and Hart-Smith *et al.* [36].

Properties	Mass accuracy (ppm)	Mass resolving power	Mass range (Da)	Linear dynamic range	Abundance sensitivity	Advantages	Limitations
Quadrupole	100	10^2 - 10^3	4000	10^7	10^4 - 10^6	Low cost Less space needed Well suited for electrospray Easy analysis of positive/negative ions	Limited mass range of detection
Ion trap	50 - 100	10^3 - 10^4	4000	10^3 - 10^4	10^3 - 10^4	Low cost Less space needed Well suited for tandem MS analysis Easy analysis of positive/negative ions	Limited mass range of detection
TOF	5 - 50	10^3 - 10^4	$>10^5$	10^6	10^6	Highest mass range of detection High scan speed Moderate cost Simple design	Low resolution Difficult to adapt for electrospray
FT-ICR	1 - 5	10^4 - 10^6	$>10^4$	10^3 - 10^4	10^3 - 10^4	Very high resolution and accuracy Well suited for tandem MS analysis	Expensive instrumentation More space required Low scan speed
Orbitrap	2 - 5	10^4 - $15{\times}10^4$	6000	10^3 - 10^4	10^4	Very high resolution and accuracy Well suited for tandem MS analysis Moderate cost	Low scan speed

Values presented in the table may vary with hybrid configurations from different instrument manufactures; please refer the manufacturer's technical specification.

mass analyzer has a poor abundance sensitivity, it will often prohibit the measurement of a small peak next to a major interfering peak.

Table 2.1 compares some of the popular mass analyzers. Below we describe the time-of-flight mass analyzer in detail as the imaging datasets used in this thesis have been acquired using this analyzer type.

Time-of-flight analyzer

Time-of-flight (TOF) analyzers are based on the concept that ions with the same kinetic energy but different mass (m) travel with different velocities (v). This can be represented as:

$$eV = \frac{1}{2}mv^2 \tag{2.3}$$

Here e is the electric charge and V is the accelerating potential. In these types of analyzers, ions are accelerated through a fixed length region (D) known as a flight tube towards the detector. In the absence of a magnetic or electric field these ions begin to separate according to their masses and have differing velocities. Lighter ions arrive earlier than the heavier ones and hence are recorded first.

In such analyzers, the flight time (t) for ions is proportional to the square root of their masses (m):

$$t = \left(\frac{m}{2eV}\right)^{\frac{1}{2}} D \tag{2.4}$$

One of the main advantages of TOF analyzers is that the upper mass range has no limit, making it suitable for soft ionization techniques. They also have a high scan speed and a high transmission efficiency that leads to a very high sensitivity. However, linear TOF analyzers suffer from a poor mass resolution. This is because in these analyzers, the travel time t is proportional to the square root of the mass m, due to which, as the mass increases it leads to the decrease in the Δt for a given $\Delta m/z$. This issue can be addressed by having a high acceleration and long flight tubes, since they increase the difference between the different m/z ratios. The issue of poor resolution has now substantially improved with further advances in technology [40].

2.2.4 Detector

Once the ions are separated according to their m/z in the mass analyzer, they reach the detector where they get detected and are transformed into a signal which is then recorded. Since the number of ions coming to the detector at a particular instant is usually quite small, it becomes necessary to perform amplification in order to record a usable signal. Ideally, the electric signal generated from the incident ions is proportional to their abundance. However, instruments in real time cannot provide this proportionality for all the masses.

Depending on the type of the detector used, ion detection is based either on charge of the ions, their mass or their velocity [35]. Some of the most common detectors used in a MS setup are: *Faraday cup*; which is based on the measurement of direct charge current that is produced when an ion hits a surface, and *electron multipliers*; that are based on the kinetic energy transfer of ions when they hit a surface that in turn generates secondary electrons, which are amplified further to give an electronic current. However, these detectors have a limitation that their efficiency generally decreases as the m/z increases. The new age detectors like *inductive detectors* and *cryogenic detectors* address this issue.

2.2.5 Mass spectrum

After ion detection, the received signals are sent to a computer-aided data system where the m/z ratios are stored together with their relative abundance measures which is presented in the form of a *mass spectrum*. It is a two-dimensional representation of the m/z which is represented on the x-axis and its corresponding signal intensity which is represented on the y-axis. A mass spectrum can also be represented as a *tabular list* with 2 columns - m/z and intensity, as *analog (profile) form* where each peak has a height and a width and is displayed in a continuous manner or as *digital (centroided) form* where each peak corresponding to a specific ion is represented as a vertical line [41]. This vertical line is drawn through the centroid of each peak profile, where the height represents the signal intensity.

2.3 Mass spectrometry imaging

Mass spectrometry can provide us with qualitative, quantitative as well as structural information about the sample analyzed. However, within complex biological systems it is also important to understand the spatial context in which molecular changes take place. This spatial information often helps in understanding the biological functions performed by different biomolecules and also can help in determining disease biomarkers. A standard MS experiment does not provide this spatial information. *Mass Spectrometry Imaging* (MSI) has emerged as an enabling technique to address this issue. The terms *mass spectrometry imaging* and *imaging mass spectrometry* are used interchangeably in a number of publications. However, in this thesis, we will only use *mass spectrometry imaging* or simply MSI.

MSI is a label-free analytical technique enabling investigation of the spatial arrangement and relative chemical concentration of different compounds in biological samples, with a high chemical specificity, in a broad mass range. Given a biological sample, MSI measures high dimensional mass spectra at each spatial position (also referred to as *pixel*) on the sample surface. These mass spectra are later reconstructed to form a hyperspectral image displaying the spatial biochemical composition within the sample of both known and unknown molecules [42, 43].

There are four main steps involved in a typical MSI workflow. First is sample preparation which highly depends on the ionization technique employed (discussed in Section 2.3.1). Ionization is performed at defined spots on the sample surface in such a way that the complete sample is rastered. Ions from every spot reach the mass analyzer and are recorded individually. The acquired mass spectrum is associated with a coordinate position on the sampled biological section. This is followed by data processing and visualization in the form of *2-dimensional (2-D) ion intensity maps* (explained in Section 2.3.2) [44].

Based on the goal of the experiment and how data acquisition is performed, MSI experiments can be classified into two different types [45–47]:

1. **Microprobe**
   In this mode, a highly focused laser beam is used to analyze a small, localized spot of the biological sample and a complete mass spectrum is recorded for the specific coordinate position on the sample. The beam is then focused on the next spot and mass spectrum is recorded for it. This process is repeated until the complete array of spots on the entire sample surface is sequentially examined in x and y directions.

2. **Microscope**
   In this mode, ions from a large sample area on the biological section are desorbed simultaneously. Then the ion optics of the instrument projects the ionized substances from this area to a position sensitive detection system such as a microchannel plate. These detectors register the signal of a specific m/z over the whole sample area, at once as well as retain the spatial information.

The microprobe mode is more widely applied and the datasets used in this thesis are acquired using this approach. A schematic of a typical MALDI-MSI experiment using the microprobe mode is shown in Figure 2.3.

2.3.1 Techniques

A typical MSI experiment is usually performed using a standard mass spectrometer. As discussed in Section 2.2, a mass spectrometer contains an ion source where the ionization of the analyte is performed. The selection of the ion source to be used is largely based on the goal of the experiment and the sample properties. Some of the commonly

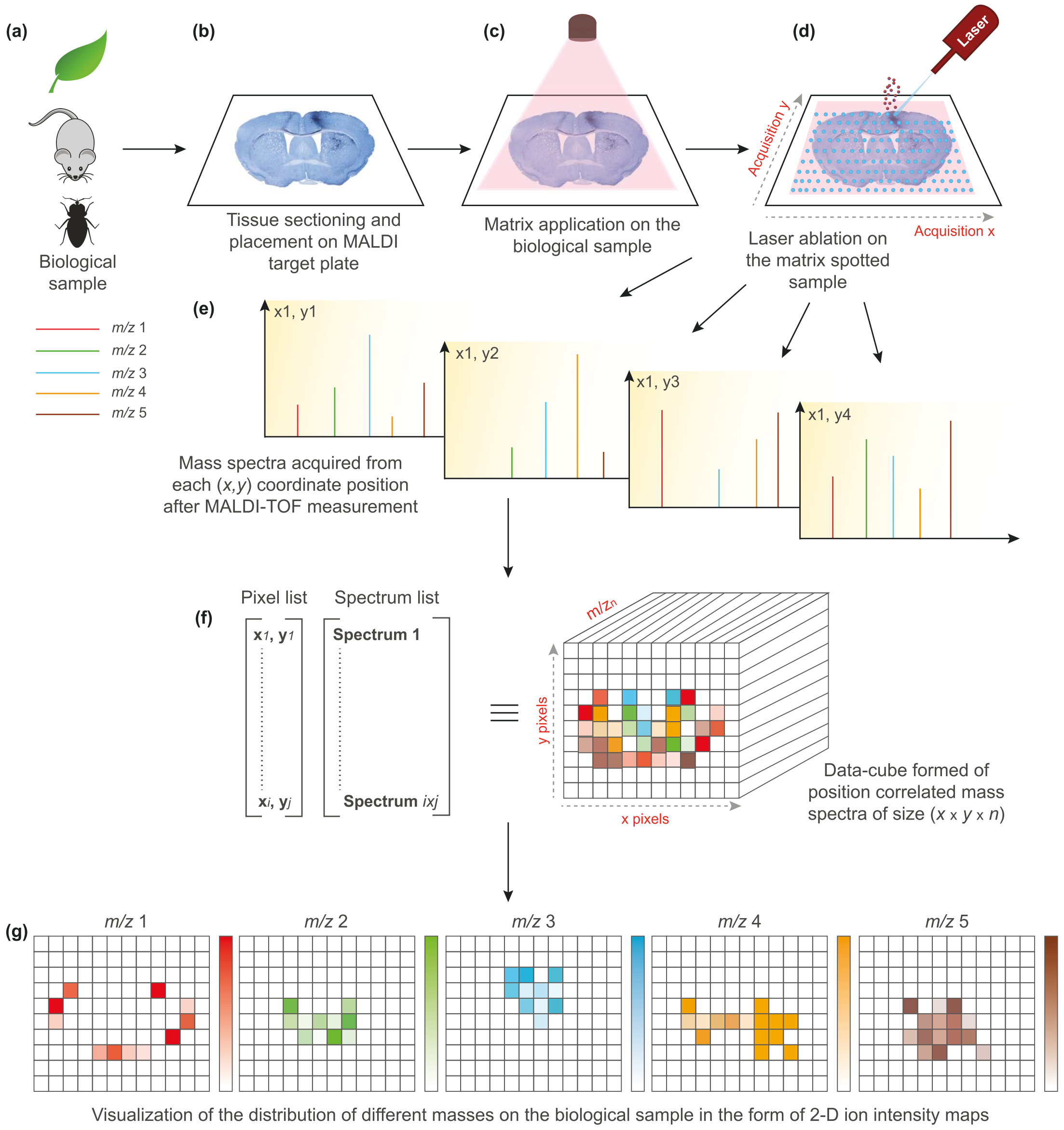

(a)
Biological sample
(b)
Tissue sectioning and placement on MALDI target plate
(c)
Matrix application on the biological sample
(d)
Laser
Acquisition y
Acquisition x
Laser ablation on the matrix spotted sample
m/z 1
m/z 2
m/z 3
m/z 4
m/z 5
(e)
x1, y1
x1, y2
x1, y3
x1, y4
Mass spectra acquired from each (x,y) coordinate position after MALDI-TOF measurement
(f)
Pixel list
Spectrum list
x1, y1
xi, yj
Spectrum 1
Spectrum ixj
≡
m/zn
y pixels
x pixels
Data-cube formed of position correlated mass spectra of size (x x y x n)
(g)
m/z 1
m/z 2
m/z 3
m/z 4
m/z 5
Visualization of the distribution of different masses on the biological sample in the form of 2-D ion intensity maps

Figure 2.3: Schematic outline of the typical work flow for a scanning microprobe MALDI mass spectrometry imaging experiment and data visualization. **(a)** Based on the aim of the experiment a biological sample is selected. **(b)** The pretreatment steps include cryo-sectioning and mounting the biological sample on a MALDI target plate. In case of whole body analysis, the sample is directly placed on the plate. **(c)** The biological sample is then evenly coated with a matrix, which helps in extracting molecular ions from the tissue. **(d)** A focused laser beam scans the complete sample area with dimensions ($x \times y$) in a raster pattern and desorbs and ionizes molecules from the surface at every coordinate position. **(e)** The resulting ions from each coordinate position are transfered to the mass spectrometer and a mass spectrum is acquired before moving to the next position. **(f)** The acquired data consisting of a list of mass spectra with their corresponding coordinate positions can be described as a three-dimensional (3-D) *data-cube* of size ($x \times y \times n$). **(g)** To visualize the distribution of a specific analyte, a *2-D ion intensity map* is generated. It is generated by extracting the signal intensity of the selected m/z from all the mass spectra acquired. The scale represents the relative intensity of the analyte at a specific coordinate position in the map.

Table 2.2: Comparison of commonly applied MSI techniques. The listing of figures of merit of the ionization techniques are taken from Wu *et al.* [48], Nemes *et al.* [49], Esquenazi *et al.* [44] and Bodzon-Kulakowska *et al.* [50].

Ion source	Type of ionization	Ionization source	Sensitivity and resolution	Mass range (Da)	Limitations	Commonly analyzed analyte classes
MALDI	Soft	Laser beam	fmol - zmol, 10 - 100 μm	0 - 100000	Matrix signals may interfere with the low m/z region	Lipids, proteins, peptides
SIMS	Hard	Primary ion beam	Varies with m/z, 10 nm - 100 μm	0 - 1000	Low sensitivity for high masses (>1000 Da)	Small molecules
DESI	Soft	Solvent spray	fmol - pmol, 40 - 400 μm	0 - 2000	Has an analyte washing effect	Small molecules, lipids, peptides
LAESI	Soft	Mid-infrared laser beam	8 - 25 fmol, less than 20 - 200 μm	0 - 2000	Water content in the sample affects ablation characteristics and ionization	Small molecules, lipids, peptides, proteins

used ionization techniques for MSI are Matrix Assisted Laser Desorption/Ionization (MALDI), Secondary ion mass spectrometry (SIMS), Desorption electrospray ionization (DESI) and Laser ablation with electrospray ionization (LAESI). Table 2.2 shows the main features of these ionization techniques.

In this section, we will discuss MALDI and SIMS, that are relevant for this thesis.

Matrix assisted laser desorption/ionization MSI

Matrix Assisted Laser Desorption/Ionization MSI or MALDI-MSI is currently the most widely used MSI technique [44, 46, 50]. In this technique, the sample usually a biological tissue, is thinly sliced approximately 5-20 μm thick and mounted on a MALDI target plate (for imaging intact biological samples which are three dimensional in nature, as discussed by Kaftan *et al.* [2], special MALDI target plates with profiled holes are also used). The target plate containing the tissue slice is then evenly coated with a matrix. This homogeneous matrix layer usually consists of small organic molecules that can absorb majority of the laser energy incident on it. Once the solvent in the matrix is evaporated, it leads to the crystallization of the matrix and incorporation of the analyte molecules into the growing crystals. This plate containing the crystals is subsequently irradiated with a laser beam typically with micron and sub-micron dimensions, by rastering across the surface. This leads to the desorption and ionization of the matrix and analyte molecules. In positive ionization mode, singly protonated molecular ions $[M+H]^+$ are generated from analytes in the sample, whereas in negative ionization mode singly de-protonated ions $[M-H]^-$ are generated [51–53]. These ions are subsequently separated by a mass analyzer based on their masses. A schematic of the MALDI process is depicted in Figure 2.4(a) [54]. The TOF analyzer (discussed in Section 2.2.3) is popularly used for MALDI-MSI studies.

Based on the goal of the study, the choice of matrix is important. Different matrices enable the ionization and desorption of different types of biomolecules, as summarized in Table 2.3.

This ionization approach has a high mass range (100,000 Da) and is capable of producing intact higher molecular weight ions since they get incorporated into the matrix crystals and do not get fragmented [46]. This makes MALDI-MSI very useful for recording intact biomolecules specially peptides, proteins and lipids.

Table 2.3: Matrices commonly used in MALDI-MSI [55–57].

Biomolecule type	Matrix used for MALDI-MSI
Metabolites	9-Aminoacridine (9-AA)
Lipids	2,5-Dihydrobenzoic acid (DHB) 2,6-dihydroxyacetophenone (DHAP) p-nitoroaniline (PNA) 9AA
Peptides	α-Cyano-4-hydroxycinnamic acid (CHCA) DHB
Proteins	Sinapinic acid (SA) CHCA
Nucleic acids, small sugar molecules	1,8-bis(dimethyl-amino)naphthalene (DMAN)
Organic acids, amino acids	1,5-diaminonapthalene (DAN)

Secondary ion mass spectrometry imaging

Secondary ion mass spectrometry (SIMS) is a destructive analytical technique that is used to analyze the composition and structural layers of biological tissues. This technique utilizes a highly focused and energetic primary ion beam (e.g. Ar^+, Xe^+, Cs^+, Ga^+, In^+) of about 5-25 KeV to bombard the sample surface that induces multiple collisions with atoms and molecules on the surface [58]. This results in the emission of mostly neutral but also charged secondary ions. This process is also known as *sputtering*. These secondary ions are then introduced into a mass analyzer after acceleration using a high voltage acceleration system. The SIMS ionization process is shown in Figure 2.4(b). The most commonly used analyzer along with SIMS is the TOF analyzer (Section 2.2.3).

One of the main features of SIMS technique is the extensive fragmentation of molecules on the sample surface, as compared to MALDI (Section 2.3.1). This happens because, the energy of the primary ions deposited into the sample surface is substantially higher than the energy deposited by the laser beam in the case of MALDI. Also, because of this extensive fragmentation, the SIMS mass spectra are usually complex and the mass range is limited to $\sim m/z$ 1000 Da.

SIMS usually requires minimum sample preparation. Specifically for SIMS imaging the sample preparation usually just involves mounting of the biological tissue section on a glass slide coated with indium tin oxide [46]. SIMS imaging can analyze a wide range of samples however it is favorable to have a flat sample surface, so as to avoid the adverse effect of surface topology on the generation of secondary ions and mass measurement [42]. Because of the small size of primary ion beam, SIMS offers a high spatial resolution of about 10 nm, making it possible to image the cellular and sub-cellular distribution of a variety of compounds in the sample. However, it is only suitable for the analysis of small molecules, because the secondary ion yield in SIMS decreases rapidly and non-linearly with increasing mass, hence offering low sensitivity for higher mass ions [51, 59].

With advancement in technology, different modifications for SIMS have been developed so as to enhance ionization and large molecule detection capabilities.

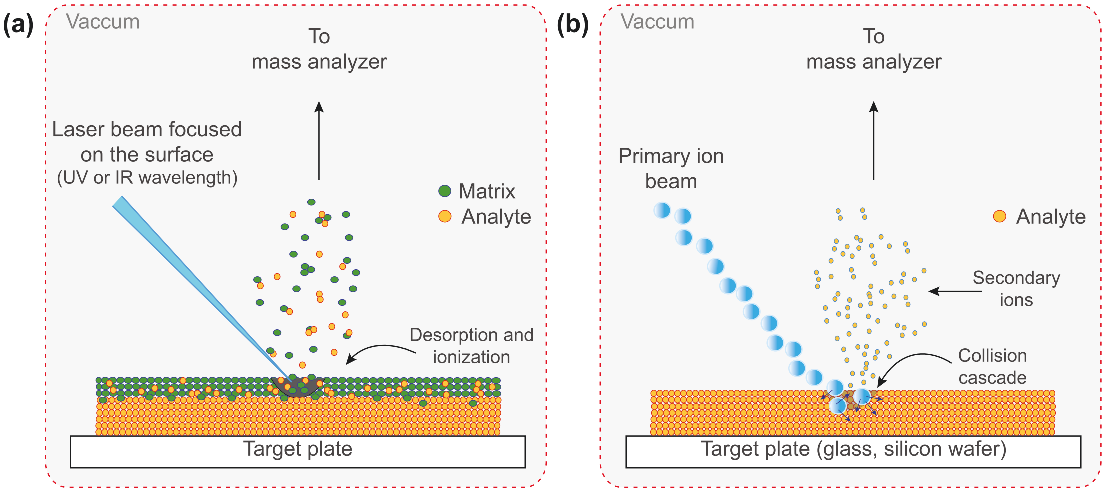


Figure 2.4: Ionization mechanisms for MALDI and SIMS ionization sources. **(a)** In a MALDI experiment, the sample mounted on a target plate is evenly coated with a matrix. The sample is then shot with a highly focused laser beam whereby the energy is absorbed by the matrix and ionization of the matrix and analyte ions is induced. These ions are then passed to the mass analyzer. **(b)** In a SIMS experiment, the sample is placed on a target plate and a high-energy primary ion beam is incident on the sample. Upon collision with the sample surface, the energy of these primary ions is transfered to the analyte, generating secondary ions which are then passed to the mass analyzer.

2.3.2 Data features, computational aspects and visualization

As discussed in Section 2.3, during MSI data acquisition, a mass spectrum is acquired from every single spatial position on the sample. For each coordinate position, the data is recorded as a profile of intensity values over a corresponding range of m/z values. With the use of modern mass spectrometers that generate high resolution molecular profiles and depending on the sample size, a single MSI experiment can generate massive amount of data often ranging a few gigabytes.

A single MSI dataset consists of $x{\times}y$ coordinate positions (pixels), corresponding to the dimensions of the sample area imaged and n m/z values which are the spectral data points or bins. So, the dataset can be described as a three-dimensional (3-D) *data-cube* or a *hyperspectral image* of size ($x{\times}y{\times}n$) (Figure 2.3(f)). This 3-D *data-cube* comprises of peak intensity values for each m/z value corresponding to a specific coordinate position (x_i,y_j) [60, 61], where i is the row number and j is the column number in the data matrix.

The distribution as well as the localization of a specific analyte of interest with a known m/z value, can be visualized using a *2-D ion intensity map* (Figure 2.3(g)). This map is constructed using the intensity values extracted from all the spectra present in the dataset, at the selected m/z value. We use the term *2-D ion intensity map* or simply *ion intensity map* throughout this thesis to refer to such an image.

One of the key parameters while performing an MSI experiment is its resolution. There are three types of resolution defined for a MSI experiment:

1. *Mass resolution* provides the degree of chemical specificity. A detailed description on mass resolution is provided in Section 2.2.3.

2. *Spatial resolution* which is also known as the *lateral resolution*, determines the size of features that can be seen as individual pixels on the 2-D ion intensity

map [50, 62]. The more is the number of pixels from which the mass spectra is acquired, the better is the lateral resolution of the analysis. This means that, on increasing the spatial resolution, smaller and more detailed features of the sample can be seen on the acquired image. But a higher spatial resolution also leads to an increase in the volume of data acquired. There are a number of factors that determine the lateral resolution. These are: *spot size* which denotes the size and shape of a single spot from which a single acquisition is performed, *raster size* or *step size* which is the distance between two consecutive spots and *pixel size* and *density* of pixels which determines the ability to distinguish two features. Most commercially available instruments offer spot size of 10-200 μm [63].

3. *Depth resolution* is specifically considered while performing SIMS imaging (discussed earlier in Section 2.3.1). SIMS provides in-depth information on the atomic constituents of the sample. The sputtering process gradually erodes the sample surface and provides information on the first few mono-layers below the initial surface, also known as the *depth profile*. The quality of this depth profile is measured by the depth resolution. It is described in terms of the ability to discriminate between atoms in adjacent thin layers [45, 64]. The depth resolution achievable depends on nature of the primary ion beam used, the depth below the surface and the uniformity of sputtering by the ion beam.

Due to the massive size of data produced in a single experiment, it is tedious to perform manual analysis. Many free as well as commercial data processing, data mining and visualization software have been developed to aid in the analysis of this data. A comprehensive list is available here: `http://ms-imaging.org/wp/sotware-tools/`. Standard data processing steps and analysis pipelines have been discussed in Chapter 3.

Data description and representation

Below we provide a standard notation to describe MSI data that is used throughout this thesis.

A single MSI dataset is stored as a 2-D matrix. Letters in upper case, bold font, denote matrices and letters in the lower case, bold font, denote vectors. Letters in italic font, denote scalars. The matrix transpose is denoted by an apostrophe. All indices are taken to run from one to their capital versions. Each spectrum is a row vector and intensities of the distribution of an ion across all the pixels is a column vector in the data matrix. The term *sample* denotes a single mass spectrum from the MSI dataset and the term *variable* denotes the mass (m/z) of a single ion.

2.3.3 Applications of mass spectrometry imaging

MSI is now commonly applied for the analysis of a wide variety of biological systems ranging from intact whole-body samples like insects [2], complex whole-body tissue sections [65, 66] to specific biological tissue samples and single-cells [67, 68]. It is also being increasingly used for the study of disease pathology, tissue-based disease classification, biomarker identification, molecular expression and drug resistance patterns in diseases like cancer [69, 70], Alzheimer's disease [71], Parkinson's disease [72], muscular dystrophy [73] and kidney disease [74].

Although, existing clinical imaging methods like immunostaining and fluorescence-based techniques are popularly used to study the spatial localization of biological substances, these techniques are highly specific and usually allow the visualization of only a single class of analyte per sample. Also, techniques such as magnetic resonance imaging provide only structural information. On the contrary, MSI allows for the label-free discovery of multiple classes of biomolecules present in the sample in a single run [75]. To add value to the results, many experiments now apply a multimodal approach and combine results obtained from multiple MSI techniques as well as other classical histological techniques [76, 77]. This allows to correlate the molecular and spatial data obtained from MSI to the structural information obtained from conventional clinical techniques.

Chapter 3

Computational processing and analysis of MSI data

As discussed in Chapter 2, the principal goal of an MSI experiment is to determine the chemical specificity and the spatial distribution of molecules which can then be correlated to the underlying biology of the sample analyzed. A single MSI dataset may contain hundreds to thousands of mass spectra, each containing signals from hundreds to thousands of mass peaks. To aid in the analysis of such high-dimensional data, robust computational algorithms are important to enable extraction and visualization of relevant information from the data.

In this chapter, Section 3.1 introduces a standard data processing workflow employed in analyzing MSI data. In Sections 3.2 and 3.3, we describe the file formats used post acquisition and explain the preprocessing steps necessary before analyzing the data. Later, in Section 3.4, we explain multivariate data analysis (MVA) approaches specifically focusing on unsupervised techniques, that can help in interpreting multi-dimensional MSI data and can extract distinct correlations within ion distributions. To conclude, in Section 3.5, we briefly discuss supervised classification and the emerging trend of combining and interpreting data from different imaging modalities to aid in areas like biomarker identification.

Selection and application of the unsupervised multivariate data analysis methods explained here is presented as a case study in Chapter 7, where we analyze co-aggregation patterns found on the leaf surface of *Populus trichocarpa* using TOF-SIMS imaging.

3.1 Overview of the data analysis workflow

For analyzing MSI data, when the aim is to study specific ions of interest, the simplest method is to visualize the spatial distribution of each ion. This can be done by constructing 2-D ion intensity maps (explained in Section 2.3.2). These maps can then be compared with optical images of the tissue section that has been analyzed. For more complex analysis like biomarker discovery, where significant molecular information or characteristic signals are not known *a priori* and the goal is to study correlations between hundreds of detected molecules, manual analysis of data is impractical. In such cases, exploratory analysis of MSI data is important to understand the underlying spatial patterns and identify significant mass features in the data.

The steps used for analyzing such data depends on the goal of analysis or the biological question. Figure 3.1 illustrates a typical data analysis pipeline that can be applied to MSI data [63, 78–80].

Every single step has subdivisions and the methods applied for each step should be chosen wisely as this can impact on the quality or effectiveness of the subsequent steps and can highly influence the results generated.

We explain individual steps comprising the data analysis workflow in the next sections.

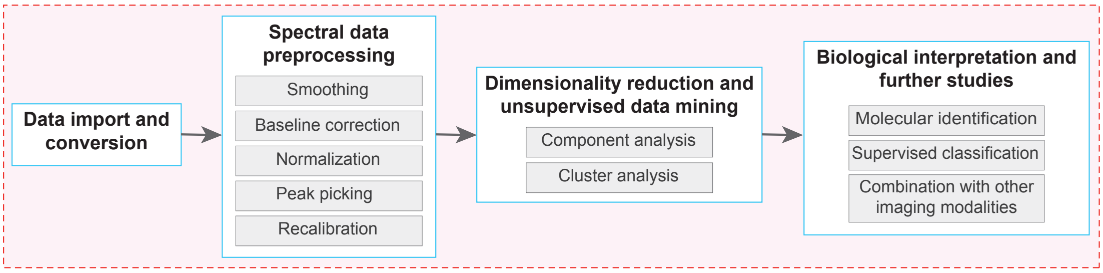

Data import and conversion
Spectral data preprocessing
Smoothing
Baseline correction
Normalization
Peak picking
Recalibration
Dimensionality reduction and unsupervised data mining
Component analysis
Cluster analysis
Biological interpretation and further studies
Molecular identification
Supervised classification
Combination with other imaging modalities

Figure 3.1: Typical pipeline for the interpretation of a MSI dataset. The workflow illustrates the key steps involved in analyzing and interpreting a MSI dataset post acquisition. It should be noted that the methods selected and their order within each step of the pipeline depends on the goal of analysis and nature of the data.

3.2 Data import and conversion

MSI data is usually acquired in a proprietary vendor specific instrument format and can be analyzed using the commercial software bundled along with the instrument. However, with increase in the application of MSI, many open source software packages have also become available that provide strong data analysis and statistical functionalities. To ensure efficient and flexible exchangeability of MSI data between different instrument setups and software, a standard data format known as *imaging mzML* or *imzML*, is now established [81]. Many available software packages now allow import and export of MSI data in *imzML* format. Converters like `imzMLConverter` [82] allow conversion of different file formats to **.imzML*.

3.3 Spectral data preprocessing

Multivariate analysis (MVA) techniques describe the underlying structure of the data and are therefore very sensitive to data scaling and transformations. This makes data preprocessing of paramount importance before applying these techniques.

The raw data produced during an MSI experiment may often be affected by *noise*: random fluctuation between measured spectra leading to unwanted peaks not belonging to the sample analyzed. There are commonly two sources of noise: *electrical noise* arising from the components of the mass spectrometer, and *chemical noise* that may be derived from contaminants or simply a chemical background that is generated during the measurement process [83, 84]. It is usually expected that the analyzed sample contains a certain number of compounds that should each produce an intense peak with a predictable shape in a mass spectrum. The electrical noise present in the mass spectrum, is expected to produce a continuously varying low-intensity signal, which may be detected as a large number of very small, badly shaped peaks. This noise may also distort the true shape of peaks in the mass spectrum. Along with this other factors like peak broadening, instrument distortion and saturation, isotopes, mis-calibration as well as contaminants may add unwanted variation in the data leading to erroneous interpretation [85]. The quality of a mass spectrum can be measured using the *signal-to-noise ratio* (S/N). It is an indicator of how much a peak corresponding to a compound present in the sample, is distinguishable from its background noise [26]. A spectrum is usually considered to have high S/N ratio if a peak detection algorithm finds a small number of high-intensity peaks, a large number of low-intensity peaks and few or no peaks with intensities in between.

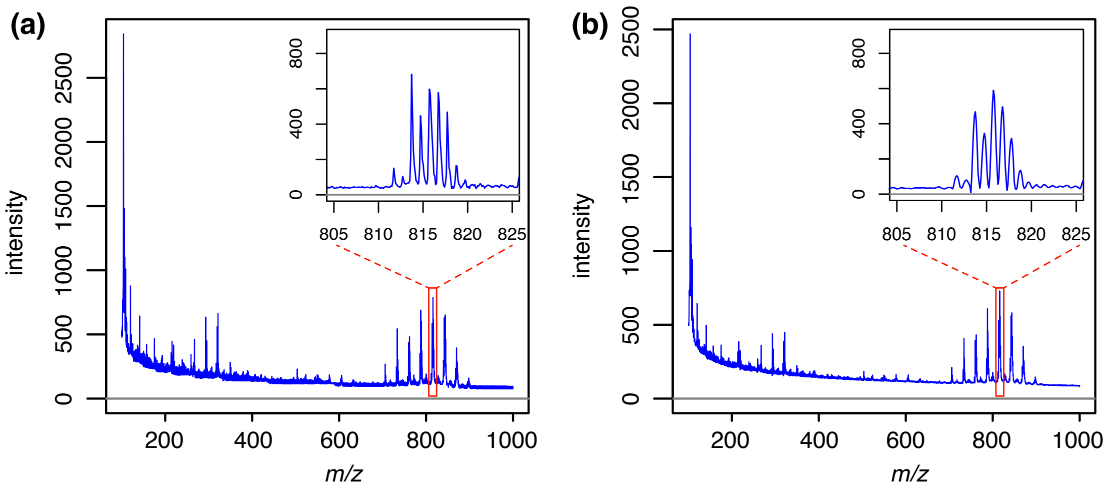


Figure 3.2: Representation of smoothing on a MALDI-TOF mass spectrum. (a) Raw mass spectrum in *blue*. **(b)** Smoothed mass spectrum after applying the Savitzky-Golay filter [89] implemented in the `MALDIQuant` R package [90].

Preprocessing is performed after data acquisition and helps in cleaning unwanted noise. It also helps in reducing the overall experimental variance in order to make all the spectra comparable within a single dataset [86]. As shown in Figure 3.1, there are multiple steps involved in preprocessing data, each one correcting for a particular artifact [87]. It should be noted that the order of baseline correction and smoothing when applied to a particular dataset can be changed depending on the nature of the data [88].

3.3.1 Smoothing

Smoothing is performed to increase S/N ratio in a mass spectrum, mainly under the assumption that high-frequency components are more likely to be noise than signal. This preprocessing step helps in mitigating the effect of noise in the spectrum by reducing the fluctuations in the observed signal, in turn improving the performance of peak-picking (explained further in Section 3.3.4). This is done by removing high-frequency noise present in the data [91, 92].

There are three smoothing algorithms that are commonly used: *Moving average filter* [93], *Savitzky-Golay filter* [89] and *Lowess filter* [94]. The moving average filter smooths the spectra by replacing each data point with the average of the neighboring data points in a specific span. A single span should always contain odd number of data points and the data points to be smoothed should always be at the center. Greater the span width, the more intense is the smoothing effect. The *Savitzky-Golay filter* performs smoothing by fitting a small subset of the data to a polynomial using least squares regression. The filter coefficients are derived by performing a linear least squares fit using a polynomial of a given degree. The advantage of using this filter is that it can preserve signal features such as the resolution and height of the peaks. The amount of smoothing can be controlled using the span size and the polynomial order selected. The *Lowess filter* smooths a mass spectrum by using a locally weighted linear regression method. This filter finds a data value by averaging the neighboring values within a span of data points. A regression weight function is defined for all the data points that are within the span. This step is repeated for every point in the signal. The amount of smoothing performed depends on the span size selected. It is important that the span size is not

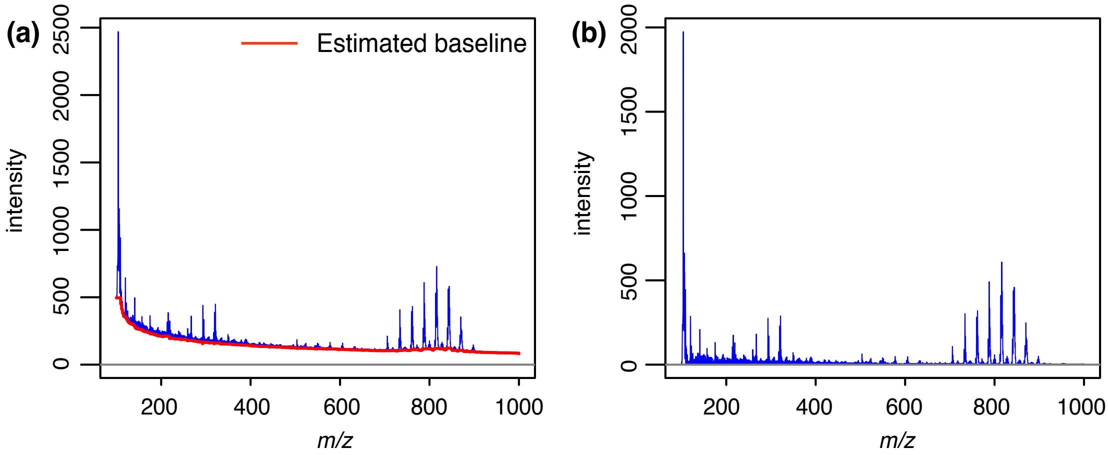


Figure 3.3: Illustration of baseline estimation and correction on a MALDI-TOF mass spectrum. **(a)** Raw mass spectrum in *blue* with the estimated baseline in *red*. Baseline estimation is performed using the *Top-Hat filter* [95] algorithm implemented in the `MALDIQuant` R package [90]. **(b)** Baseline-corrected mass spectrum after subtraction of the estimated baseline.

large, since this may lead to loss of information. Figure 3.2 shows smoothing performed on a raw MALDI-TOF spectrum using the *Savitzky-Golay filter*.

3.3.2 Baseline correction

Baseline correction is a preprocessing step performed to identify the baseline and remove it from the mass spectrum. Theoretically, an acquired mass spectrum is composed of the true signal, the baseline and noise. This can be represented in the form of the following equation:

$$f_{(i,j)} = b_{(i,j)} + s_{(i,j)} + \varepsilon_{(i,j)} \tag{3.1}$$

where $f_{(i,j)}$ is the observed value (i.e the raw spectrum), $b_{(i,j)}$ is the baseline value, $s_{(i,j)}$ is the true signal and $\varepsilon_{(i,j)}$ is the noise for the *ith* sample at the *jth* m/z ratio.

Baseline correction helps to reduce the noise in the data by flattening the base profile of the spectrum. In a raw mass spectrum, there is typically a large amount of background noise at lower m/z that decays exponentially as the mass increases. Specifically for MALDI-MS data, prominent chemical noise is generated by the matrix molecules in the sample. The effect of this chemical noise can be suppressed by estimating a best fit *baseline* for a spectrum in an iterative manner. This estimated baseline is then subtracted from the spectrum to get the baseline-corrected spectrum. Since, the baseline varies from one spectrum to the other, every spectrum must therefore be treated separately. For baseline correction in a single MSI dataset, the baseline is heuristically estimated for each spectrum and is then subtracted from the raw spectrum intensities, making sure that no peak information is removed from the spectrum.

Many approaches have been developed for baseline correction. The *Top-Hat filter* algorithm proposed by Sauve *et al.* [95] is a fast morphological operation which applies a moving minimum and a moving maximum filter on the intensity values in a selected mass window. Shin *et al.* [96] propose a wavelet-based method that computes the decreasing baseline in the highest approximation of the wavelet domain. Also, baseline estimation algorithms based on polynomial fitting by Williams *et al.* [97] and nonlinear iterative peak-clipping (known as the *SNIP* algorithm) by Ryan *et al.* [98] have been proposed.

An algorithm proposed by Andrade and Manolakos [99] assumes a probabilistic mixture model to find the mean height of the baseline in each window. This algorithm is implemented in the MATLAB Bioinformatics Toolbox (MATLAB and Statistics Toolbox Release R2012b, The MathWorks, Inc., Natick, Massachusetts, United States) as the `msbackadj` routine and is commonly used. Figure 3.3 illustrates baseline correction performed on the smoothed MALDI-TOF spectrum (shown in Figure 3.2) using the *Top-Hat filter*.

3.3.3 Normalization

Normalization is a crucial step to make spectral measurements comparable by transforming them to a common intensity scale. For routine MS experiments, systematic intensity differences are observed in the data acquired from repeated measurements for the same sample. This can be due to different experimental factors like intensity shift leading to differences in the *total ion current* (TIC) of each measurement, loss of sensitivity, sample degradation and homogeneity differences between runs of the same sample [84, 91, 100, 101]. To perform normalization, the mass spectrum $x[n]$, with n data points is divided by a certain normalization factor (f) to generate a normalized mass spectrum $y[n]$, as shown in Equation 3.2 [102]:

$$y[n] = \frac{1}{f(x)} x[n] \tag{3.2}$$

Different normalization approaches reported in literature can be classified in two categories: global and local [95]. In the global approach, it is assumed that intensities of all spectra are related by a constant factor and hence all features in the spectra are used simultaneously to compute a single normalization coefficient. Some common approaches are *TIC normalization* and *median normalization.* TIC represents the sum of all the separate ion currents carried by the ions of different *m/z* contributing to a complete mass spectrum [103]. In TIC normalization, each spectrum is normalized to the sum of all the ion currents as shown in Equation 3.3 where $p^1 = 1$.

$$f(x) = \left(\sum_{n=0}^{N} |x[n]|^p \right)^{\frac{1}{p}} \tag{3.3}$$

Median normalization provides an approximate normalization to the baseline of the mass spectrum. It can be represented as given in Equation 3.4.

$$f(x) = median(x) \tag{3.4}$$

TIC and median normalization assume that the measured peak intensities are directly proportional to the concentration of metabolites in the sample. And, if an equal amount of sample was used in each run, then the summed intensity in the spectrum should be the same across all spectra. Therefore the TIC normalization approach multiplies each spectrum with a scaling factor so that all spectra have the same total intensity [104]. However, this approach is only recommended when the TIC's do not vary significantly

---

[1] A *norm* of a vector is the measure of its size. Normalization is the process of scaling a vector so that its norm is unity. This can be carried out in any norm and can be achieved by simply dividing the vector by its norm. There are several different types of norms one of which is *p-norm*, for any integer $p \geq 1$. Normalization on TIC is a special case of *p-norm*.

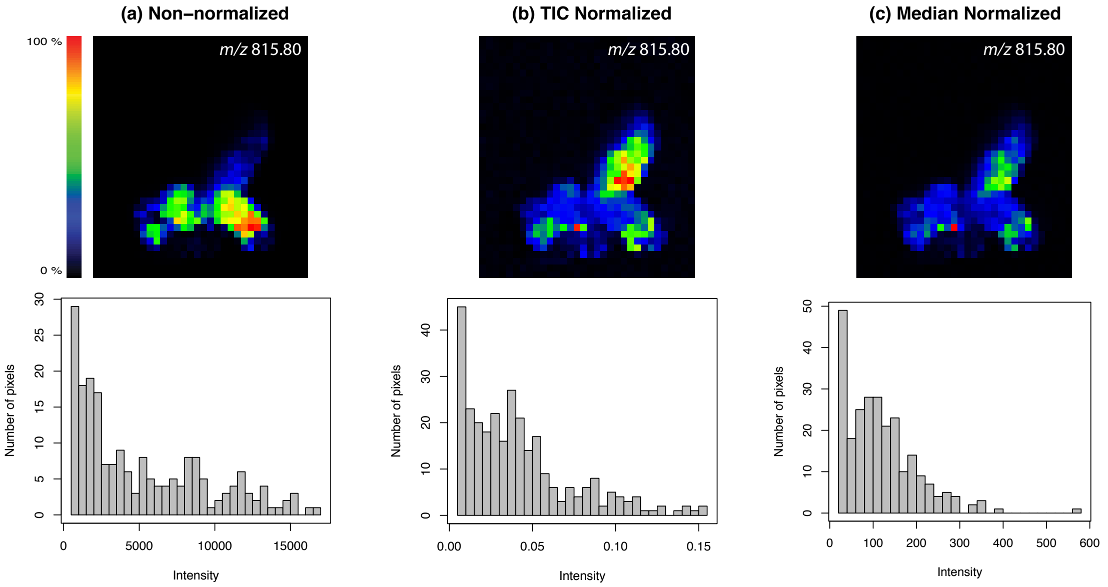


Figure 3.4: TIC and median-based normalization on a MALDI-TOF dataset obtained by MSI of virgin male *Drosophila melanogaster*. The 2-D ion intensity maps are generated for the distribution of *m/z* 815.80 and their corresponding histograms display the intensity distribution over the pixels **(a)** for smoothed and baseline-corrected data without any normalization applied to it **(b)** for smoothed and baseline-corrected data that has been TIC-normalized **(c)** for smoothed and baseline-corrected data that has been median-normalized. Smoothing is performed using the *Savitzky-Golay filter*, baseline correction using the *Top-Hat filter* implemented in the `MALDIQuant` R package [90]. TIC and median-based normalization are also performed using the functions implemented in the `MALDIQuant` R package. Pixels outside the fly body are not considered while plotting the histogram to allow better visualization of the effect of normalization.

across the dataset [105]. In median normalization, instead of summing all the signals, the median is used.

In the local approach, each spectrum is normalized to some reference or a specific feature independent of the collective dataset. One such popular approach is to normalize intensities relative to the *base peak* (peak with the maximum intensity) [106]. Another normalization approach uses the height of an isotopically labeled internal standard peak. This mass peak corresponds to a compound with a known mass and with the same amount of compound added to each sample. The intensity of each peak in the mass spectrum can then be normalized to that of the internal standard. Another popularly used local normalization approach is *locally weighted scatter plot smoothing* normalization. An in-depth comparison of commonly applied normalization approaches was performed by Deininger *et al.* [102].

For MSI data, it is important to normalize each spectrum in a single dataset, since the data acquired in each laser shot can vary. It may happen that for a certain coordinate position on the imaged sample, more ions were detected, resulting in *hotspots* as compared to other positions. This may occur due to factors like uneven matrix coating, differential ionization efficiencies and crystal inhomogeneity, especially in the case of MALDI-MSI [107]. This can result in spot-to-spot variance in the signal intensities within the same imaged section. Inter-spectrum normalization helps in removing these systematic differences in the data. However, for MSI data, the normalization methods discussed above may not necessarily be optimal. This is because these methods rely on

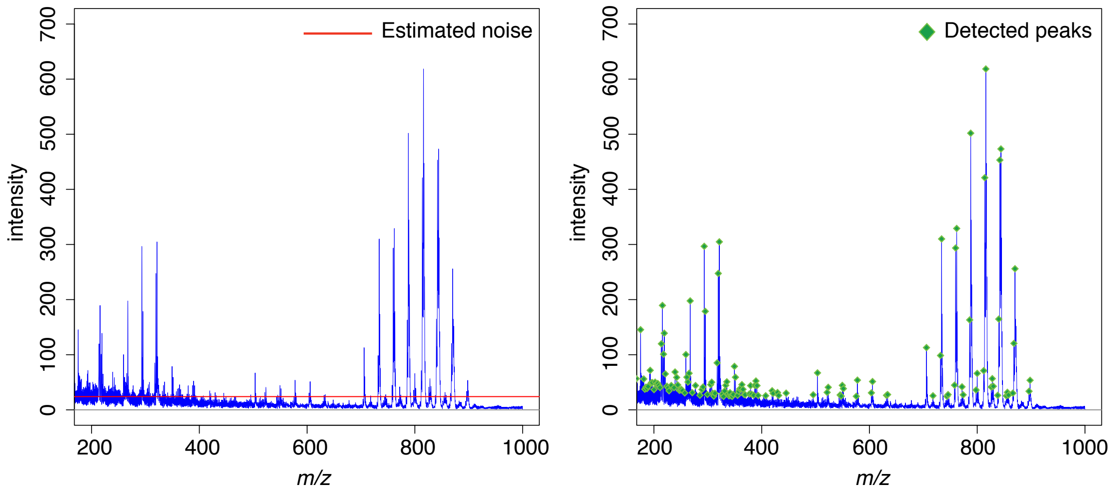


Figure 3.5: Noise estimation and peak detection on a MALDI-TOF mass spectrum. **(a)** Smoothed and baseline-corrected mass spectrum in *blue* with the estimated level of noise in *red*. Noise estimation is performed using the *Median absolute deviation* approach implemented in the `MALDIQuant` R package [90]. **(b)** Detected peaks labeled with *green* diamonds, within the mass spectrum based on the estimated signal to noise ratio.

a set of assumptions that may not be fulfilled due to the varying molecular composition in the sample imaged. TIC normalization tends to equalize the intensities of the distinct biological regions, which may lead to inaccurate representation of ion distribution in 2-D ion intensity maps. It has also been observed that performance of these methods is compromised noticeably due to the presence of single large molecular ion peak intensities because of their substantial contribution to the total peak intensity [108, 109]. Figure 3.4 shows the effect of TIC and median based normalization on the distribution of m/z 815.80 on *Drosophila melanogaster*. As can be seen, the two normalization strategies applied generated different spatial distribution. This shows that it is important to carefully select the appropriate normalization approach depending on the experimental design. Considering the nature of MSI data, specialized normalization approaches are also being developed, like the one proposed by Fonville *et al.* [23], which is based on the median pixel intensity of every single coordinate position in the imaged dataset.

3.3.4 Peak picking

A *peak* in a mass spectrum can be defined as a local maximum above a user defined noise threshold. The main aim of peak picking is to reduce the number of m/z values by extracting informative peaks from the dataset and neglecting any noise or baseline signals present [63, 110]. This can prove to be a challenging task specifically in the case of data with low S/N ratio and it may also happen that certain low abundance peaks may remain buried within noise, leading to a high false positive rate of peak detection. Considering these factors the peak picking method to be used should be carefully selected.

Peak detection algorithms mainly use one or more of the following criteria to identify peaks: *S/N ratio*, which considers a signal above a fixed ratio as a true positive; *intensity threshold*, which removes all small peaks below a certain user-defined threshold; *peak shape*, which helps in filtering out false peaks; *local maximum*, which helps in selecting a peak that is a local maximum of n neighboring peaks; *peak width*, which is the mass difference of the right end point and left end point of a peak above a certain noise threshold and is important to be considered for low-resolution data where peak width varies a lot; *peak shape ratio*, which helps in selecting a peak if its shape ratio, a quantity computed as the peak area divided by the maximum of all peak areas, exceeds a selected threshold and *model-based criterion*, which uses a model function to fit peaks. A comparison of different peak detection approaches presented by Yang *et al.* [88] showed that methods that consider the peak shape along with its intensity perform better than those that consider only the peak intensity.

Peak picking specifically for MSI data can pose new problems because of the immense volume of the data. A peak detection method in this case should be selected considering both efficiency and sensitivity. To this end, many approaches have been proposed that either allow selection of a region of interest from the complete dataset to perform peak picking [111] or select those peaks that are present in at least 1% of spectra [112]. Figure 3.5 shows peak detection performed on a smoothed and baseline-corrected MALDI-TOF spectrum using the *Median absolute deviation* approach, that estimates noise as the median of the absolute deviation of points within a selected window.

3.3.5 Spectral recalibration

For every mass spectrometry experiment, external calibration is performed. Even then, the systematic shifts during the measurement can affect the acquired mass spectrum. These mass shifts can interfere with interpretation of the data and can lead to erroneous annotation of peaks. Mass shifts in the data can be corrected using spectral recalibration. In Chapter 4, we provide a detailed description of spectral recalibration and the existing approaches proposed in literature. Within that chapter, we also present our recalibration method (see Section 4.3) and discuss its performance.

3.4 Dimensionality reduction and unsupervised data mining

After applying the preprocessing steps and peak picking to MSI data, considerable noise is eliminated and S/N ratio is assumed to be improved. However, individual MSI datasets (specifically those acquired at a high resolution or from whole body samples) can still be extremely large in size, making manual interpretation infeasible. This becomes even more difficult when little or no information is available about the sample that is analyzed. In such a scenario, unsupervised techniques are often helpful for exploratory data analysis and to extract important features without prior knowledge of the sample.

The high-dimensionality of MSI data introduces complexity in analysis as well as increases the computation time. To give an example of the complexities of analysis, clustering techniques are often applied to mass spectra. These techniques, which may be applied to many kinds of data, use a distance measure to describe the similarity of pairs of objects. Under very high dimensionality, all distances calculated by these similarity measures tend to equal one another and the discriminatory power of these measures fails [113, 114]. These problems arising due to high dimensionality are collectively known as the *curse of dimensionality.* It is a term introduced by Bellman [115] to describe the problem caused by the exponential increase in volume associated with adding extra dimensions to a Euclidean space.

Many *dimensionality reduction* approaches have been developed that aim to represent the data by a small subset of extracted features with minimum information loss, while preserving as much of the variance present in the data as possible. These features capture the informative portions of the signal whilst discarding the noise and redundant measurements. Some of the dimensionality reduction as well as MVA approaches are principal component analysis, independent component analysis, partial least squares discriminant analysis, non-negative matrix factorization and multivariate curve resolution. These techniques are also now being popularly applied to analyze MSI data [24, 105, 116–118].

3.4.1 Component analysis

Component analysis techniques help in the simultaneous statistical study of the dependence (covariance) between different variables, using a small number of factors.

Given a data matrix (X) composed of I samples (mass spectra) and K variables (mass units), component analysis performs bilinear decomposition on this matrix X to generate three new matrices, which is represented in Equation 3.5:

$$X = TP' + E \tag{3.5}$$

The *scores matrix* (T) is a projection of the samples onto the factors and is of dimensions $I \times N$, where N is the number of factors[2]. This scores matrix consists of the contributions. The *Loadings matrix* (P) is the projection of a factor onto the variables and is of dimensions $K \times N$ and contains the spectra. The *Residual matrix* (E) mainly represents noise and is of dimensions $I \times K$.

In this section, we will discuss Principal Component Analysis (PCA) and Multivariate Curve Resolution (MCR; also known as Alternating Least Squares regression) since these are applied to analyze the TOF-SIMS imaging data discussed in Chapter 7. PCA and MCR are both methods of component analysis or factor analysis, where the data is described using a small number of factors, however they have different applications. The aim of PCA is usually to explore and interpret the dataset whereas MCR helps in defining contributions of the constituents (components) using concentration and spectral profiles [119, 120].

These two techniques differ in the way the factors are extracted. To explain them, we use terminologies that are consistent with ISO standard vocabulary [121].

Principal component analysis

PCA is probably the most popular factor analysis method and has been widely applied to analyze MSI data [109, 116, 122–124]. Described first in 1901 by Pearson [125], PCA looks at the variance pattern within a dataset to find the direction of greatest variance. It extracts a set of uncorrelated orthogonal factors that re-orients the data onto a new set of perpendicular axes in such a manner that the axes are aligned along directions of maximum variance within the data [126]. These orthogonal factors are known as principal components (PCs) and are ordered such that the first PC accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. Not all PCs generated provide valuable information and it is preferable to discard higher PCA factors. This step is often referred to as *factor compression*. It is important to carefully determine the number of factors required to describe the main features of the data. A wrong selection of the number of components can lead to loss of information (underestimation)

---

[2] A *factor* is basically an axis in the data space of a factor analysis model, representing an underlying dimension that contributes to summarizing or accounting for the original data set. Individual factors typically represent interesting properties of a dataset.

or the inclusion of noise components (overestimation). The number of factors can be deduced by inspection of the eigenvalue[3] plot, also known as the *scree test* [127], and the percentage of total variance captured by the first N PCA factors. Hence, by selecting a smaller number of factors to represent the data, PCA is able to achieve dimensionality reduction for large MSI datasets. These generated PCs are linear combinations of all the original variables and, therefore, usually capture much more information than any of the original variables considered individually.

The *scores* describe the relationship among the samples (i.e. the different mass spectra in the dataset). The *loadings* are simple correlations between the components and the original variable. They define the contributions of the original variables (i.e. individual m/z values) to the new PCs and describe which variables are responsible for the differences seen within the samples. The *residual matrix* describes the random variations not described by the new PC axes and usually represents noise in the data.

Multivariate curve resolution

While PCA calculates factors based on mathematical properties (like capturing maximum variance), the factors may display negative intensities due to imposition of orthogonality. These negative intensities are often difficult to interpret because they are not directly related to chemical properties of the mass spectral data. Also, PCA loadings for a data set of measured spectra generally are not pure component spectra. Instead, the loadings are typically linear combinations of pure analyte spectra that have positive and negative intensities.

MCR is an approach to decompose a hyperspectral data matrix and is designed to identify pure components from a multi-component mixture [128, 129]. This bilinear decomposition is usually performed by repeated application of multiple least squares regression. The technique extracts chemically meaningful information in the form of factors that resemble the spectra of chemical components and contributions. Applying MCR to multivariate images yields information about what analytes are present and where in the image they are located [130]. MCR assumes a linear combination of chemical spectra (MCR loadings) and contributions (MCR scores) to describe each spectrum. Since MCR factors are not required to be mutually orthogonal, by applying a non-negativity constraint[4] to the loadings and scores matrices during optimization, the MCR components are directly interpretable as spectra of pure compounds, as they have positive values [131–133].

The downside of MCR when compared to PCA is that it is computationally more intensive and requires more user input prior to analysis. Also, in the case of PCA, the orthogonality constraint ensures that only one linear combination of original variables gives the optimal approximation of the data. However, this is not the case in MCR. The solutions generated by MCR are not unique. This is because the original data matrix may be reproduced in an infinite number of ways by using component profiles differing in shape (rotational ambiguity) or in magnitude (intensity ambiguity) from the true ones. Hence, quality and uniqueness of MCR solutions is strongly dependent on the number of chemical components or species assumed to be causing the data variance. Also, using appropriate constraints can help in limiting the ambiguities in MCR solutions.

---

[3] An *eigenvalue* is the amount of variance described by each factor.

[4] *Constraints* are defined as systematic properties or mathematical conditions that help MCR to generate a optimal and chemically meaningful data matrix. Some commonly used constraints for MCR are non-negativity, unimodality, closure and hard-modeling.

The very first indication of the number of chemical species present in a dataset can be obtained directly from the rank[5] of the data matrix or from the number of significant singular values associated with the data matrix [134]. The singular values that are related to chemical species are usually larger than noise, systematic errors or baseline values. The initial number of components can also be roughly estimated based on the number of components in a PCA or singular value decomposition that are sufficient to explain the systematic changes in the data variance, i.e. by selecting the number of eigenvalues higher than those associated with the noise level [119, 135]. This method works well for uniformly distributed homoscedastic noise, but fails when this is not the case. This is where it is important to clean the data of any baseline and instrumental contributions and apply suitable preprocessing steps [136]. A different approach to identify the number of components was applied by Motegi *et al.* [137]. They perform a comparison of the concentration profiles generated from MCR results obtained by sequentially changing the number of components for a single dataset. They observed that similar components emerged repeatedly. Based on this observation, it was considered that reliable components emerged repeatedly irrespective of the number of components selected whereas unreliable components emerged only once or just a few times. These reliable components were considered to be informative.

To obtain initial estimates of the spectral profiles, a straightforward approach is to choose some pure spectra from the original data matrix. The SIMPLISMA [138] method is also popularly used for selecting pure spectral variables as initial estimates. To determine the initial estimates for concentration profiles, methods such as evolving factor analysis and evolving window factor analysis are used [139].

3.4.2 Cluster analysis

To perform exploratory data analysis, clustering is yet another technique that can aid in highlighting unknown patterns. It is a class of unsupervised methods that allow classification and grouping of objects based on their similarity (or differences). These methods can be classified into two main categories: *agglomerative* and *partitional* (also known as segmentation or divisive methods). *Agglomerative methods* such as hierarchical clustering analysis (HCA) begin with each object being its own cluster, and progress by combining existing clusters into larger ones. *Segmentation methods* such as discriminative cluster analysis (DCA) start with a single cluster containing all objects, and progress by dividing existing clusters into smaller clusters or segments based on their similarities or homogeneous composition [105, 116, 140–142].

Hierarchical clustering analysis

The principal behind HCA is that the more similar two objects are, the closer they are in multidimensional data space. For MSI data, the objects referred to in HCA are either mass spectra or ion intensity maps. HCA starts with N clusters, each of which includes one data object. The algorithm successively merges clusters in an iterative manner, based on the selected similarity criterion, until no more merges are possible. The main advantage of using HCA is that multi-dimensional data can be easily summarized based on its distinguishing features and can be visualized in 2-D space. The procedure to perform HCA can be summarized in the following steps [143, 144]:

---

[5] The *rank* of a matrix is the maximum number of rows or columns that are linearly independent. It represents the number of independent parameters that are needed to fully describe the data.

1. Start with N objects and calculate pairwise similarity between these objects, based on a defined similarity measure, to generate a distance matrix.
2. Search the pair of objects C_i and C_j with the minimum distance in the matrix and merge them into a single cluster C_{ij}.
3. Update the distance matrix by computing the distance between the cluster C_{ij} and other objects. The distance between an object and a cluster or two clusters is calculated using the linkage function (explained in detail further below).
4. Repeat steps 2 and 3 in an iterative manner until all objects are contained in one large cluster.

Results from HCA are represented in the form of a tree diagram called *dendrogram*, where clustering on different levels can be visualized, and/or as a heat map which represents the individual values in the sample matrix with a color code, showing how various patterns can segregate defined groups. The dendrogram can be broken at different levels to yield different clusterings of the data. Each leaf in the dendrogram corresponds to one of the original objects used for clustering and each internal node represents a subset or cluster of objects. Cutting the dendrogram at a certain height gives the number of clusters, that the input objects have been divided into. However, there is no objective way to say how many clusters are generated, as this can change with the change in height at which the dendrogram is cut. Although, there have been some methods reported [145], but in any case, there is a fair amount of subjectivity in determining which branches to be cut or not to be cut to form separate clusters.

Measure of similarity: There are various measures to express (dis)similarity between pairs of objects. The most commonly used distance measure is the *Euclidean distance* (d). It measures the shortest path, basically the length of the straight line, connecting two points and can be defined as the square root of the sum of the squared distances between two points. In general, the distance between two points x and y in a Euclidean space $\mathbb{R}^n$ is given by:

$$d = |x - y| = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \tag{3.6}$$

Other types of distance measures used are Manhattan distance, Mahalanobis distance and Pearson's correlation.

Cluster linkage: In addition to a distance measure, a measure to compute the distance between two clusters or a object and a cluster is also needed. Such a measure is referred to as a *linkage*. It is a function that takes two cluster nodes as input and provides the distance between these nodes. Some of the popularly used linkage measures are:

1. *Single linkage*, where the distance between two clusters is the distance between the two closest data points in these clusters (each point taken from a different cluster).
2. *Complete linkage*, where the distance between two clusters is the distance between the two furthest data points in these clusters. To cluster ion intensity maps based on their spatial similarity (discussed in Chapter 6), we have used the complete linkage function.

3. *Average linkage*, where the distance between two clusters is the distance between the centers of the two clusters.

4. *Ward's linkage*, which aims to minimize the within-cluster variation such that the resulting groups are as homogeneous as possible [146–148]. It uses an analysis-of-variance approach to evaluate the distances between clusters. In this approach, the linkage function estimates the distance between the two clusters, which is the increase in the *error sum of squares* (ESS) after fusing two clusters into a single cluster [149]. ESS can be expressed mathematically as:

$$ESS(X) = \sum_{i=1}^{N_x} \left| x_i - \frac{1}{N_x} \sum_{i=1}^{N_x} x_i \right|^2 \tag{3.7}$$

where X is a cluster, x is an object in cluster X and N_x is the total number of objects in cluster X. Using Equation 3.7, Ward's linkage can be represented as:

$$D_w(X,Y) = ESS(XY) - [ESS(X) + ESS(Y)] \tag{3.8}$$

where X and Y are two clusters, $|\cdot|$ is the absolute value of a scalar value or the length of a vector and XY is the combined vector obtained by fusion of clusters X and Y. This linkage approach tends to produce clusters with similar numbers of observations, but it is sensitive to outliers. To analyze the TOF-SIMS data (discussed in Chapter 7), we have used the Ward's linkage function.

Discriminative cluster analysis

DCA is a segmentation approach. In the context of ion maps corresponding to a MSI dataset, segmentation is a process of separating an image into regions or *segments*, based on their homogeneous chemical composition. This can be performed by the use of clustering approaches that group the data in a few segments or clusters so that the data points within a cluster exhibit similar characteristics whereas data points across clusters are different. DCA performs clustering by obtaining a single partition of the data instead of generating a dendrogram as in HCA [142, 150]. There have been many DCA approaches proposed like k-means, k-medians, mixture of Gaussians, graph theoretic, however, the k-means approach is the most widely applied. The k-means algorithm starts with k objects (clusters), where k is specified by the user *a priori*. During each cycle of this clustering method, the remaining objects are assigned to one of these clusters, based on distance from each of the k targets. New cluster targets are then calculated as the means of the objects in each cluster, and the procedure is repeated until no objects are re-assigned after the updated mean calculations. When k-means is applied to MSI data, the algorithm classifies each pixel into one of the k clusters either by minimizing the sum of distances from their respective centers or by maximizing interclass distance, which usually leads to the most distinct clusters possible.

The k-means algorithm is simple to implement and offers good performance. However, it also suffers from some drawbacks. It is sensitive to the initial number of clusters selected, meaning regions of the biological sample revealed are strongly dependent on the number of clusters selected by the user. It can be difficult to estimate the optimal number of clusters in advance. Selecting a good number requires a combination of statistical reasoning (for example, use of a Silhouette plot [151] to study the separation between the resulting k clusters), some knowledge about the sample data being used and also depends on human judgement. In practice, k-means clustering is performed by using different values of k to get a series of solutions. The final choice of k is made based on qualitative criteria of the clusters obtained [152]. Also, its complexity in time is $O(nkl)$ and in space is $O(k)$, where n is the number of samples, k is the number of clusters, and l the number of iterations. This degree of complexity could at times be impractical for large datasets [153].

One major disadvantage of applying HCA and k-means to MSI data is that these methods treat each pixel independently and ignore similarities of spectra acquired from spatially proximate locations, i.e. they do not take into account any spatial relationships [22, 105]. This can adversely affect the quality of segmentation.

Spatially aware segmentation

To address the disadvantage of the traditional clustering approaches explained above, spatially aware segmentation methods have been developed. Alexandrov *et al.* [43, 112] have proposed two such approaches that incorporate spatial relations between pixels, so that pixels are clustered together with their neighbors:

1. *Spatially aware clustering* is a distance-based clustering approach where the distance between two spectra, obtained from two pixels in a MSI dataset, depends on the neighboring pixels of the selected pixels. This method is based on the assumption that mass spectra acquired from neighboring pixels in a morphologically defined region on the biological sample most likely should represent similar biochemical composition and so should be similar. To take into account this spatial context, a pixel neighborhood radius is used which is selected by the user. The method uses specific Gaussian weights corresponding to pixels from the neighborhoods. These weights decrease with increasing distance from the neighborhood center. One drawback of this approach is that it selects the neighboring pixels in the same manner for every pixel in the dataset, which can lead to elimination of smaller details, particularly in complex tissue samples.

2. *Spatially aware structure-adaptive clustering* has been developed to counter the drawback of the spatially aware clustering approach. This type of clustering is based on the same principle. However, in this the weights of pixels in its neighborhood are not simply Gaussian but are calculated adaptively, i.e. they take into account similarities of pixels and the structure observable in the data.

3.5 Biological interpretation and further studies

It is a crucial as well as a challenging task to make sense of the huge amount of data generated in a MSI experiment, so as to answer a specific biological question. Unsupervised methods discussed in the previous section can provide an overview of the dataset and can highlight variance-causing features in the data. However, several studies also require identification of unknown molecules as well as data-dependent classification of the samples analyzed, leading to biomarker discovery. This section briefly describes these approaches used for MSI data analysis.

3.5.1 Molecular identification

As described previously, a single MSI experiment generates volumes of spectral data that reflects complexity of the sample analyzed and requires identification of the unknowns. However, molecular identification is inherently challenging, particularly when low-resolution instruments have been employed to generate the data. With advancements in instrumentation, experiments are now performed that combine high-resolution accurate-mass imaging with data-dependent tandem MS (MS/MS) [154–156]. It is important to have high mass-resolving power and high mass accuracy so that compounds can be identified with a higher confidence based on their accurate mass and MS/MS fragmentation patterns.

Recently, Alexandrov *et al.* [157] developed a computational framework for false discovery rate (FDR)-controlled compound annotation for high-resolution MSI data. This framework is based on three principles: database-driven annotation by screening for metabolites with known sum formulas, calculation of a specific metabolite signal match score that quantifies the likelihood of the presence of a metabolite with a given sum formula, and a FDR estimation approach with a decoy set generated using implausible adducts.

3.5.2 Supervised classification

Supervised classification approaches in MSI are mainly employed to identify molecular profiles or specific ions that can act as potential biomarkers in order to discriminate groups of samples, e.g. a tumor tissue from a benign tissue in clinical and pathological studies. These approaches require prior knowledge about the samples to be analyzed, and also require sample annotation in order to generate a model and identify candidate biomarkers [105]. Model generation usually involves training a classifier using specific input features of the samples in order to discriminate spectra of different types [158]. Some of the popularly used supervised classification approaches for MSI data include linear discriminant analysis [159], support vector machines [160], artificial neural networks [161] and random forests [162]. These approaches are now popularly employed as primary tools to study and characterize a number of cancer types, Alzheimer's disease, arthritis and other disease biomarkers [163].

3.5.3 Combination with other imaging modalities

As described in Chapter 2, there are several different ionization approaches developed for MSI, including SIMS, MALDI, and DESI. Other popular imaging modalities, such as magnetic resonance imaging (MRI), high resolution magnetic resonance spectroscopic imaging, etc are also used in clinical studies. All these imaging modalities differ in their analytical capabilities and two or more of these can be used to analyze a single biological question. However, until recently, analyses using these approaches were treated as separate entities [62].

An emerging trend is to combine two or more imaging modalities for validation and confirmation of findings. This approach has been termed *multimodal imaging*. This imaging approach can be a combination of MSI with an independent imaging modality like MRI or can use a combination of different ionization modalities in MSI. Several studies using multimodal imaging have been reported for glycan and protein analysis [164], lipids [165], small molecule analysis for cancer studies [166] and many more.

Chapter 4

Recalibration of mass spectrometry imaging data

When analyzing biomolecules using mass spectrometry, mass accuracy and mass resolving power play a crucial role.

Mass accuracy increases with the increase in mass resolving power and high mass resolution. High mass resolution is important because it minimizes the possibility of overlap of two closely placed mass peaks [167]. With high mass accuracy it is possible to achieve accurate peak annotation and uniquely identify the elemental compositions of observed ions.

In Section 4.1, we discuss about mass shifts that can compromise mass accuracy, leading to erroneous compound identification. Mass recalibration is an important preprocessing step before performing any analysis so as to eliminate these mass shifts. In Section 4.2, we briefly report the different recalibration methods available for mass spectrometry data. To correct these mass shifts in MSI data, we present our lock mass-free recalibration method in Section 4.3 [1]. In Section 4.4, we evaluate the performance of our method by applying it to correct mass shifts present in three MSI datasets.

4.1 Mass shifts and their correction

With advancements in technology, performance of mass spectrometers have considerably improved in terms of mass resolving power and mass accuracy. However, systematic errors still get introduced leading to mass mis-assignments also known as *mass shifts*, whereby the mass spectrum of the ejected ions exhibits peaks that do not have the correct assignment of m/z . These mass shifts can be positive or negative in nature [168, 169]. It is important to limit these systematic mass measurement errors. *Mass calibration* is a critical parameter to be considered in order to correct mass shifts and achieve high mass accuracy.

Every mass spectrometer requires some calibration prior to an experiment, a process commonly known as *external calibration.* In this form of calibration, several peaks of known standard compounds are evenly distributed over the mass range of interest, hence creating a mass reference list. Calibration is then performed by recording a mass spectrum of the standard compound and subsequent correlation of experimental m/z values to the mass reference list [35, 170]. Another protocol used for calibration is *internal calibration.* This form of calibration usually requires the introduction of an internal mass reference of known molecular formula, commonly known as the *lock mass.* It is introduced separately using a second inlet system in the mass spectrometer or together with the sample before analysis [35, 41, 171–173]. It should be taken care that the lock mass peak does not suppress the analyte peaks and vice-versa.

Even then, mass accuracy can be compromised due to different experimental factors like outdated calibration coefficients, ion intensity and temperature changes during the measurement, charge accumulation, etc. For an MSI experiment, things can be more complex, specifically when the sample under investigation is not flat. This causes absorption variation between different histological areas of the biological sample. Due to this, the measured mass of the same ion may not be exactly the same in the spectra acquired from different regions of the sample [2, 107, 174, 175]. Just, internal or external calibration is not ideal for a MSI experiment because we do not have a standard reference spectrum that can be used for the entire section that has been imaged. Secondly, if a lock mass is used, then this mass peak may not be present in all the spectra acquired at discrete spatial positions of the imaging surface. This makes it difficult to calibrate the spectra that do not contain the lock mass peak. Also, at times, the use of multiple internal calibrants to cover the complete mass range can overlap the masses of interest and may preclude their detection.

Hence, in order to eliminate mass shifts in MSI data spectral recalibration becomes an important preprocessing step. The concept of spectral recalibration basically is to use the hypothetical knowledge of the sample under investigation so as to compute an accurate calibration function [6]. Unlike external or internal calibration, which are performed before starting the experiment, spectral recalibration is performed after data acquisition as an additional data preprocessing step.

4.2 Previous contributions

There are different spectral recalibration approaches that have been reported for both MS and tandem MS (MS/MS) data in the literature.

A recalibration method developed by Kozhinov *et al.* [176] estimates a mass calibration function using *m/z* ratios and abundances of internal calibrants. Only monoisotopic species are used as internal calibrants and a high number of these calibrants are obtained by using an iterative approach, so as to cover the entire mass range. The authors have applied this method to analyze petroleum samples acquired using orbitrap fourier transform mass spectrometry and achieved sub-ppm level mass accuracy for the evaluated datasets.

Barry *et al.* [177] have developed a frequency shift-based recalibration approach that they have applied to fourier transform ion cyclotron resonance MSI data. The method corrects systematic mass shifts introduced due to space-charge effects observed while using fourier transform mass spectrometry instruments. The method applies cyclotron frequency correction which is determined based on the frequency shift of polydimethycyclosiloxanes ions, that are found in ambient laboratory environment.

Another method proposed by Petyuk *et al.* [175] is developed to reduce mass shifts in liquid chromatography mass spectrometry-based-proteomics data. The method uses *a priori* knowledge of the sample being analyzed as well as employs non-parametric regression models so as to incorporate additional explanatory variables such as elution time, *m/z* and ion intensity. The same group has developed yet another method that performs multidimensional recalibration for highly complex mixtures [178]. This method identifies a subset of effective calibrants for recalibration by statistically matching measured masses to accurate masses of putative known compounds likely to be present in the mixture that is being analyzed.

A recalibration method proposed by Gobom *et al.* [179] is specifically developed for MALDI-TOF data. In their work, they pointed out that delayed ion extraction in MALDI-TOF often distorts the linear relationship between *m/z* and square of ion flight time. This leads to the consequence that the calibrants have to be present close enough to the analyte signal in order to achieve high mass accuracy, ultimately leading to the requirement of several calibrant signals and a higher-order calibration function. As specified earlier, increasing the calibrant signals may lead to analyte signal suppression. To

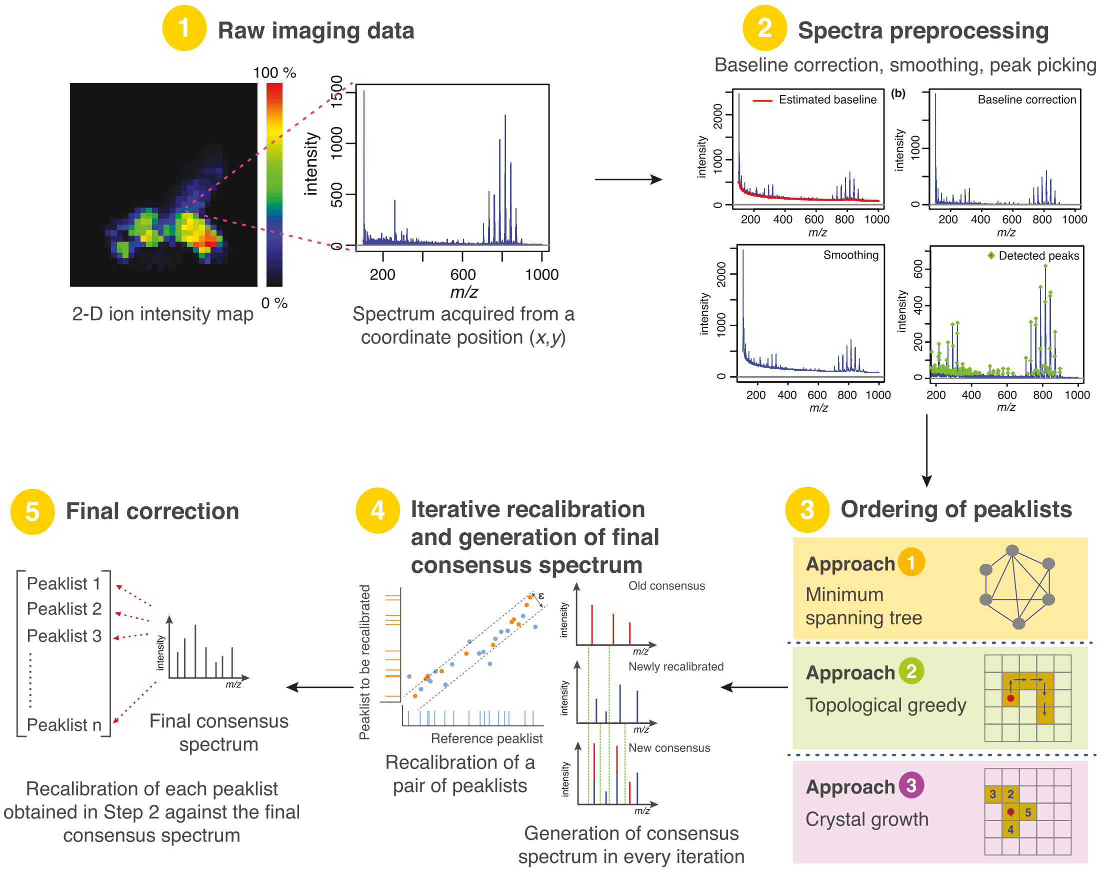


Figure 4.1: Schematic illustrating our lock mass-free recalibration method. This 5-step recalibration method is generic as it can be applied to different MSI datasets recorded using different types of mass spectrometers, equipped with different ion sources.

address these issues, Gobom *et al.* propose an approach that combines the use of external polynomial calibration and a first order internal correction to improve the accuracy of MALDI-TOF data.

Wolski *et al.* [180] have developed two recalibration approaches that they have applied to correct spectra of protein samples resulting from two-dimensional gel electrophoresis separation. These two methods do not use any external or internal calibrants, instead the method is based on the use of peaklist similarities to perform recalibration.

A complete work flow comprising automated recalibration and processing of tandem mass spectra is developed by Stravs *et al.* [181]. This method generates calibration curves by setting intensity-cutoffs to perform recalibration of MS/MS and MS^n spectra. The high quality spectra of standard compounds generated post-recalibration and processing can then be uploaded to the MassBank [182] repository using this workflow.

4.3 Lock mass-free recalibration

In this section we present our lock mass-free computational recalibration method to reduce mass shifts present in MSI data. As mentioned earlier, MSI data does not have a single reference spectrum against which all acquired spectra can be recalibrated. Our method generates a reference spectrum, termed here as a *consensus spectrum*, using all the spectra present in the dataset. An overview of our method is shown in Figure 4.1. Before applying our recalibration method to the MSI dataset, we first subject the raw data to the following preprocessing steps - Baseline correction, smoothing and peak picking (refer to Section 3.3 for detailed explanation of these steps), to obtain peaklists. Our method is then applied to these peaklists in order to generate a consensus spectrum. Our recalibration method can be summarized in three steps:

1. We first decide on an order to process the peaklists. This can be achieved using one of the three approaches:
    (a) Minimum spanning tree (MST)
    (b) Topological greedy (TG)
    (c) Crystal growth (CG)

    These ordering approaches take into account pairwise spectral similarity that has been explained in Section 4.3.1.
2. We then perform recalibration along this order, in an iterative manner, to generate a consensus spectrum.
3. This consensus spectrum acts a reference against which we recalibrate all the preprocessed peaklists.

4.3.1 Spectral similarity measure

To achieve high-quality results while recalibrating a peaklist against another, several common masses should be present in a user-defined region (δ), between the two peaklists, covering a wide mass range. For this, our method exploits spectral similarity to generate a ranking order and perform recalibration. We use a distance function d to find local pairwise peaklist similarity. It is defined as:

$$d_{(x,y)(x',y')} = \frac{1}{p_{(x,y)(x',y')} \cdot r_{(x,y)(x',y')}} \tag{4.1}$$

where $p_{(x,y)(x',y')}$ is the peak counting score between the two peaklists $P_{(x,y)}$ and $P_{(x',y')}$. The mass range of the common peaks amongst the two peaklists is represented using $r_{(x,y)(x',y')}$. Lower the value of the distance function, more is the similarity between the peaklist pair.

4.3.2 Spectral ordering approaches

Before performing recalibration, it is important that all the peaklists are ordered in a specific manner. This especially is helpful to minimize the error within the growing consensus spectrum. For this, we describe three peaklist ordering approaches:

4.3.2.1 Minimum spanning tree ordering

In Minimum Spanning Tree (MST) ordering, we use local pairwise peaklist similarity (see Section 4.3.1), to generate an order of ranking the peaklists. We calculate distances between all pairs of peaklists using the distance function (see Equation 4.1), to generate a distance matrix.

The distance matrix can be represented as a complete weighted graph $G(V,E)$, where the vertex set V corresponds to individual peaklists and the edges E are weighted by the distance corresponding to peaklist pairs, in the distance matrix. We then calculate

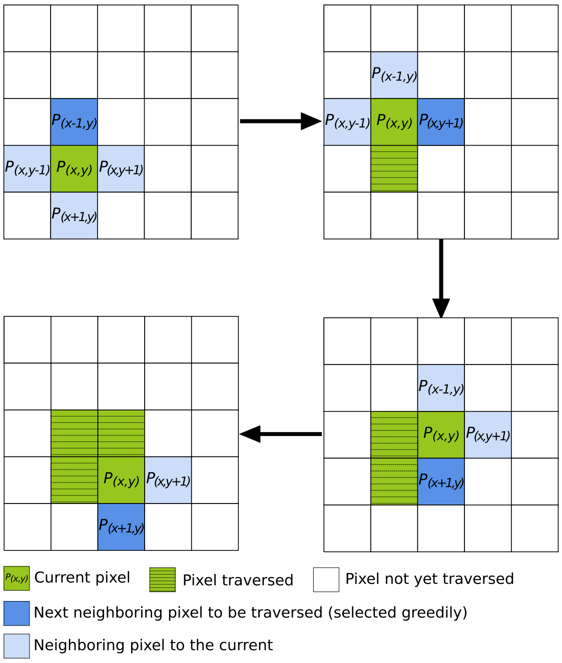

P(x-1,y)
P(x,y-1)
P(x,y)
P(x,y+1)
P(x+1,y)
Current pixel
Pixel traversed
Pixel not yet traversed
Next neighboring pixel to be traversed (selected greedily)
Neighboring pixel to the current

Figure 4.2: Representation of the steps involved in Topological Greedy ordering. To order the peaklist using TG ordering approach, the peaklist with the highest pixel score is the starting point. The next peaklist to be traversed is added greedily by selecting the immediate neighboring pixel that has the smallest distance to the current pixel. This iterative process continues till all the pixels in the dataset have been traversed.

a minimum spanning tree. To connect all the vertices in the graph with their edges, the Dijkstra-Prim algorithm [183] is used. Recalibration of spectra proceeds along this tree.

We start with the pair of peaklists that have the minimum distance in the spanning tree. We then add the edge with the next smallest distance that is connected to the current subtree, to the growing subtree, in an iterative manner. The above process iteratively adds one peaklist at a time until all peaklists have been added. In each step, we create a consensus spectrum which is used in the next step for recalibration. The process to recalibrate the next peaklist against the consensus spectrum generated in the previous iteration is discussed in Section 4.3.3. This process continues until all the peaklists have been added and a final consensus spectrum is generated. The procedure to generate the consensus spectrum is discussed in Section 4.3.4.

4.3.2.2 Topological greedy ordering

In a MSI dataset, for many neighboring spatial positions (also known as *pixels*) on the imaged sample, the spectra obtained most likely represent similar molecular composition, and thus should be similar [112]. We have developed Topological Greedy (TG) ordering based on this concept. We assume that mass deviation in spectra obtained from neighboring pixels should also be similar.

To order the peaklists based on the above assumption, we first calculate a score (termed as *pixel score*) for the peaklist acquired at every single coordinate position (say peaklist $P_{(x,y)}$ acquired at a specific coordinate position (x,y)) based on similarity with its four neighboring peaklists ($P_{(x,y-1)}$, $P_{(x,y+1)}$, $P_{(x-1,y)}$ and $P_{(x+1,y)}$). We calculate the similarity between peaklist $P_{(x,y)}$ and all the four neighboring peaklists individually, using the distance function defined in Equation 4.1 (see Section 4.3.1).

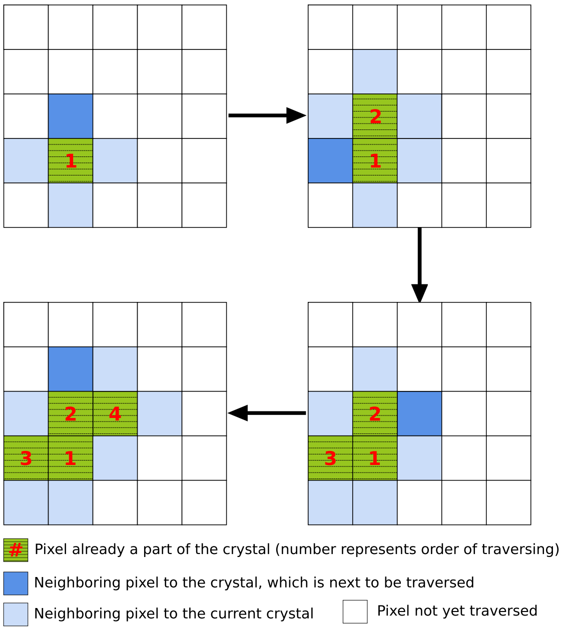

1
2
1
2
1
3
2
4
3
1
Pixel already a part of the crystal (number represents order of traversing)
Neighboring pixel to the crystal, which is next to be traversed
Neighboring pixel to the current crystal
Pixel not yet traversed

Figure 4.3: Representation of the steps involved in Crystal Growth ordering. To order the peaklist using CG ordering approach, the peaklist with the highest pixel score is the starting point. The next peaklist to be traversed is added by selecting the non-traversed neighboring pixel of the growing crystal that has the smallest distance to the crystal. This iterative process continues till all the pixels in the dataset have been traversed.

The pixel score calculated for peaklist $P_{(x,y)}$ is represented as:

$$Score_{P_{(x,y)}} = \frac{d_{(x,y)(x,y-1)} + d_{(x,y)(x,y+1)} + d_{(x,y)(x-1,y)} + d_{(x,y)(x+1,y)}}{4} + i_{P_{(x,y)}} \tag{4.2}$$

where $i_{P(x,y)}$ corresponds to the intensities of the peaks in $P_{(x,y)}$.

Based on these pixel scores, all peaklists are ordered such that the peaklist with the highest score is placed at the top of the list. This peaklist is the starting point of our TG ordering and is considered as the initial consensus spectrum. To proceed to the next pixel, we select the peaklist with the smallest distance d from the current pixel, greedily from the four neighboring pixels. Distance d is calculated using Equation 4.1 (see Section 4.3.1). A step-wise pictorial representation of how pixels are traversed in TG ordering is shown in Figure 4.2. We recalibrate the new peaklist against the current consensus spectrum and merge the two to generate a new consensus spectrum, in every iteration. We repeat this process until we reach a dead-end, where all neighbors of the current pixel have been traversed. In this case, we select a new starting point as described above, among all the pixels not yet traversed. This process continues till all the pixels have been traversed.

4.3.2.3 Crystal growth ordering

The concept of CG ordering is based on the same assumption as described for TG ordering (see Section 4.3.2.2). In this approach we score the peaklist acquired at every single coordinate position using Equation 4.2. The peaklist with the highest score is the starting point to build the crystal and is the initial consensus spectrum. To proceed to the next pixel, we select the peaklist with the smallest distance d, from the four neighboring pixels. This new pixel now becomes a part of the crystal. We now look for

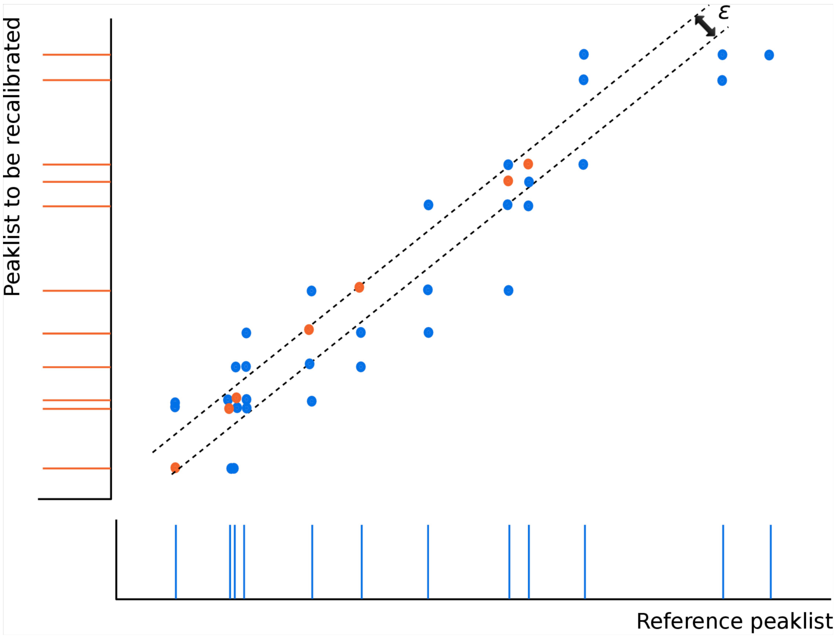


Figure 4.4: Geometric interpretation of linear recalibration using Maximum Line Pair Stabbing algorithm. The algorithm uses a set of mass values where *blue* represents masses from the reference peak list and *orange* represents masses from the peaklist that has to be recalibrated. The algorithm then finds a pair of parallel lines within distance ε from each other such that the number of input points that intersect (stab) the area between the two lines is maximized. Figure reproduced from [6].

all non-traversed neighbors of this crystal and calculate distance d for each neighboring pixel. The next pixel to be traversed is the one with the smallest distance to the crystal. Distance d is calculated using Equation 4.1(see Section 4.3.1). A step-wise pictorial representation of how pixels are traversed in CG ordering is shown in Figure 4.3.

In every iteration, we add a new neighboring pixel to the growing crystal, recalibrate the peaklist corresponding to this pixel against the current consensus spectrum and merge the two to generate a new consensus spectrum. This process continues till all the pixels are added to the growing crystal.

4.3.3 Recalibrating a pair of peaklists

For robust recalibration it is very important to detect and remove outliers, since recalibration can easily be corrupted if two peaks are wrongly matched. We detect outliers and perform recalibration for a pair of peaklists using the *Maximum Line-Pair Stabbing* (MLS) algorithm proposed by Böcker and Mäkinen [6].

Given a list of mass values, obtained from the two peak lists, this combinatorial and deterministic approach, uses a computational geometry interpretation of the problem to find a pair of parallel lines within a constant distance, epsilon (ε), such that, the number of points in space between the two lines is maximum (see Figure 4.4). The value of epsilon (ε) can be estimated using on the measurement divide and other conditions. Ordinary least squares regression is then used on these subset of points, to fit a regression line. We finally use this regression line to recalibrate the second peak list, using the first as a reference.

This recalibration procedure is applied in an iterative manner to the ordered peaklists. A consensus spectrum is generated at every iteration (see Section 4.3.4), which is used

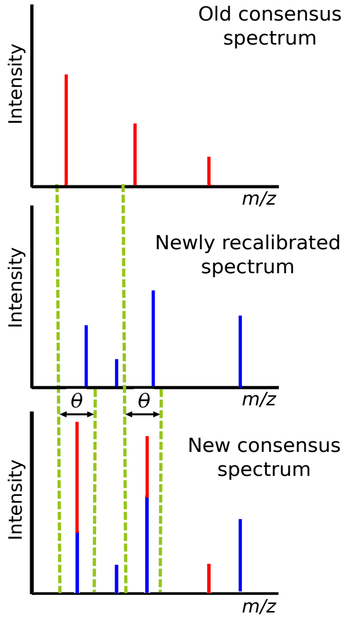


Figure 4.5: Representation of the generation of consensus spectrum. This approach involves merging matching peaks between old consensus spectrum and the newly recalibrated spectrum in a defined mass threshold $\boldsymbol{\theta}$, to generate a new consensus spectrum

to recalibrate the next peaklist in the ordered list. This process continues till all the peaklists are traversed and a final consensus spectrum is generated.

4.3.4 Generation of consensus spectrum

As mentioned in the previous sections, in each step of pair-wise peaklist recalibration, a consensus spectrum is generated. We perform this using the following approach: Initially, the first peaklist (peaklist at the top of the ordered list) is assumed as the consensus spectrum (this acts as a reference). Against this consensus spectrum, we recalibrate the next spectrum in the list using the MLS algorithm described in Section 4.3.3. After recalibration, the old consensus spectrum (assumed initially) and the newly recalibrated spectrum are merged using the following steps:

Lets assume that a peak in the old consensus spectrum with mass m and intensity i matches with a peak in the newly recalibrated spectrum with mass m' and intensity i', i.e. $|m - m'| \leq \boldsymbol{\theta}$ for some user-defined mass threshold $\boldsymbol{\theta}$. We then update the mass and intensity in the consensus spectrum using the following equation:

$$m_{new} = \frac{i \cdot m + i' \cdot m'}{i + i'} \qquad i_{new} = i + i' \tag{4.3}$$

If there is no peak in the newly recalibrated spectrum within the selected mass threshold $\boldsymbol{\theta}$, then the peak with mass m and intensity i, in the old consensus spectrum is added as it is, to the new consensus spectrum. The same holds true when a peak exists in the newly recalibrated spectrum, but not in the old consensus spectrum. The generation of a new consensus spectrum is represented diagrammatically in Figure 4.5.

This new consensus spectrum is used in the next iteration to perform recalibration. We continue this process until all peaklists have been traversed. In the last iteration, we receive the final consensus spectrum. This final consensus spectrum generated does not necessarily contain the correct masses of the analytes present in the sample, but it

Table 4.1: Summary of imaging datasets used in this study.

Name	Mode of m/z	Laser frequency	Laser shots per spot	Mass range	Total spectra
Dataset 1	LDI	10 Hz	50	100.00-1000.00	1184
Dataset 2	MALDI	15 Hz	120	100.00-1000.00	2109
Dataset 3	MALDI	15 Hz	90	100.00-1000.00	3408

contains average masses for peaks repeatedly observed in the individual spectra. Using intensity values of mass peaks in merging the two peaklists assures that the mass of intense peaks is given higher confidence than low-intensity peaks.

4.3.5 Final correction

In the last step of our recalibration method, we use the final consensus spectrum (see Section 4.3.4) as a reference to recalibrate all the peaklists in the MSI dataset. To perform the final correction, every peaklist is independently recalibrated against the final consensus spectrum using the MLS algorithm described in Section 4.3.3. This final consensus spectrum spans the entire collected mass range of the imaging dataset on which recalibration is performed.

4.4 Evaluation of mass shift correction

In this section, we evaluate the recalibration method for its performance and accuracy and also present a comparative evaluation of the three spectral ordering approaches discussed in Section 4.3.2.

4.4.1 Experimental datasets

We apply our recalibration method to three datasets acquired using MSI of whole-body *Drosophila melanogaster* flies (see Table 4.1). Details related to the insect treatment, chemicals used and sample preparation can be found in Appendix Section A.1.

The MSI experiments corresponding to all the three datasets were performed using MALDI Micro MX (Waters, www.waters.com, Milford, MA, USA) operated in a reflectron mode with the acceleration and plate voltages at -12 kV and +5 kV, respectively. Delayed extraction time was 500 ns. Desorption and ionization was realized by nitrogen (337 nm) UV laser MNL 103-LD (LTB Lasertechnik Berlin GmbH, www.ltb-berlin.de, Germany). Matrix ions were suppressed with a low mass cut-off set at 150 Da.

All the samples were imaged using a step size of 100 μm (lateral resolution of 254 dpi). The number of laser shots per spot was optimized and set to different values as shown in Table 4.1. The range of the measured masses was set from 100 - 1000 Da.

4.4.2 Data acquisition and processing

The data was acquired using the MassLynx 4.0 software (Waters, www.waters.com, Milford, MA, USA) and processed with custom-made software MALDI Image Converter (Waters) to obtain spatially differentiated data in Analyze 7.5 imaging file format.

Before applying our recalibration method we subject the datasets to certain preprocessing steps. We import the acquired imaging dataset in MATLAB (MATLAB and Statistics Toolbox Release R2012b, The MathWorks, Inc., Natick, Massachusetts, United States) and parse the image data using `analyze75read`() routine. We then apply the

Table 4.2: Summary of specific ions of interest used in each dataset along with the recalibration parameters used by the three ordering approaches.

Dataset	Compound of interest	Form of adduct	Ion atomic composition	Exact mass (Da)	Mass used for recalibration (Da)[a]	Parameters		
						ε^b (Da)	δ^c (Da)	θ^d (Da)
Dataset 1	cVA Triacylglycerol	$[M+K]^+$ $[M+K]^+$	$C_{20}H_{38}O_2K$ $C_{49}H_{92}O_6K$	349.25034 815.65255	349.40 ± 0.30 815.80 ± 0.35	0.5	0.5	0.5
Dataset 2	DHB Tricosadiene	$[M+K]^+$ $[M+Li]^+$	$C_7H_6O_4K$ $C_{23}H_{44}Li$	192.98977 327.35976	193.15 ± 0.30 327.15 ± 0.10	0.2	0.2	0.2
Dataset 3	DHB Triacylglycerol	$[M+K]^+$ $[M+K]^+$	$C_7H_6O_4K$ $C_{49}H_{92}O_6K$	192.98977 815.65255	193.15 ± 0.30 815.80 ± 0.30	0.2	0.2	0.2

[a] m/z values for recalibration were chosen based on the observed ion abundances in a specific mass range for a particular analyte in the dataset;
[b] Mass threshold used for Maximum Line-Pair Stabbing algorithm to perform recalibration;
[c] Mass threshold to find common masses in a pair of peaklists;
[d] Mass threshold used while merging the old consensus spectrum and the newly recalibrated spectrum.

following preprocessing steps: baseline correction using `msbackadj`(), then smoothing using `mslowess`() routine and at last peak picking using `mspeaks`() routine, to generate peaklists. We then export the preprocessed data as peaklists along with their coordinate position information to individual text files.

4.4.3 Recalibration accuracy and performance

We evaluated the performance of our method for two analytes of interest in each of the three imaging datasets. The details of all the ion species of interest used for each dataset along with the parameters used to perform the recalibration are listed in Table 4.2. The chosen analytes for dataset 1 are the male anti-attractant 11-*cis*-vaccenyl acetate (cVA) [184], with confirmed biological significance [185] and male-specific triacylglycerol [186], both in their potassiated form. For dataset 2, we used potassium adduct of 2,5-dihydroxybenzoic acid (DHB) matrix ion and lithium adduct of tricosadiene [184]. Dataset 3 included the same potassiated DHB and triacylglycerol analyte, as used in dataset 2 and dataset 1 (see Table 4.2).

Representation of results We demonstrate the mass shift correction for each selected analyte of interest using 2-D pseudo color plots. We generate these plots in a selected mass window of ± 0.4 Da for dataset 1 and a mass window of ± 0.3 Da for dataset 2 and 3. These mass windows have been chosen based on the maximum shift observed in the raw imaging datasets.

For evaluation, we construct two types of plots, one which shows the observed mass shift for the selected mass in the preprocessed data and the second showing mass shift correction generated using the recalibrated peaklists obtained after applying our method. The 2-D plots are generated using a customized version of `plotImsSlice.R` function available in the `MALDIquant` R package.

Observed mass shift for selected ions We observed that the mass shifts for the selected analytes (shown in Table 4.2) ranged between 0.15 - 0.25 Da, from their exact masses. These masses used for recalibration were chosen by an expert, based on the observed ion abundances, in a selected range, from the experimental data. The ion abundances for each analyte were selected based on the 3-D profile of the imaged area and were studied using 2-D ion intensity maps generated using the raw imaging data. To confirm the abundance of these analytes, we generated 2-D ion intensity maps using BioMap software (Novartis, Basel, Switzerland) and we found the distribution to be

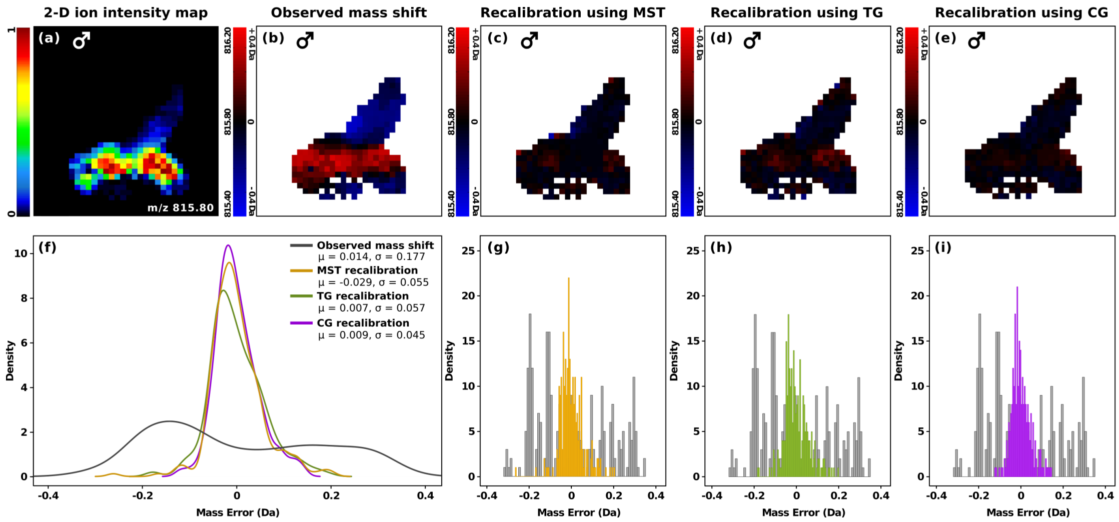


Figure 4.6: Demonstration of mass shift and its correction using data obtained during LDI-MSI of six-day-old virgin male *D. melanogaster* fly in lateral position. **(a)** 2-D ion intensity map of triacylglycerol *m/z* 815.80 ([M+K]$^+$), signal intensity in voxel view with rainbow-color scale, in a mass window size of ± 0.4 Da. **(b)** 2-D pseudo color plot of the observed mass shift with PlusMinus 1 color scale (*red* represents a positive mass shift and *blue* represents a negative mass shift, the pixels in *black* signify no mass shift, whereas the pixels in *white* correspond to no signal from that specific pixel in the selected mass range). Mass shift correction by the recalibration method **(c)** using MST ordering, **(d)** using TG ordering, **(e)** using CG ordering. **(f)** Pseudo density plot of the mass shift observed in preprocessed data (*gray*) and its correction (MST ordering - *yellow*, TG ordering - *green*, CG - *violet*). **(g - i)** Mass error distribution histograms of the preprocessed data before and after recalibration using MST, TG and CG ordering respectively (Observed mass error - *gray*, correction using MST ordering - *yellow*, correction using TG ordering - *green*, correction using CG ordering - *violet*).

identical. The analyte signals were evaluated considering the fly's body parts and its biological significance, as suggested by the expert.

Dataset 1 For this dataset, we apply our recalibration method to correct mass shifts observed for cVA ion at *m/z* 349.40 [(M+K)$^+$] and triacylglycerol ion at *m/z* 815.80 [(M+K)$^+$]. Results of the observed mass shift and its correction for *m/z* 815.80 [(M+K)$^+$] are shown in Figure 4.6.

As can be seen in Figure 4.6(a), triacylglycerol is seen to be distributed in the abdomen, thorax and the wing region of the fly body. The abdomen and the thorax region represent a high intensity distribution whereas the wing region and the region below the thorax represent low intensity distribution of the *m/z* 815.80 [(M+K)$^+$]. The signals observed in area below thorax show the distribution of triacylglycerol originating from damaged tissues around cut-off legs. Figure 4.6(b) shows a high positive mass shift, a maximum of 0.4 Da, on the thorax and the tip of the abdomen. The abdominal tip, wing region and the region near the legs shows a negative mass shift of about 0.15 - 0.3 Da.

The 2-D pseudo color plots of the mass shift correction are shown in Figures 4.6(c-e). As can be seen, the CG approach for ordering the peaklists performs better mass shift correction as compared to MST and TG ordering. For TG ordering, the 2-D pseudo color plot of mass shift correction shows several pixels in red in the central abdomen, thorax and tip of the wing indicating a positive mass shift in the range of about 0.15 - 0.25 Da for these pixels and a few pixels in blue representing a negative mass shift of about 0.15 - 0.2 Da.

Table 4.3: Mean mass errors (ME) and standard deviations (SD) for all masses of interest calculated before and after recalibration.

Dataset	Mass used for recalibration (Da)	Before recalibration		MST ordering		TG ordering		CG ordering	
		ME (Da)	SD (Da)	ME (Da)	SD (Da)	ME (Da)	SD (Da)	ME (Da)	SD (Da)
1	349.40 ± 0.4	+ 0.062	± 0.114	+ 0.062	± 0.067	+ 0.050	± 0.068	+ 0.052	± 0.069
	815.80 ± 0.4	+ 0.014	± 0.177	- 0.029	± 0.055	+ 0.007	± 0.057	+ 0.009	± 0.045
2	193.15 ± 0.3	+ 0.103	± 0.047	- 0.014	± 0.029	- 0.007	± 0.028	- 0.017	± 0.028
	327.15 ± 0.3	- 0.007	± 0.038	- 0.002	± 0.015	+ 0.013	± 0.026	+ 0.016	± 0.028
3	193.15 ± 0.3	+ 0.141	± 0.054	+ 0.101	± 0.046	+ 0.232	± 0.053	- 0.033	± 0.128
	815.80 ± 0.3	- 0.006	± 0.141	- 0.017	± 0.059	+ 0.001	± 0.120	- 0.045	± 0.078

There is a significant difference observed in the 2-D pseudo color plots generated before and after recalibration and the range of mass error distribution is seen to be reduced. This is evident from the pseudo-density plot shown in Figure 4.6(f), which compares the performance of the three peaklist ordering approaches. Plots 4.6(g-i) represent the mass error distribution histogram for the preprocessed data before and after recalibration, using MST, TG and CG ordering, respectively. The pseudo density plot shows that CG ordering displays the least standard deviation of the three ordering approaches. The mass error distribution histograms in the Figure 4.6(g-i) support this observation.

A similar observation was also found for the mass shift correction performed for the distribution of cVA ion at *m/z* 349.40 $[(M+K)^{+}]$, results for which are discussed in Appendix Section A.2.

Dataset 2 For the second dataset, we evaluated the performance of the recalibration method on the distribution of DHB matrix ion, at *m/z* 193.15 $[(M+K)^{+}]$ and tricosadiene at *m/z* 327.15 $[(M+Li)^{+}]$. Results for this dataset are discussed in Appendix Section A.3.

Dataset 3 For the third dataset, we evaluated the performance of the recalibration method on the distribution of DHB matrix ion, at *m/z* 193.15 $[(M+K)^{+}]$ and triacylglycerol at *m/z* 815.80 $[(M+K)^{+}]$. Results for this dataset are discussed in Appendix Section A.4.

Mean mass error and standard deviation We also compared the average mass and standard deviation (SD) before and after applying our recalibration method. This comparison was performed for the three datasets using all the three ordering approaches, results for which can be seen in Table 4.3. For the first dataset, recalibration using MST, TG and CG ordering performed equally well for *m/z* 349.40 $[M+K]^{+}$ by reducing the mass error distribution as observed from 0.114 Da before recalibration to 0.067 - 0.069 Da after recalibration. However, for *m/z* 816.0 $[M+K]^{+}$, which had a SD of ± 0.177 Da before recalibration, recalibration using CG ordering results in a slightly better SD of 0.045 as compared to MST ordering with SD 0.055 and TG ordering with SD of 0.057. Table 4.3 also shows the better performance of applying CG ordering and in some cases MST ordering for the mass values in the other two datasets, used in this study. The SD values clearly indicate an advantage of using CG as well as MST ordering for recalibration, over TG ordering.

Ion intensity distribution We also evaluated the change in ion intensity distribution before and after recalibration. Since, after applying our recalibration method, mass shift for the analytes of interest was considerably reduced, we expected to see an increase in the number of pixels for the selected analyte in a narrower mass window.

A comparison of the 2-D ion intensity maps representing the distribution of *m/z* 349.40 $[M+K]^{+}$ for dataset 1, before and after performing recalibration using CG ordering is

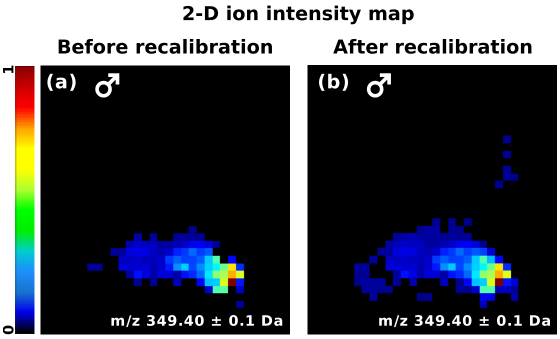


Figure 4.7: Comparison of ion intensity distribution using preprocessed data before and after recalibration. **(a)** 2-D ion intensity map before recalibration representing the distribution of m/z 349.40 $[M+K]^+$ for dataset 1 using the rainbow-color scale in a lower mass accuracy threshold of ± 0.1 Da. **(b)** 2-D ion intensity map after recalibration for m/z 349.40 $[M+K]^+$ in the same threshold shows increase in the number of pixels, representing a reduction in the mass shift error.

shown in Figure 4.7. We used a lower mass accuracy threshold of ± 0.1 Da to represent this comparison. As can be seen in Figure 4.7(b), additional pixels are present as a part of the fly body as compared to Figure 4.7(a), representing a reduction in the mass shift error after applying our recalibration method.

4.4.4 Evaluation of running times

We have implemented our recalibration method in Java, version 1.7.0-65 (Sun Developer Network). In Table 4.4, we report the running times of our recalibration method presented as a comparison of the three ordering approaches. All the evaluations have been performed on a 64-bit laptop, equipped with Intel i5 2.67 GHz quad-core processor and 3.5 GB RAM. The values reported here only apply for the recalibration method, without including the time required for initial preprocessing of mass spectra to generate peaklists, that has been performed using MATLAB.

Table 4.4: Running times per recalibration for the three imaging datasets measured on a 64-bit laptop, equipped with Intel i5 2.67 GHz quad-core processor and 3.5 GB RAM.

Datasets	Total spectra	MST recalibration	TG recalibration	CG recalibration
Dataset 1	1184	2 min 10 s	1 min 2 s	2 min 5 s
Dataset 2	2109	36 min 26 s	1 min 47 s	6 min 6 s
Dataset 3	3408	3 h 44 min	26 min 15 s	46 min 43 s

Using, CG ordering results in higher running times as compared to TG ordering however, it takes less running time when compared to MST ordering. Based on the results discussed in the previous section, it can be said that CG ordering seems to be slightly more robust for recalibration and requires moderate running time.

Chapter 5

MSICorrect: Mass spectrometry imaging data recalibration tool

In this chapter, we present `MSICorrect`, a Java-based tool that implements the lock mass-free recalibration method described in Chapter 4. In Section 5.1, we provide an overview of the tool layout, the main user window and its implementation. In Section 5.2 and Section 5.3, we describe the different input file formats and the user-defined parameters necessary to perform recalibration. We then describe the functionalities to visualize the recalibration results and export the recalibrated spectra and plots in Section 5.4.

5.1 Graphical user interface

The interface for `MSICorrect` is built using Swing components in the Java programming language. The back-end recalibration method as described in Chapter 4 is implemented in Java, version 1.7.0-65 (Sun Developer Network). The tool also uses scripts implemented in R programming language (version 3.2) as well as available R packages for certain file reading and visualization functionalities. The integration between Java and R is established using the `Rserve` [187] package, which acts as a socket server that allows applications written in various other programming languages to use facilities of R. `MSICorrect` is platform-independent and has been successfully tested to work with systems running Linux, Windows and Mac OS X. An overview of the tool and its requirements is presented in Table 5.1.

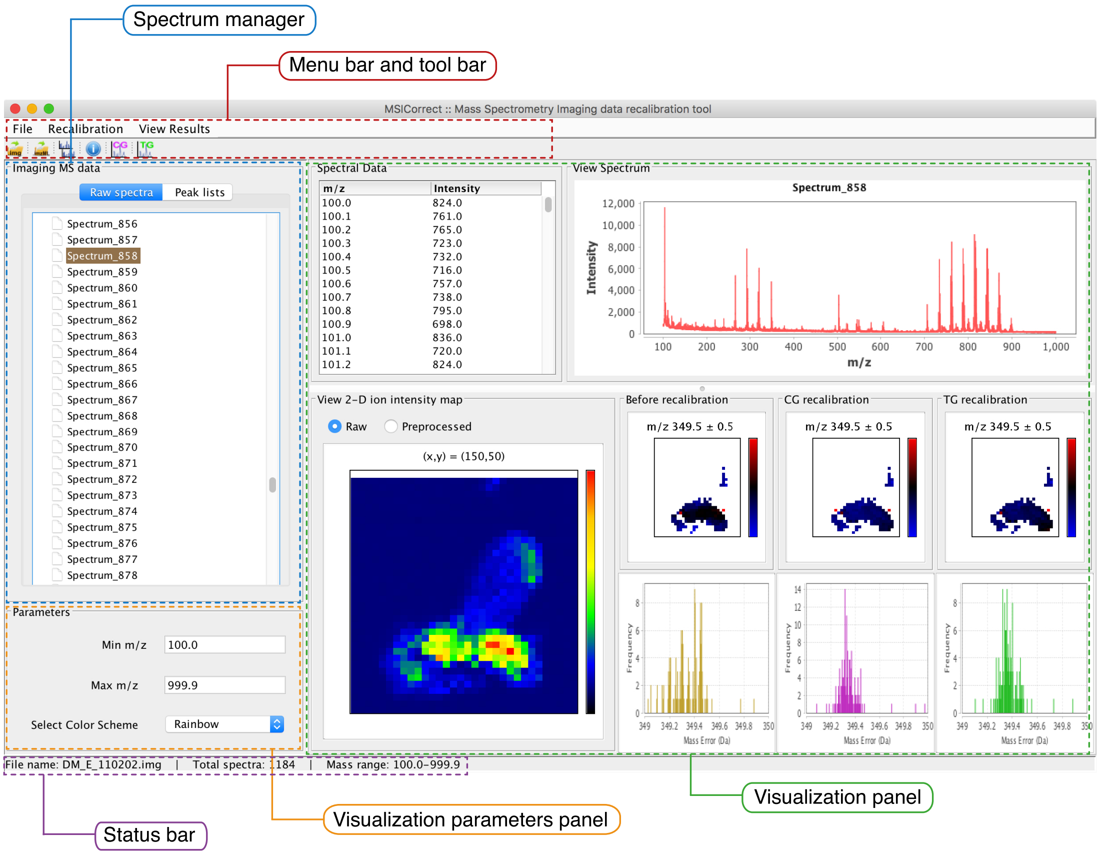
Table 5.1: An overview and list of requirements for `MSICorrect`.

Tool name: MSICorrect
Functionality: Recalibration and visualization of mass spectrometry imaging data
Operating system(s): Platform independent
Programming language: Java (JRE ≥ version 1.7 to run the tool)
Other requirements: R statistical package (≥ version 3.2.4)
Distribution: Executable `.jar` file

The main GUI window is divided in five main parts, as can be seen in Figure 5.1. Each of this is explained below briefly:

1. **Menu bar and tool bar:** Displays options for data file import, view input file properties, recalibration using Crystal Growth (CG) and Topological Greedy (TG) ordering (details on these ordering approaches are explained in Section 4.3.2) and export of recalibration results.

2. **Spectrum manager:** It mainly contains two tabs named: `Raw spectra` and `Peak lists`. When an imaging file in imported in `MSICorrect`, the spectrum manager lists all the raw spectra under the `Raw spectra` tab. Once the preprocessed peak-lists for the corresponding raw file are imported, they are listed under the `Peak lists` tab.



Spectral Data

m/z	Intensity
100.0	824.0
100.1	761.0
100.2	765.0
100.3	723.0
100.4	732.0
100.5	716.0
100.6	757.0
100.7	738.0
100.8	795.0
100.9	698.0
101.0	836.0
101.1	720.0
101.2	824.0

Min m/z: 100.0
Max m/z: 999.9
Select Color Scheme: Rainbow

File name: DM_E_110202.img | Total spectra: 1184 | Mass range: 100.0–999.9

Figure 5.1: Screenshot of the `MSICorrect` graphical user interface. The figure shows MSI data obtained during LDI-MSI of a 6-day-old virgin male *D. melanogaster* fly in lateral position. The ion intensity map shows distribution of 11-*cis*-vaccenyl acetate (cVA) at m/z 349.40 [M+K]+, observed mass shift and its correction using the implemented recalibration method.

3. **Visualization panel:** This panel is further divided in four sections. The first sub-panel displays the m/z and intensity of the spectrum number selected by the user. The second sub-panel displays a plot of the corresponding selected spectrum. The plot can be zoomed to visualize closely placed peaks. The third sub-panel displays a pseudo colored ion intensity map across the complete mass range of the dataset, when the data file is imported in the tool. The mass range and color scheme of the ion intensity map can be changed under the visualization parameters panel. Ion intensity maps for raw and preprocessed data can be visualized by clicking the respective radio buttons. The last sub-panel displays the recalibration results in the form of pseudo color plots as well as histograms. Results are displayed for the mass shift observed at a selected mass value before recalibration and the correction in this shift after performing recalibration using CG and TG ordering.

4. **Visualization parameters panel:** This panel controls the visualization of the pseudo-colored ion intensity map. The `Min m/z` and `Max m/z` parameters regulate the mass range for which the pseudo-colored ion intensity map is generated. Users can select a specific color gradient to visualize the ion intensity map using the `Select Color Scheme` drop down menu.

5. **Status bar:** It displays the name of the uploaded imaging data file, the total number of mass spectra and the mass range of the dataset.

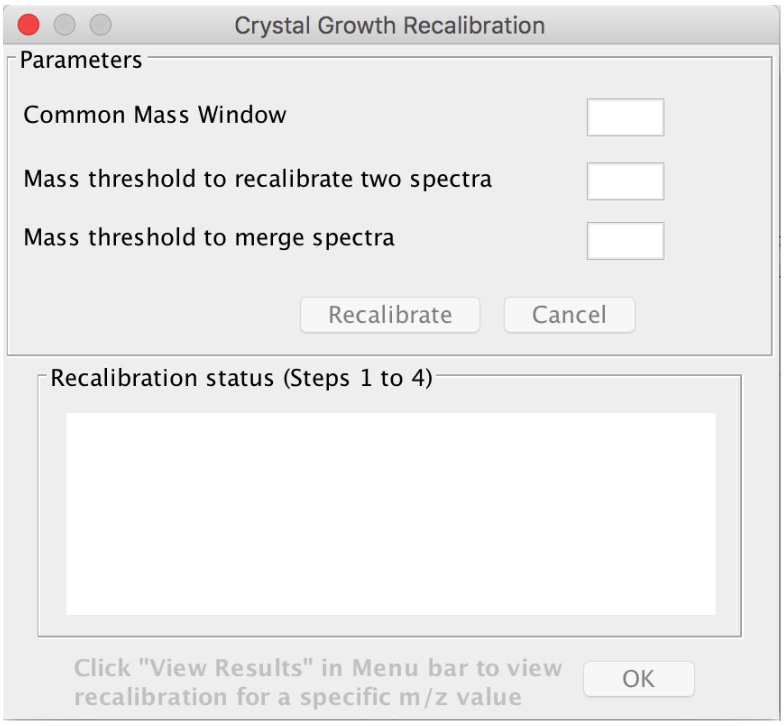

Crystal Growth Recalibration
Parameters
Common Mass Window
Mass threshold to recalibrate two spectra
Mass threshold to merge spectra
Recalibrate
Cancel
Recalibration status (Steps 1 to 4)
Click "View Results" in Menu bar to view recalibration for a specific m/z value
OK

Figure 5.2: Screenshot of the user-defined parameters window to perform recalibration. The screenshot shows the recalibration parameters where a user-input is necessary as well as the window that displays the status of the recalibration process. The parameters are identical for Crystal Growth and Topological Greedy ordering approaches.

5.2 Data import

`MSICorrect` can import common MSI data formats: ANALYZE 7.5 (Mayo Clinic) and imzML [81]. This is supported using file parsing scripts available in the R package `MALDIQuant` [90]. Along with the imaging data file, it is necessary to import preprocessed peaklists in `.txt` format corresponding to the data file. Mass shift correction is performed on these peaklists.

Data preprocessing can be performed on MSI data files using routines available in `MATLAB` (Mathworks, Inc., Natick, MA, USA) or using the `MALDIQuant` R package. After preprocessing steps, peaklists can be exported as individual text files to a specific location. These peaklist files can then be imported in `MSICorrect`.

5.3 Lock mass-free recalibration

To perform lock mass-free recalibration, two spectral ordering approaches are available in `MSICorrect`: Crystal Growth and Topological Greedy (explained in Section 4.3.2). Mass shift correction is performed on the dataset using user-defined parameters as shown in Figure 5.2.

The parameters include selecting a suitable mass window to calculate pairwise similarity between the two mass spectra, mass thresholds to recalibrate two mass spectra and to merge two spectra. To achieve best recalibration results, it is important to select appropriate mass thresholds used at various steps in the recalibration method. The selection of these thresholds mainly depends on the quality of the input dataset in terms of mass resolution and accuracy.

Once these parameters are provided by the user, recalibration is initiated using the `Recalibrate` button. The progress is displayed as intermediate steps in the panel below the parameters section. The user is notified once the process is complete. Results can be viewed using the `View Results` menu option, for the specific ordering approach selected.

5.4 Visualization of recalibration results

After completion of recalibration, the corrected mass spectra are saved at the same location where the `.jar` file is placed. Mass shift correction results can be visualized by specifying a m/z value of interest and a mass window, which is selected based on the maximum shift observed in the corresponding MSI dataset. Results are displayed as pseudo-colored plots of observed mass shift before and after recalibration based on the selected approach. These plots are represented by blue/red color (PlusMinus 1 color scale), where *red* represents a positive mass shift and *blue* represents a negative mass shift. The pixels in *black* signify no mass shift, whereas the pixels in *white* correspond to no signal from that specific position in the selected mass range. Mass error distribution histograms of the preprocessed data before and after recalibration are also plotted to visualize the mass correction performed by the method.

`MSICorrect` also provides the functionality to export the pseudo-colored plots and histograms as `.png` images for presentation and publication purposes.

The development of `MSICorrect` has been motivated by the need to create a tool which performs mass shift correction specifically for mass spectrometry imaging datasets. `MSICorrect` provides a user-friendly interface for recalibration and visualization functionalities, making it possible to easily integrate it in an existing MSI data analysis pipeline.

Chapter 6

Towards an automated method to characterize biologically relevant spatial patterns in MSI data

MSI enables simultaneous detection of thousands of known and unknown ions present in the biological sample that has been imaged. This commonly leads to the acquisition of several gigabytes (even terabytes) of multi-dimensional data from a single experiment. A typical representation of the acquired MSI data is in the form of 2-D pseudo-colored ion intensity maps to visualize the distribution of specific ions (or a selected mass range) on the biological section. Generation of ion intensity maps becomes easy if we have some prior knowledge about ions of interest in the imaged sample. However, for untargeted analysis where the aim is to extract rich information from the data, manually looking through each and every ion intensity map for biologically interesting distribution can be tedious, practically infeasible and may require expert judgment.

In this chapter, we present our first attempt to characterize biologically-relevant spatial patterns in large MSI datasets, with the final objective of automation, rather than manual analysis. The main aim of our simplistic approach is to select and rank the ion intensity maps. This ranking is based on the abundance of spatially-resolved information pertaining to structures, similar to the manner in which a naked eye would distinguish these ion images. Furthermore, our approach also aims to group ion maps with similar spatial patterns (structures) in order to obtain a list of ions that exhibit similar spatial distribution. In Section 6.1, we start with a brief introduction of the well-established field of image processing and pattern recognition. Next, in Section 6.2, we report the work that has been performed to detect structured intensity patterns, specifically for MSI data. In Section 6.3, we present our method to recognize significant spatial patterns in ion intensity maps and rank them based on a calculated score. In this section, we also discuss the steps to group ion maps based on their spatial similarity. In Section 6.4, we evaluate our method using biological MSI datasets.

6.1 Image processing and pattern extraction

Visual information is the most important type of information perceived by the naked eye, which is later processed and interpreted by the human brain. Digital image processing is an established yet rapidly growing technology that captures this visual information and evaluates or manipulates it electronically using computers. Image digitization is an approach that converts an image from its pictorial form to a matrix containing numerical data [188]. Some of the objectives of digital image processing are to detect important geometric structures present in an image, remove noise as well as enhance extracted features. This technology has played an important role in a wide variety of disciplines with applications in space research, robotics, industrial inspection, remote sensing and medical diagnosis.

Pattern extraction from an image is a method to capture the visual content in the form of local features (or patterns) present in an image. This can be performed computationally in an automated manner and has considerable application in fields like face recognition, object recognition, analyzing fingerprints, medical decision support and many more.

In our study, the main aim is to computationally exploit the overall spatio-temporal information present in an ion intensity map, in a completely unsupervised manner. This would mainly involve extracting the local patterns present within the ion intensity map in the spatial domain and then provide an overall score based on the pattern abundance.

6.2 Previous contributions

In the recent years, there has been a significant increase in the number of groups working on developing computational methods for data reduction, compression and feature identification for MSI data. There are some approaches that have been reported in this direction.

McDonnell *et al.* [189] have developed and demonstrated data reduction routines that are based on automated feature extraction for peptide, protein and lipid imaging. These routines summarize each MSI dataset by determining a list of masses representing each dataset, extracting the spatial variation of each ion above a set of user-defined peak thresholds and highlighting localized features. For feature identification, a set of mass spectral representations are calculated to distinguish all features in the imaging MS dataset; which includes the datasets mean spectrum, base peak spectrum and their TIC normalized analogues. Post this, each mass spectral representation is smoothed and baseline corrected followed by peak detection. The peaks detected in each spectrum are then collated into a final peak list, which is subsequently used to extract each peaks intensity from every pixel and the reduced image cube dataset is generated.

A clustering and feature extraction approach for spatially distributed high-dimensional data was proposed by Winderbaum *et al.* [79]. This method uses preprocessed binary data and initially performs noise removal and dimension reduction. This is followed by applying k-means clustering which yields spatially localized clusters that help in distinguishing tissue types. The features from the binary data are extracted using their *difference in proportions of occurrence* approach that identifies discriminating m/z bins and then ranks these bins to select the best variables in a data-driven manner. This approach was applied to detect significant masses and separate tissue types in ovarian cancer.

An approach reported by Alexandrov *et al.* [190] allows to find structured molecular images that helps in listing unknown molecules with significant patterns, present in a MSI dataset. This method uses a measure termed as *spatial chaos*, which is calculated for every ion intensity map and later all the maps are ranked based on this score. They define *spatial chaos*, as “the lack of a spatial pattern in the pixel intensities”. Using this measure, ion intensity maps that have many pixels compactly located and display clear edges exhibit a lower measure of spatial chaos as compared to those that have pixel intensities randomly located and are not associated to form a defined structure.

6.3 Our pattern extraction and image ranking approach

In this section, we present our simplistic approach to extract significant patterns from ion intensity maps and rank these maps based on the pattern abundance. This method is based on applying morphological transformations to ion intensity maps to extract the amount of structural information present in the image. This approach can mainly be summarized in four steps:

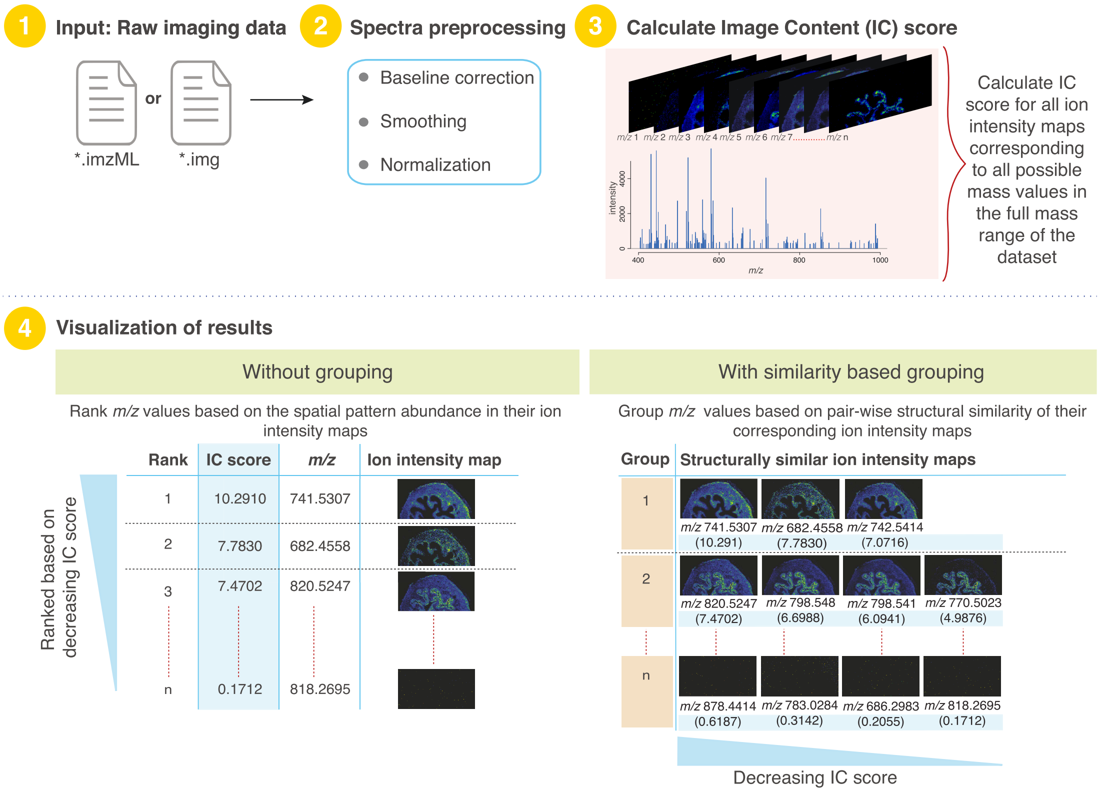

1 Input: Raw imaging data
or
*.imzML
*.img
2 Spectra preprocessing
Baseline correction
Smoothing
Normalization
3 Calculate Image Content (IC) score
m/z 1 m/z 2 m/z 3 m/z 4 m/z 5 m/z 6 m/z 7 m/z n
intensity
0
2000
4000
400
600
800
1000
m/z
Calculate IC score for all ion intensity maps corresponding to all possible mass values in the full mass range of the dataset
4 Visualization of results
Without grouping
Rank m/z values based on the spatial pattern abundance in their ion intensity maps
Ranked based on decreasing IC score
Rank
IC score
m/z
Ion intensity map
1
10.2910
741.5307
2
7.7830
682.4558
3
7.4702
820.5247
n
0.1712
818.2695
With similarity based grouping
Group m/z values based on pair-wise structural similarity of their corresponding ion intensity maps
Group
Structurally similar ion intensity maps
1
m/z 741.5307 (10.291)
m/z 682.4558 (7.7830)
m/z 742.5414 (7.0716)
2
m/z 820.5247 (7.4702)
m/z 798.548 (6.6988)
m/z 798.541 (6.0941)
m/z 770.5023 (4.9876)
n
m/z 878.4414 (0.6187)
m/z 783.0284 (0.3142)
m/z 686.2983 (0.2055)
m/z 818.2695 (0.1712)
Decreasing IC score

Figure 6.1: Schematic illustration of our pattern extraction and image ranking approach This 4-step approach takes a raw imaging data file and performs preprocessing on it. In the next step, an Image Content (IC) score is generated for each ion intensity map corresponding to every possible m/z in the full mass range of the dataset. Results can be visualized as a list of mass values with their respective ion intensity map, ranked based on their spatial pattern abundance, with or without spatial similarity-based grouping.

1. Read the raw imaging data file.
2. Preprocess the raw data to minimize the effect of artifacts.
3. Calculate an *Image Content* (IC) score for each ion intensity map corresponding to every possible m/z value in the full mass range of the dataset.
4. Visualize the results as a list of masses, ranked according to their IC score with or without performing image similarity-based grouping. The ion intensity map with the highest IC score is provided rank 1.

An overview of our method is presented in Figure 6.1.

Before applying our image ranking method to the MSI dataset, we first subject the raw data to the following preprocessing steps - Baseline correction, smoothing and normalization (refer to Section 3.3 for a detailed explanation of these steps). These help in cleaning the data of baseline as well as noise related artifacts and make the spectral intensities comparable. Data preprocessing is performed using the functions available in the `MALDIQuant` R package [90].

6.3.1 Image Content score for an ion intensity map

Our method calculates a score for an ion intensity map, based on the spatial pattern abundance visible within the image. This measure of pattern abundance has been termed as the *Image Content* (IC) score. We estimate the overall pattern present in the ion intensity map using a simplistic approach. The steps to calculate the IC score are illustrated in Figure 6.2.

Given a m/z -specific ion intensity map as a true color image in RGB color space (Figure 6.2**(a)**), we first convert it to a grayscale image, as shown in Figure 6.2**(b)**. This grayscale image represents a color for each pixel with a grayscale value between 0 and 255. Next, this grayscale image is segmented to generate a binary image, shown in Figure 6.2**(d)**. In this image, all the pixels are split into two classes: One class representing the foreground, constituting the areas of interest in the image and the other represents the background. In a binary image, pixels with color values larger than a defined threshold t are set to 1 (displayed in white) that represent the foreground patterns, and other pixels are set to 0 (displayed in black) that represent the background. Selection of an optimal threshold in an automated manner for a grayscale image is explained in Section 6.3.2. Afterwards, a flood-fill step is applied to fill the internal holes in the binary image, resulting in Figure 6.2**(e)**. Filling holes in the binary image helps to create one solid region which is presumed to be a rough estimate of the actual spatial pattern present in the image. The identified regions in the binary image are displayed with visible boundaries in Figure 6.2**(f)**. In the end, we calculate the number of connected components[1] on the binary image and determine the total area fraction of the identified regions. This value is the IC score for a single ion intensity map. We display the connected components identified in the binary image using a pseudo-color image, as shown in Figure 6.2**(g)**.

All morphological operations on the ion intensity maps were performed using various routines available in the Image Processing Toolbox within MATLAB (MATLAB and Statistics Toolbox Release R2012b, The MathWorks, Inc., Natick, Massachusetts, United States).

6.3.2 Selection of an image threshold

In order to extract visible spatial patterns from an image, it is important to segment the image into regions using an optimal threshold t, that is applied to every pixel in the image. Setting the threshold too high can cause loss of pattern associated pixels and setting the threshold too low can increase the number of spurious and undesirable pixels that may appear in the foreground. For a gray scale image $g(i,j)$, when a threshold t is given, the binary output $b(i,j)$ can be calculated as follows [192]:

$$b(i,j) = \begin{cases} 1 & \text{if } g(i,j) \geq t \\ 0 & \text{otherwise} \end{cases} \tag{6.1}$$

A threshold can be selected by inspecting the histogram of an image as shown in Figure 6.2**(c)**. Histogram for an image is a graph displaying the distribution of color variance in an image where the x-axis represents different intensity values and the y-axis represents the number of pixels that have such values. In the ideal case of a symmetric bimodal histogram, with deep valleys between two peaks, the optimum threshold is usually detected near the bottom of the valley [193]. In such a scenario, often termed as *Global thresholding*, the threshold value is held constant throughout the image and the binary image is generated using Equation 6.1. However, this is not the case with

---

[1] *Connected components* in a binary image are individual components or objects that are formed by pixel connectivity. The pixels in a connected component share similar intensity values and are in some way connected with each other.

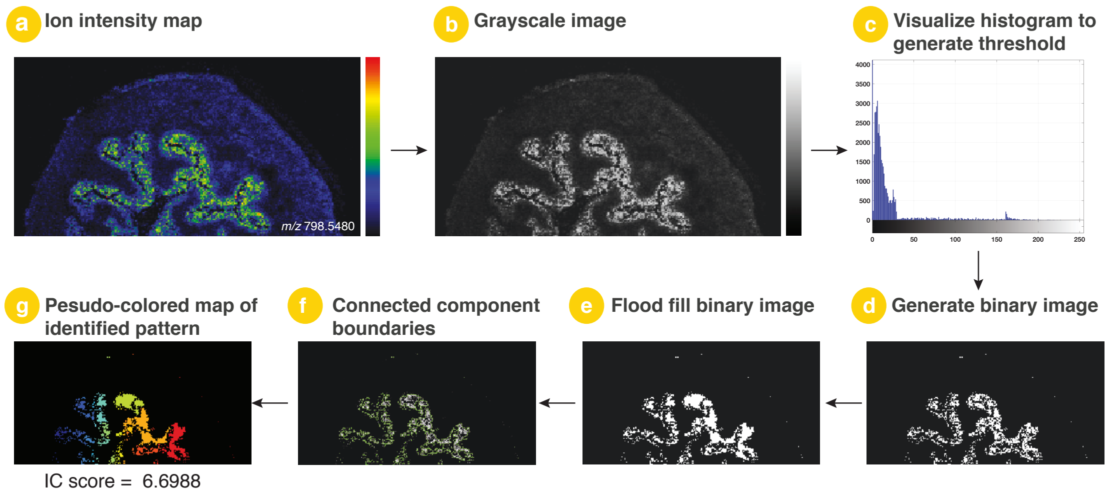


Figure 6.2: Steps to calculate the Image Content (IC) score for an ion intensity map. **(a)** Given the intensity values corresponding to the distribution of an ion on the imaged sample, a pseudo-colored ion intensity map is generated and exported as an image file. **(b)** This is then converted to a gray scale image. **(c)** Next, an image threshold is calculated in an automated manner using the Isodata algorithm [191]. **(d)** This generates a binary image. **(e)** To identify the connected components in this binary image, patterns obtained after thresholding are first flood-filled. **(f)** These connected components are then labeled with boundaries. **(g)** Finally, area for the over-all identified spatial pattern (as can be seen in the pseudo-colored map) is computed. This is the IC score for a specific ion intensity map.

majority of the images that have regions whose statistics differ considerably, making selection of a threshold difficult. There have been many approaches reported for automatic threshold selection that exploit many factors such as histogram shape, entropy, image attributes, spatial information, etc to generate an optimal threshold [194, 195].

We have applied the Isodata algorithm developed by Ridler and Calvard [191], that performs automated threshold selection in an iterative manner. This algorithm was chosen due to its unsupervised characteristic and its capability of selecting an optimal threshold value in histograms that have extremely uneven peak distribution. In this approach, the histogram is initially segmented into two parts using a starting estimate of the threshold value, like the mean intensity of the whole image. This splits the set of pixels into two groups: background and foreground, both of which should be non-empty. Next, the means of both these groups is calculated and the threshold is repositioned to their average. This is carried on iteratively, until the threshold does not change any longer. This procedure is summarized in Algorithm 1.

6.3.3 Spatial similarity-based grouping of ion intensity maps

From the hundreds to thousands of compounds that are measured in a single MSI experiment, many ion species exhibit similar spatial distribution profiles. Grouping ion intensity maps with similar distribution is useful to obtain knowledge about the dataset in an efficient manner. In order to identify ion species with similar distribution patterns and group them, we use an image-similarity metric to calculate pairwise similarity between all the ion intensity maps.

To measure the degree of similarity between two images, we use the structural similarity (SSIM) index proposed by Wang *et al.* [196]. The SSIM index ($S_{(a,b)}$) compares the structural information between two images a and b of the same size based on three

Algorithm 1 Automated threshold selection using Isodata algorithm [191]

```
procedure SELECTTHRESHOLD($h$)    ▷ $h$ is the histogram for grayscale image $g(i, j)$
    $t \leftarrow Mean(h)$    ▷ set initial threshold to overall mean intensity
    repeat
        $P_b = \{g(x,y) \colon g(x,y) < t\}$    ▷ Set of background pixels
        $P_f = \{g(x,y) \colon g(x,y) > t\}$    ▷ Set of foreground pixels
        if $(P_b = 0) \vee (P_f = 0)$ then    ▷ background or foreground is empty
            return -1
        end if
        $\mu_b \leftarrow Mean(P_b)$    ▷ background mean
        $\mu_f \leftarrow Mean(P_f)$    ▷ foreground mean
        $t' \leftarrow t$    ▷ keep previous threshold
        $t \leftarrow \left(\frac{\mu_b + \mu_f}{2}\right)$
    until $t = t'$    ▷ terminate loop if equal
    return $t$
end procedure
```

characteristics: local luminance[2] ($l_{(a,b)}$), global contrast[3] ($c_{(a,b)}$) and structural features ($s_{(a,b)}$). The overall index is a multiplicative combination of the three characteristics as shown in Equation 6.2:

$$S_{(a,b)} = S_{(b,a)} = f[l_{(a,b)} \cdot c_{(a,b)} \cdot s_{(a,b)}] \tag{6.2}$$

The SSIM index always lies between 0 and 1, where an upper bound serves as an indication of how close the images are to being perfectly identical.

Given the pairwise similarities between the ion species stored in the form of a symmetric similarity matrix, we use hierarchical agglomerative clustering with complete linkage function (explained in Chapter 3 Section 3.4.2) to classify masses showing similar intensity distribution patterns across the imaged sample.

6.4 Evaluation of our pattern extraction and image-ranking approach

In this section, we evaluate the pattern extraction and image ranking approach presented in Section 6.3.

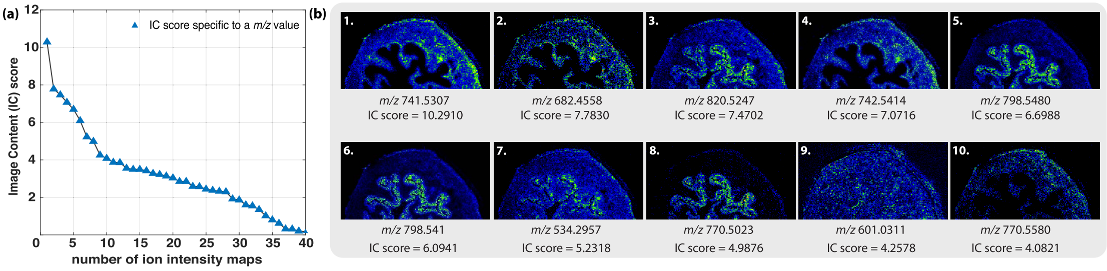
6.4.1 Experimental datasets

We have performed the evaluation using two datasets. For dataset 1, we use publicly available high-resolution MSI data (dataset identifier: PXD001283) deposited to the ProteomeXchange Consortium (`http://proteomecentral.proteomexchange.org`) via the PRIDE partner repository (`http://www.ebi.ac.uk/pride`) [199]. This dataset is acquired from the MALDI-MSI of mouse (*Mus musculus*) urinary bladder with high mass accuracy at a pixel size of 10 μm. Experimental details pertaining to the dataset

---

[2] *Luminance* corresponds to brightness and is a basic property of human vision. It can be defined as a photometric measure that specifies the amount of light emanating from an object or image [197].

[3] *Contrast* is defined as the difference in brightness between the light and dark areas of an image. The global contrast for an image is calculated as the mean of all local contrast values of smaller image fractions [198].



m/z 741.5307 IC score = 10.2910

m/z 682.4558 IC score = 7.7830

m/z 820.5247 IC score = 7.4702

m/z 742.5414 IC score = 7.0716

m/z 798.5480 IC score = 6.6988

m/z 798.541 IC score = 6.0941

m/z 534.2957 IC score = 5.2318

m/z 770.5023 IC score = 4.9876

m/z 601.0311 IC score = 4.2578

m/z 770.5580 IC score = 4.0821

Figure 6.3: Ranking of the 40 selected ion intensity maps from dataset 1, based on their spatial pattern abundance. (a) IC scores for all the 40 selected ion intensity maps sorted in descending order. **(b)** Representation of the top 10 mass values along with their corresponding ion intensity maps, ranked based on their decreasing spatial pattern abundance.

can be found in the original work by Römpp *et al.* [174]. The dataset comprises of 34840 spectra acquired within the slice area of 260×134 pixels. An optical image of the mouse urinary bladder section, available along with the MSI dataset, can be found as Figure B.1 in the Appendix B. The optical image of the tissue section can be broadly categorized in three regions: outer muscular layer, connective tissue and epithelium. From this dataset, we created a subset of 40 ion intensity maps showing the distribution of specific m/z values on the tissue section. Some of these ion maps contain visible patterns (groups of pixels forming a defined structure), whereas some have no visible structure (pixels randomly located). These ion maps were selected on the basis of visual examination and previous knowledge about the dataset [174].

For the second round of evaluation and demonstration of results, we use another MALDI-MSI dataset of a mouse urinary bladder section (dataset 2). This dataset is available for download, as an example file in `*.imzML` format on the `https://ms-imaging.org` website. The dataset comprises of 16200 spectra acquired within the slice area of 180×90 pixels, with 300 mass features measured in the range of m/z 225-250.

6.4.2 Method performance

The first aim of this evaluation was to arrange the ion intensity maps, in a decreasing order of their visible image pattern, based on their IC scores. The second aim was to group these ion intensity maps based on their structural similarity.

Dataset 1

For the chosen set of ion maps, we first calculated the IC scores. Figure 6.3**(a)** displays the IC scores for all the 40 masses, arranged in a descending order. For illustration, the ion maps for the mass values with the first 10 highest IC scores are shown in Figure 6.3**(b)**. As can be seen, our IC score calculation approach is able to precisely extract the visual pattern present in the maps and provide them a score based on their area fraction. In the set of the selected images, m/z 741.5307 tops the list with the highest IC score=10.2910. This compound appears to be mainly distributed in the outer lining of the muscle and parts of the inner lining of the connective tissue. A similar distribution is shown by m/z 682.4558 with IC score=7.7830, ranked on the second position. Next, patterns representing localization of ions in the epithelial region of the tissue section are highlighted, these are ranked between 5 and 8 in the list. The ion intensity maps for m/z 601.0311 (IC score=4.2578, rank=9) and m/z 770.5580 (IC score=4.0821, rank=10), do not display a defined pattern. However these have numerous high intensity

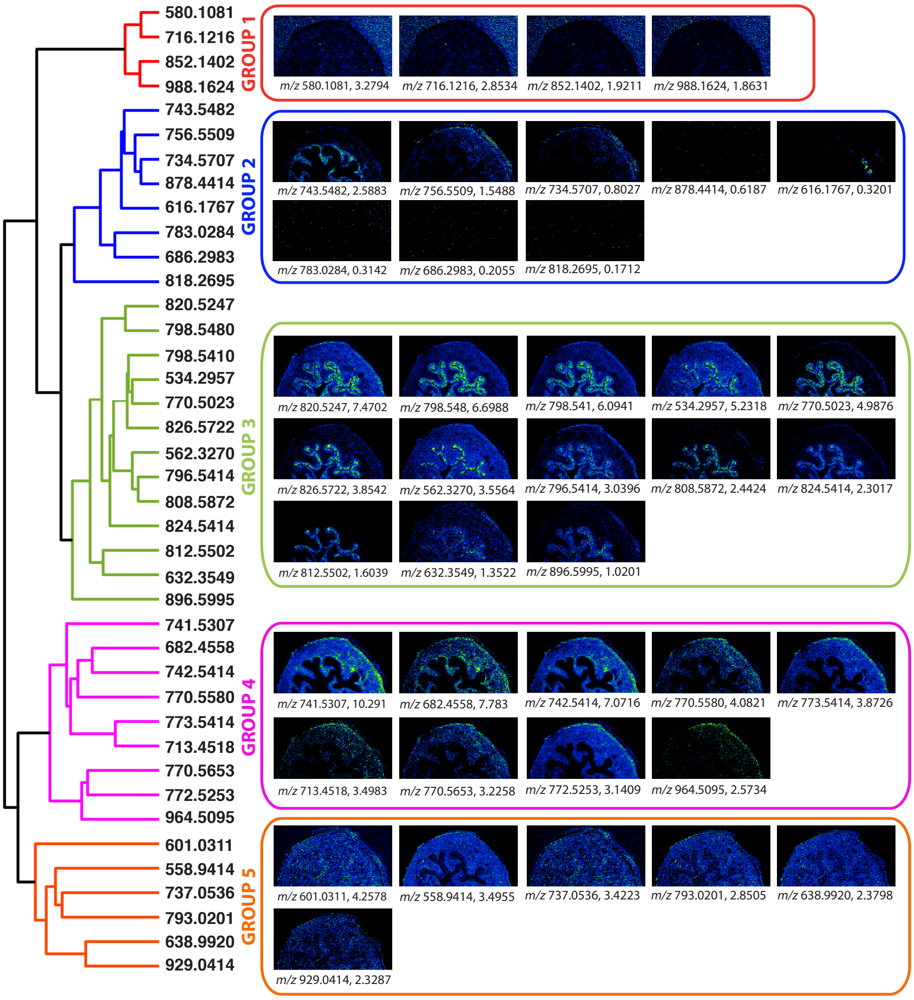


Figure 6.4: Dendrogram representing the result of hierarchical clustering performed for dataset 1. Each nested group formed within the dendrogram contains ion intensity maps that exhibit similar spatial distribution on the imaged tissue section.

pixels distributed all over the tissue section specifically for m/z 601.0311 and mainly in the muscular layer in the latter image.

The ion intensity maps for ion species ranking between 11 to 40 are displayed in the Appendix B Figure B.2, in decreasing order of their IC scores. Ranks 11 to 20 list ion intensity maps with ion distributions in the outer lining of the muscle (m/z 773.5414 and m/z 772.5253), all over the muscular layer and connective tissue (m/z 713.4518, m/z 770.5653, m/z 558.9414), the epithelium (m/z 826.5722, m/z 562.327 and m/z 796.5414). The ion intensity map for m/z 580.1081 (IC score=3.2794, rank=17), displays matrix related signals mainly present outside the tissue section. The IC scores for ion intensity maps ranked between 21 to 40, decrease further due to the presence of limited spatial patterns that can be extracted using our approach. The ion map for m/z 743.5482 (IC score=2.5883, rank=23) clearly highlights the connective tissue present in the imaged bladder section. Further down the list, ion map for the highly localized distribution of m/z 616.1767 (IC score=0.3201) is ranked 37. However, one can see that, even though it could be a biologically relevant spatial patten, it is placed after the ion intensity map for m/z 878.4414 (IC score=0.6187, rank=36), that displays no defined pattern. Providing a lower rank to ion intensity maps containing highly localized distribution patterns, that

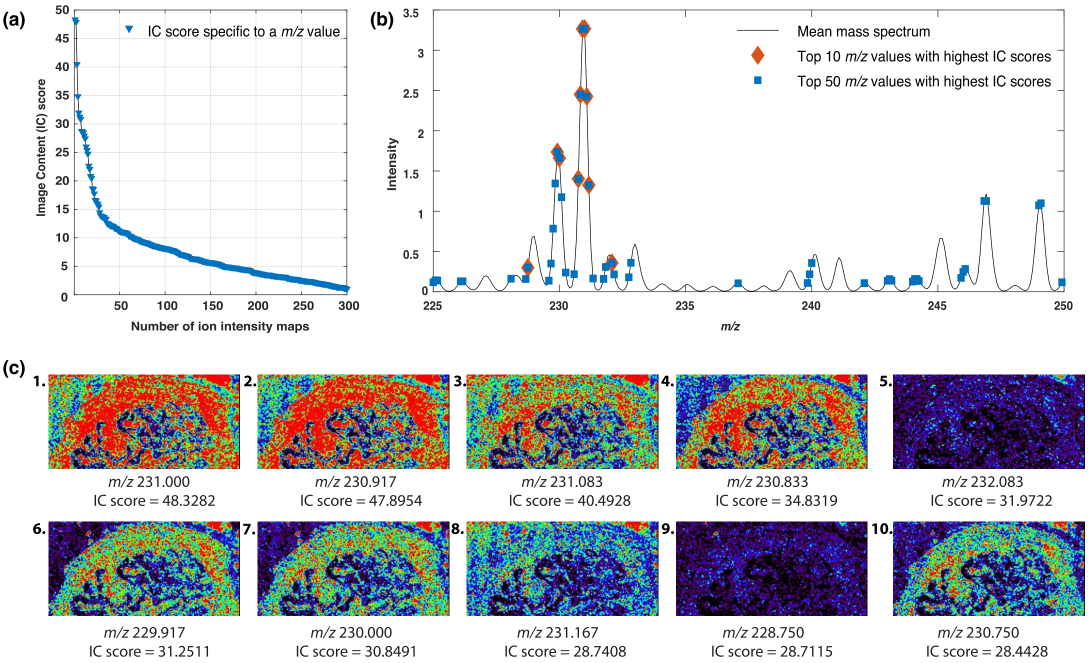


Figure 6.5: Ranking of the 40 selected ion intensity maps based on their spatial pattern abundance. (a) Sorted values of IC scores for all the 40 selected ion maps. **(b)** Representation of the first 10 mass values along with their corresponding ion maps, ranked based on their decreasing spatial pattern abundance.

comprise only a few pixels of the overall image, could be a limitation of the presented approach. Lastly, in the ranked list, ion maps with no visible pattern and randomly located high intensity pixels are provided the lowest ranks.

In order to group these 40 ion intensity maps based on their spatial similarity, we calculated pairwise similarities amongst all the maps and generated a similarity matrix. An illustration of this matrix is provided in the form of a heat map in Figure B.3, in Appendix B. A total of 5 groups containing ion species having similar spatial patterns were identified after hierarchical clustering. A dendrogram representing the nested groups of mass values is presented as Figure 6.4. The list of masses in each group along with their IC scores is provided in Table B.1 in Appendix B. As can be seen in the dendrogram, group 1 that contains four ion intensity maps representing signals from the matrix ions, cluster very well together. Group 3 mainly represents spatial distribution of ions in the epithelium of the mouse bladder. Group 4 represents ion species showing distribution in the muscular layer and the outer lining of the tissue section. And, Group 5 mainly contains ion species that display a distribution all over the tissue section, except for m/z 558.9414 (IC Score=3.4955), that has no signals present in the epithelium region. However, one can see that group 2 consists of ion maps showing diverse spatial distribution patterns. This group mainly consists of maps with lowest IC scores, showing pixels randomly distributed on the tissue section. This group also contains two maps with ion distribution on the outer lining of the tissue section (m/z 756.5509 and m/z 734.5707) and a map with distribution on the outer lining of the epithelium (m/z 743.5482). We assume that this issue could be resolved by enhancing the SSIM index by considering parameters like object edges and distortions as well as by improving our approach to generate the hierarchy of image clusters.

Dataset 2

We then applied our approach to the mouse urinary bladder MSI dataset constituting 300 mass values. Figure 6.5(a) shows the IC scores for all the 300 ion maps, sorted in a descending order. As can be seen, amongst the 300 ion images, about 50 images have an IC score of more than 10. The mass values with the top 10 IC scores and top 50 IC scores are labeled in the mean mass spectrum shown in Figure 6.5(b). As can be seen, most major peaks are present in the top 10 and top 50 ranked list of mass values. For illustration, the ion maps for the mass values with the first 10 highest IC scores are shown in Figure 6.5(**c**). One can see that the ion maps for *m/z* 232.083 (rank 5) and *m/z* 228.750 (rank 9), contain a randomly dispersed distribution on the whole tissue section, with no visible pattern. Also, some groups of high intensity pixels can be seen on the upper region of both the ion maps. These two mass values however have a high IC score. We assume that for such images, with randomly dispersed high intensity pixels all over the image, selecting the mean intensity as an initial estimate for automatic image thresholding could be a drawback. This needs further improvement by using additional robust image-feature related parameters. The IC scores for all the 300 mass values are listed rank-wise in Appendix B Table B.2.

In order to group the ion maps based on their spatial similarity, we calculated the pairwise similarities and generated a similarity matrix. These groups were calculated for top 10, top 50 and top 100. Table 6.1 shows three groups generated using the mass values with top 10 IC scores. A corresponding dendrogram can be found in Appendix B Figure B.4. As can be seen, all the three groups contain mass values that display high intra-group similarity in spatial distribution. Our spatial similarity-based grouping approach is able to differentiate between images containing visual structural pattern from those that have randomly distributed pixels all over the tissue section (*m/z* 228.750 and *m/z* 232.083).

For the mass values with top 50 IC scores, on applying spatial similarity-based grouping, a total of 6 groups were generated. These can be found in Table 6.2 and the corresponding dendrogram can be found in Appendix B Figure B.5. As can be seen in the dendrogram, our method is able to group ions showing similar spatial distribution with high signals arising from matrix and epithelial lining (Group 1), even signal distribution in the matrix region (Group 2 and Group 4) as well as signals arising from the muscular region and connective tissue (Group 5 and Group 6). However, one can see that within Group 4, ions with different spatial distribution are placed together. This group contains mass values showing distribution in the background of the imaged section (*m/z* 246.083, *m/z* 244.167, *m/z* 240.000, *m/z* 226.167 and *m/z* 244.083), all over the tissue section (*m/z* 228.750, *m/z* 230.583, *m/z* 230.250, *m/z* 231.750, *m/z* 229.667 and *m/z* 231.833) as well as distribution in the muscular and epithelium region (*m/z* 231.167). Also, it can be seen that *m/z* 232.833 that has been placed in Group 5 has a different spatial distribution as compared to other ion species in this group. As mentioned previously, this issue can possibly be resolved by using additional information related to the image structure.

6.4.3 Evaluation of running times

We have implemented our IC score calculation and similarity-based grouping approach within the MATLAB (MATLAB and Statistics Toolbox Release R2016a, The MathWorks, Inc., Natick, Massachusetts, United States) framework. In Table 6.3, we report the running times of our approach. All the evaluations have been performed on a Macintosh operating system (version 10.10.5) equipped with 8GB RAM, using 3.1 GHz Intel

Table 6.1: Spatial similarity-based grouping of the ion maps corresponding to the top 10 ranked mass values from dataset 2. The top 10 mass values have been selected based on their IC scores.

m/z	IC score	m/z	IC score	m/z	IC score
Group 1		Group 2		Group 3	
228.750	28.7115	231.083	40.4928	230.000	30.8491
232.083	31.9722	230.917	47.8954	231.000	48.3282
				231.833	34.8319
				231.167	28.7408
				230.750	28.4428
				229.917	31.2511

Table 6.2: Spatial similarity-based grouping of the ion maps corresponding to the top 50 ranked mass values from dataset 2. The top 50 mass values have been selected based on their IC scores.

m/z	IC score	m/z	IC score	m/z	IC score	m/z	IC score	m/z	IC score	m/z	IC score
Group 1		Group 2		Group 3		Group 4		Group 5		Group 6	
246.833	12.1040	244.250	11.3859	226.083	13.6300	228.750,	28.7115	232.833	27.9242	231.000	48.3282
249.000	11.6121	249.917	11.8050	225.083	13.8926	230.583	20.8312	229.750	16.0282	230.917	47.8954
246.917	12.0171	228.083	11.6692	237.083	11.1480	231.167	28.7408	230.083	24.7494	231.083	40.4928
249.083	11.2824	243.083	13.3245	229.583	13.2473	246.083	15.4536	230.000	30.8491	230.833	34.8319
		244.000	13.5568	228.667	13.6997	244.167	14.4451	229.833	27.4424		
				245.917	16.5166	230.250	22.6309	230.750	28.4428		
				231.333	12.5903	231.750	16.6236	229.917	31.2511		
				243.000	12.4736	240.000	22.0618				
				225.167	12.4150	229.667	26.0017				
				246.000	17.7613	226.167	13.7967				
				232.167	20.4686	244.083	15.8267				
				239.917	18.6668	231.833	25.4084				
				242.083	11.6672						
				239.833	12.1126						
				225.000	12.1530						
				243.167	12.2519						
				232.083	31.9722						
				232.750	18.5229						

Table 6.3: Running times measured for IC score calculation and spatial similarity-based grouping for the two MSI datasets used in this evaluation.

Category	Running times
IC score calculation	
Dataset 1 (40 ion intensity maps)	3 s
Dataset 2 (300 ion intensity maps)	36 s
Structural similarity-based grouping	
Dataset 1 (40 ion intensity maps)	19 min 6 s
Top 10 ion intensity maps in dataset 2	2 min 30 s
Top 50 ion intensity maps in dataset 2	1 h 4 min
Top 100 ion intensity maps in dataset 2	4 h 44 min

Core i7 processor. The values reported here only apply for the IC score calculation and spatial similarity-based grouping for individual datasets, without including the time required for initial preprocessing of mass spectra.

As can be seen in Table 6.3, our IC score calculation and ranking approach is extremely fast (ca. 75 ms for a single ion map of dimensions 260×134 pixels and ca. 120 ms for a single ion map of dimensions 180×90 pixels), even when the number of ion intensity maps increases. However, this is not the case while performing spatial similarity-based grouping. As the number of ion maps used as an input to perform spatial similarity-based grouping increases, the running time increases almost exponentially.

Chapter 7

Preprocessing and multivariate analysis of TOF-SIMS imaging data

To derive reliable conclusions from MSI data, it is of utmost importance to apply appropriate preprocessing steps and perform subsequent multivariate analysis (MVA). Over the years, several methods for both preprocessing and MVA have been developed. In this chapter, we present a case study to probe the complex leaf surface chemistry of *Populus trichocarpa* (*P. trichocarpa*) by applying these techniques to the acquired TOF-SIMS imaging data. The applied approach revealed a set of unique crystal formation patterns in epicuticular waxes (EWs), present on the surface.

In Section 7.1, we first introduce the objective of the study and its ecological importance. In Section 7.2, we briefly provide the experimental details regarding the acquired TOF-SIMS imaging data. Later in Section 7.3, we explain the data analysis pipeline that includes data preprocessing and different MVA approaches. Within this section, we also highlight the importance of selecting specific data preprocessing methods based on the nature of the data, since this is highly critical for the downstream analysis. In Section 7.4, we present the results of our analysis.

7.1 Introduction

Studying the physicochemical properties of plant surfaces has been the subject of a number of studies in the past [200–202]. Leaf surfaces are composed of cuticular waxes where-in spatial differentiation exists [203][1]. These possess considerable ultra-structural and chemical diversity.

Black cottonwood, *P. trichocarpa* (Torr. & A. Gray), is an economically and ecologically relevant tree and the first woody plant whose genome has been sequenced [204]. *P. trichocarpa* nowadays is considered a model for long-living trees [205–207]. However, knowledge is still lacking regarding the extent of the leaf surface represented by EWs, the chemical characterization of EWs and their role in interactions with herbivorous insects.In this study, we apply TOF-SIMS imaging followed by MVA to investigate the chemistry of EWs present on the leaf surface of *Populus trichocarpa.*

From literature, we know that after crystallization of EWs, the leaf surface properties appear to be dramatically modified in comparison to the amorphous EW layer [208, 209]. Crystallization may be initiated as a mono-layer self-assembly, later rising over the original layer by an under-flow of EWs at the center. The crystallization of EWs was also studied for pure compounds as well as isolated and partially purified EW mixtures [210]. Crystals tend to form diverse geometric shapes and have been shown to regulate water flow from the surface (lotus effect) or to act as plant defenses against herbivory [211].

---

[1] The plant surface which is exposed to the environment is coated with a layer known as the cuticle. It is a lipophilic structure formed of highly non-polar compounds, the cuticular waxes, on the outside of epidermal cells. Within cuticular waxes, different layers exist [203]: an inner layer, in which cutin is embedded, is composed of intracuticular waxes; an outer layer, which has direct contact with the environment and thus creates the actual plant surface, is formed by the EWs (See Appendix Figure C.1**(a)**).

7.2 TOF-SIMS imaging

The leaf sample surface was prepared to perform TOF-SIMS imaging. Details related to this can be found in Appendix Section B.1.

TOF-SIMS imaging was performed on a standard commercial TOF.SIMS 5 instrument (ION-TOF GmbH, Münster, Germany, `www.iontof.com`). The spectrometer was equipped with a Bi cluster primary ion source and a reflectron type time-of-flight analyzer. The UHV base pressure was $< 5\times10^{-9}$ mbar. For high mass resolution, the Bi source was operated in the high-current bunched mode providing short Bi^{+} or Bi^{3+} primary ion pulses at 25 keV energy and a lateral resolution of approximately 4 μm. The pulse length of 1.1 to 1.3 ns allowed high mass resolution. The primary ion beam was rastered across 700×700 μm^2, 500×500 μm^2 and 100×100 μm^2 sample areas, and 700×700, 128×128 and 100×100 data points were recorded. Images larger than the maximum deflection range of the primary ion gun were obtained using the manipulator stage scan mode with a lateral resolution of 100 pixel/mm. Primary ion doses were kept below 10^{11} ions/cm^2 (static SIMS limit). Spectra were calibrated on C^{-}, C^{2-} and C^{3-} peaks for negative ion mode and on C^{+}, CH^{+}, CH^{2+} and CH^{3+} peaks for positive ion mode. Based on these datasets, the chemical assignments for characteristic fragments were determined. The experiments were performed with 5 biological and 2 technical replicates for 700×700 μm^2 area, 1 biological replicate for 500×500 μm^2 area and 3 biological replicates for 100×100 μm^2 area. The following standards were used: hexacosanoic acid, ethyl stearate, methyl tricosanoate, 1-hexacosanol, tetracosane (Sigma-Aldrich, Germany, `www.sigmaaldrich.com`). The data acquisition software was IonSpec (ION-TOF GmbH, Münster, Germany, `www.iontof.com`). Negative ion mode data at 1 μm lateral resolution was used for the data analysis.

7.3 Data processing and analysis

A MSI experiment usually generates a large data cube with two spatial dimensions and one *m/z* dimension. Depending on the biological question, if one is interested in observing the spatial distribution of a few specific ions of interest individually, then 2-dimensional (2-D) molecular ion intensity maps can be generated from this data cube.

However, if the goal is to understand the overall chemical composition of the sample or analyze correlations between regions and study multiple analytes, then a different approach is necessary. As a first step, the mean spectrum of the dataset can be generated to identify the major peaks. However, the mean spectrum at times under-represents mass peaks that are generated in only a small portion of the spectra, leading to overlooking the peaks that may be biologically significant [112]. It has also been observed that within a typical TOF-SIMS spectrum, multiple peaks are generated from the same surface molecules, and their relative yields are often interrelated. This makes exploratory analysis important.

MVA techniques such as PCA and MCR, exploratory factor analysis, maximum autocorrelation factors, neural networks, latent profile analysis and mixture models are useful for identifying chemically significant areas on these 2-D ion intensity maps [212–216]. MVA and cluster analysis techniques have been extensively used to distinguish spatial structures and establish correlation patterns for data obtained from the SIMS imaging of several biological samples [217] such as proteins [218], lipids [219, 220], bio-materials [213] and single cells [221].

Based on the nature of the acquired data, we first perform data preprocessing using carefully selected approaches and later interpret the information in the data using MVA

Table 7.1: Selected ions of interest for multivariate and classification analysis of TOF-SIMS imaging data.

Chemical class	Chemical formula	Monoisotopic mass (Da)
Alcohol (Alc)	$C_{21}H_{44}O$	312.3392
Alc	$C_{22}H_{46}O$	326.3548
Alc	$C_{23}H_{48}O$	340.3705
Alkane (Alk)	$C_{25}H_{52}$	352.4069
Alc	$C_{24}H_{50}O$	354.3861
Alc	$C_{25}H_{52}O$	368.4018
Alc	$C_{27}H_{56}O$	396.4331
Alk	$C_{29}H_{60}$	408.4695
Alc	$C_{28}H_{58}O$	410.4487
Alk	$C_{30}H_{62}$	422.4852
Alc	$C_{29}H_{60}O$	424.4644
Alk	$C_{31}H_{64}$	436.5008
Alc	$C_{30}H_{62}O$	438.4800
Alc	$C_{31}H_{64}O$	452.4957
Alc	$C_{33}H_{68}O$	480.5270
Wax ester(WE)	$C_{44}H_{88}O_2$	648.6784
WE	$C_{46}H_{92}O_2$	676.7097
WE	$C_{47}H_{94}O_2$	690.7253
WE	$C_{48}H_{96}O_2$	704.7410

and segmentation approaches. This is explained in the coming subsections. It should be noted that the approach discussed in this work is applied to a subset of the data, containing only a few selected ions of interest (often referred to as targeted analysis [222]). When there is no prior knowledge about the surface chemistry of the analyzed sample, then the full spectrum data has to be used. Even then, this approach should be equally adaptable for exploratory/untargeted analysis.

For preliminary inspection of the TOF-SIMS imaging data, we generated pseudo-colored ion intensity maps using SurfaceLab 6.3 software (ION-TOF GmbH, Münster, Germany, `www.iontof.com`). We selected a total of 19 peaks of interest, comprising of alcohols (Alc), alkanes (Alk) and wax esters (WE), in the range of m/z 200-800, to perform multivariate and classification analysis. These are shown in Table 7.1.

Spectral data for these 19 peaks were extracted and converted from the vendor file format and exported as individual text files using SurfaceLab 6.3 software (ION-TOF GmbH, Münster, Germany, `www.iontof.com`). The data for these 19 peaks were arranged in the form of a $n \times m$ matrix where the n rows are *samples* which denote the spectra acquired at every single coordinate position (x, y) and the m columns are *variables* denoting the mass peaks of interest. For the 100×100 μm^2 TOF-SIMS imaging dataset, this matrix had dimensions of [10,000 × 24]. The complete data analysis was performed on a computer running Mac OS (version 10.10.5) with 8 GB of RAM and a 3.1 GHz Intel Core i7 processor. Data preprocessing and multivariate analysis was carried out using the chemometrics software Solo+MIA (Eigenvector Research, Inc. Wenatchee, WA, USA, `www.eigenvector.com`) and the R statistical package [223].

7.3.1 Data preprocessing

Data preprocessing steps consisted of *mean centering* the variables, as a first step. In *mean centering*, each variable is centered by the subtraction of its mean value across all samples. This is typically done so that the differences among the peak variances are emphasized over the differences in the peak area means. Additionally, we also scaled the data to account for *Poisson noise* [224–227]. The data from TOF-SIMS instruments are collected in a pulse-counted manner, which is subject to an uncertainty explained by Poisson statistics. This uncertainty is equal to the mean of the signal intensity. Multivariate approaches such as PCA are designed to account for variance in the data, and non-normalized variables with large variance have stronger weights and are more likely to be addressed in the modeling than are low variance variables. Since TOF-SIMS data usually has variance which is related to the signal intensity, these approaches perform sub-optimally. *Poisson scaling* (also called square *root mean scaling*) scales each variable by the square root of its mean value so that the estimated variance due to counting statistics is equal on all variables.

One important scaling method often necessary for mass spectrometry imaging data, irrespective of the ionization method used, is normalization. This technique helps in identifying and removing sources of systematic variation amongst pixels in the dataset, which in turn is beneficial to minimize inter-spectra variance. Normalization is usually performed by multiplying every mass spectrum in the dataset with an intensity-scaling factor, in order to make all spectra comparable. There have been many methods proposed for normalization, the most commonly applied being total ion current normalization [23, 102, 159]. We applied normalization to this dataset, however it led to the loss of contrast and distinct biological features on the sampled leaf surface, hence was not included as a part of data preprocessing.

7.3.2 Multivariate analysis and clustering

PCA (explained in Section 3.4.1 of Chapter 3) was applied to the preprocessed TOF-SIMS imaging dataset. Score plots show the values for each coordinate position (pixel) on the associated PC axis. The pseudo-color scale indicates the level of contribution of each pixel to the axis. Pixels that correspond to the same histochemical structure (i.e., pixels showing similar mass spectra) are expected to have a similar contribution to different PCs, and these pixels appear with the same color. Loading plots show the positive and negative correlations of each original variable with the respective principal component. The score plots, loadings and the observed correlation of individual peaks are discussed in Section 7.4.

MCR (explained in Section 3.4.1 of Chapter 3) was also applied to the preprocessed TOF-SIMS imaging dataset. Score plots reveal the distribution of the components of interest on the surface. These score plots are represented using pseudo-colors so as to display the level of contribution of each component. The MCR loadings on each factor resemble an actual SIMS spectrum of the component, showing its characteristic peaks. The score plots, the loadings and the observed correlation of individual peaks are discussed in Section 7.4.

For classification studies based on cluster analysis (explained in Section 3.4.2 of Chapter 3), first we performed HCA of the TOF-SIMS imaging data. HCA was applied to group the selected ions of interest based on their spatial patterns (specifically localization of crystals on the leaf surface). The dendrogram plot and heat map representation of the HCA was generated using the functions `hclust` and `heatmap.2` within

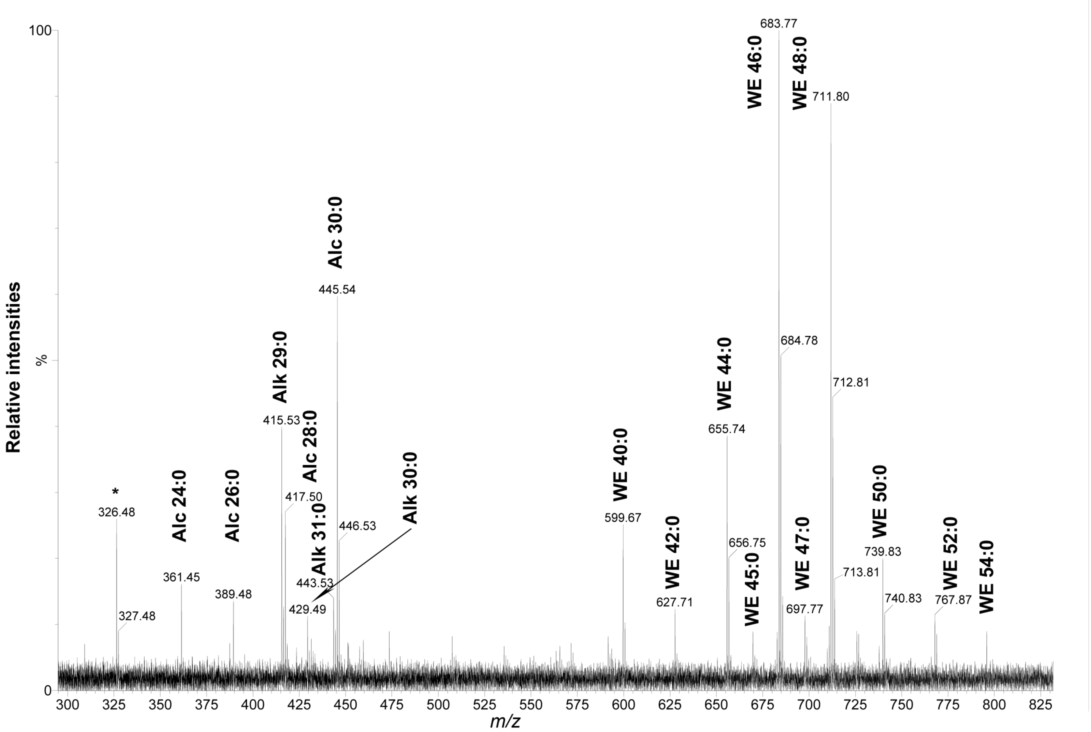


Figure 7.1: MALDI-TOF MS spectrum of EWs isolated from the adaxial leaf surface of *P. trichocarpa* using the cryo-adhesive isolation method; Alk - alkanes, Alc - alcohols, WE - wax esters.

the gplots package in R version 3.2.3 (2015-12-10, R Foundation for Statistical Computing, `www.r-project.org`). DCA was performed using multiple k values, to obtain the pattern which best correlates the PCA and MCR results. DCA results were generated using Solo+MIA software (Eigenvector Research, Inc. Wenatchee, WA, USA, `www.eigenvector.com`). Spatially aware segmentation was performed using the available `spatialKMeans` function and results were generated using the `plot` function implemented in the `CARDINAL` package [228] in R version 3.2.3 (2015-12-10, R Foundation for Statistical Computing, www.r-project.org).

7.4 Results

Prior to TOF-SIMS imaging, the composition of blotted EWs was studied using MALDI-TOF MS. Positive ion MALDI-TOF MS spectra $[M+Li]^+$ from the adaxial surface of *P. trichocarpa* leaves showed three characteristic series (Figure 7.1). Ions within each series were separated by 14 Da (CH_2), indicating the presence of consecutive homologues within the group. Their masses were compared with standards and previously published data [229, 230]. The first series observed was assigned to C_{29}-C_{31} hydrocarbons, the next series with an increment of 28 or 14 Da was assigned to C_{24}-C_{33} alcohols and the third series was assigned to long-chain C_{40}-C_{54} saturated wax esters. The observed intensities in the MALDI-TOF spectrum might be influenced by the fact that WEs will bind Li^+ ions more strongly than hydrocarbons [229] and therefore the intensities of WEs might be biased. Within this series, additional signals with a decrement of 2 Da were observed and assigned to unsaturated esters (see Appendix C Table C.1).

Later, an extensive TOF-SIMS imaging analysis using two lateral resolutions of 10 and 1 μm in both positive and negative ion modes was performed to validate the MALDI

Table 7.2: Percentage of variance captured by PCA performed on the preprocessed data obtained from TOF-SIMS of *P. trichocarpa* leaf surface.

PC	Eigenvalue	Variance captured by PC (%)	Cumulative variance (%)
1	1.83e+01	42.02	42.02
2	7.93e+00	18.23	60.25
3	2.36e+00	5.43	65.68
4	2.11e+00	4.85	70.53
5	1.80e+00	4.14	74.67
6	1.28e+00	2.94	77.61

data. Molecular ions $M^{+\cdot}$ and $M^{-\cdot}$ for the 19 selected compounds of interest, including 11 alcohols, 4 hydrocarbons and 4 wax esters, are shown in Table 7.1. Here some signal intensities were biased and the low abundance of WEs observed could be explained by their low stability under SIMS experimental conditions. On the other hand, Alk and Alc series were well represented but higher than intensities of the corresponding ions in MALDI-TOF MS spectra (Figure 7.1 and Table C.1 in Appendix C).

Principal component analysis

For this study, we selected six principal components (PCs) based on the total variance captured (see Figure C.2 in Appendix C). PC 1 captures 42% of the total variation, while PCs 2 to 6 capture a total of 36%, making a total of 78% variance captured by all the selected PCs. Details related to the percentage of contribution by selected PCs are provided in Table 7.2. The variance captured by each factor decreases quickly for factors that contain chemical features, and then reaches a gently declining slope for factors that describe noise variations. As PCs higher than 6 do not provide any additional information and do not seem to contain any clear systematic structure, it is appropriate to consider that these components contain disproportionately more noise as compared to true signal. The typical computation time is less than 5 seconds due to the small size of the data matrix. The score and loading plots for all the selected PCs are shown in Figure 7.2.

The score plot for PC 1 (Figure 7.2(a)) captures the overall variation in intensity arising from the topography of the leaf surface. Such topography includes the lateral leaf vein, background and small islets (here, crystals) and represents the largest source of variance in the data. The loading plot, which indicates the contribution of each ion to PC 1 (Figure 7.2(b)), shows high positive loading for pentacosanol C25-Alc, m/z 368 and C23-Alc, m/z 340 with negative loadings for C21-Alc, m/z 312 and C22-Alc, m/z 326.

The score plot for PC 2 (Figure 7.2(c)) shows an enhanced chemical contrast between the leaf surface background and certain localized pixels (crystals). The high-contrast yellow regions in the score plot can mainly be attributed to the increased contribution from C29-Alc; for m/z 424, on the other hand, the pixels in black can be correlated with the high negative loadings exhibited by C23-Alc, m/z 340, as shown in the loading plot (Figure 7.2(d)).

The score plot for PC 3 (Figure 7.2(e)) mainly highlights the other crystal-forming regions (shown in yellow) on the leaf surface. These can be attributed to increased C23-Alc, m/z 340 in loadings for PC 3 (Figure 7.2(f)). The sharp negative loadings for C25-Alc, m/z 368 can be attributed to the black pixels in the upper right corner of each

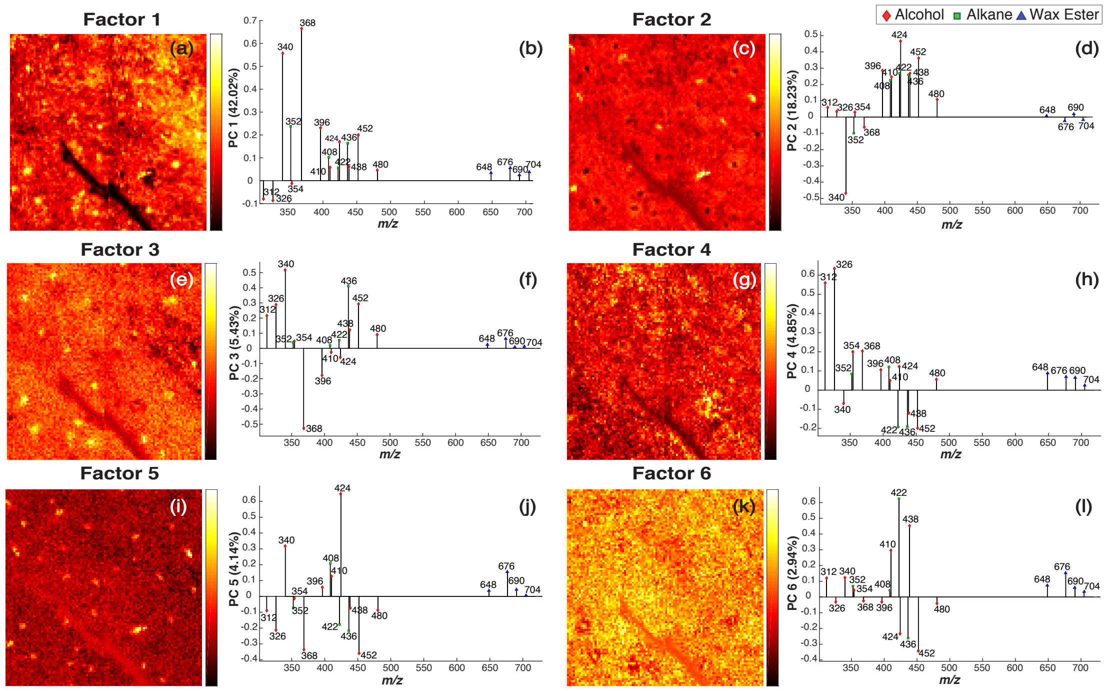


Figure 7.2: PCA results for data obtained using negative mode TOF-SIMS imaging of the surface of *P. trichocarpa* leaves at 1 m step size. **(a-b)** Score (left) and loading (right) plots corresponding to PC1. **(c-d)** Score and loading plots corresponding to PC2. **(e-f)** Score and loading plots corresponding to PC3. **(g-h)** Score and loading plots corresponding to PC4. **(i-j)** Score and loading plots corresponding to PC5. **(k-l)** Score and loading plots corresponding to PC6. PC 1-6 are the six selected principal components. Scores are plotted using a standard hot color gradient scale where black represents high negative loadings and going from red, yellow to white represents high positive loadings.

leafs surface area in its corresponding score plot. The individual crystal patterns from PC 2 and PC 3 show different spatial organization accompanied with distinct chemistry. On calculating the distance in pixels for the crystal patterns observed for contribution of C29-Alc, m/z 424 in PC 2 and C23-Alc, m/z 340 in PC 3, it was found that the distance among the crystals in the two patterns differs distinctly (ca 10 μm vs ca 25 μm, as seen in Figure C.3 in Appendix C).

The score plot for PC 4 (Figure 7.2(g)) does not highlight any crystal patterns but shows increased contributions from C21-Alc, m/z 312 and C22-Alc, m/z 326. This is evident from the corresponding loading plot (Figure 7.2(h)).

The loadings for PC 5 (Figure 7.2(j)) again highlight slightly increased contributions from C23-Alc, m/z 340 and C29-Alc, m/z 424, also reflected in the score plot with evident crystal patterns (Figure 7.2(i)). The score and loadings plot for PC6 do not provide any information related to crystal formation; however, the score plot (Figure 7.2(k)) highlights the background of the leaf surface, and this background can be correlated with sharp increases in the loadings for triacontane C30-Alk, m/z 422, C28-Alc, m/z 410 and C30-Alc; m/z 438.

Multivariate curve resolution

After applying Poisson scaling to equalize the noise variance of each data point, the preprocessed data was subjected to MCR. The chemical contrast in the score plots improved visibly after Poisson scaling was applied to raw data. For this study, we

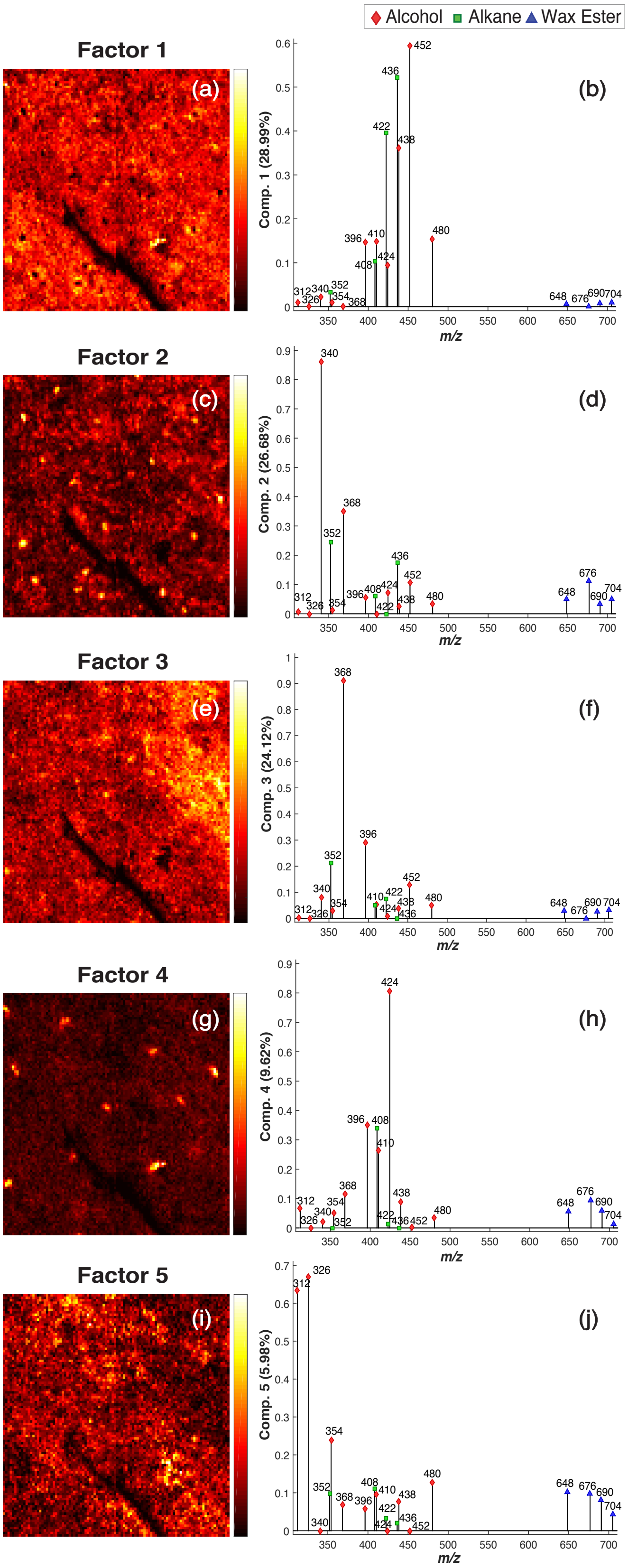


Figure 7.3: MCR results for data obtained using negative mode TOF-SIMS imaging of the surface of *P. trichocarpa* leaves at 1 m step size. (a-b) Score (left) and loading (right) plots corresponding to factor 1. **(c-d)** Score and loading plots corresponding to factor 2. **(e-f)** Score and loading plots corresponding to factor 3. **(g-h)** Score and loading plots corresponding to factor 4. **(i-j)** Score and loading plots corresponding to factor 5.

Table 7.3: Results of MCR analysis performed on the preprocessed data obtained from TOF-SIMS analysis of *P. trichocarpa* leaf surface.

Factor	Fit (%X)	Cumulative Fit (% X)
1	28.99	28.99
2	26.68	55.67
3	24.12	79.78
4	9.62	89.40
5	5.98	4.14

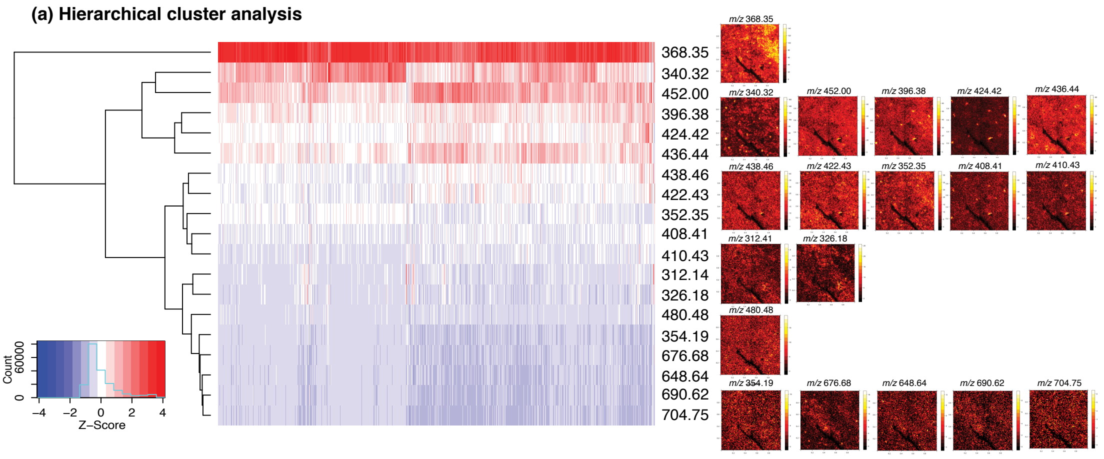
selected five factors for the MCR model. This selection was based on an estimate of the number of chemical species present in the dataset and also on the number of PCs that explained the majority of the variance in the data, as shown previously.

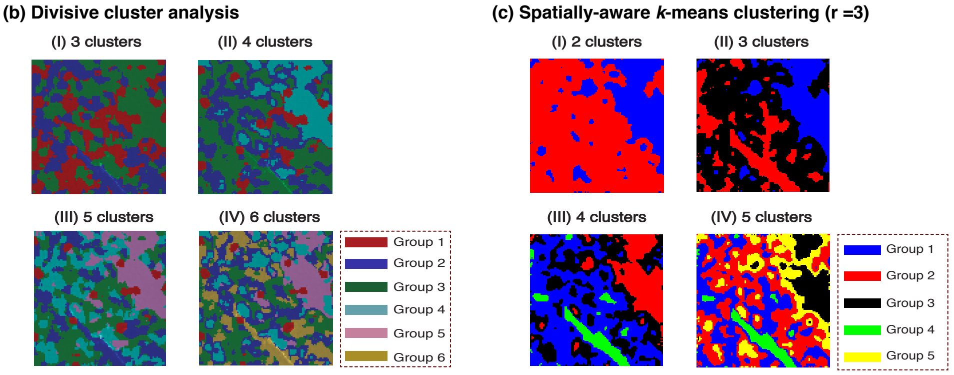
The score and loading plots obtained after applying MCR are shown in Figure 7.3. Details related to the percentage of contribution by each factor in the 5-factor MCR model are provided in Table 7.3. The loading plots of the MCR analysis, unlike those of PCA, were interpreted as normal spectra by applying non-negativity constraints. Because these spectral responses show only positively correlated species, interpretation is made easier.

The score plot for factor 1 (Figure 7.3(a)) does not provide much insight into the crystal formation patterns or the determination of which pure components are responsible for the pattern. The corresponding loading plot (Figure 7.3(b)) shows high contributions from C31-Alk, m/z 436, C30-Alk, m/z 422 and C30-Alc, m/z 438. These contributions may contribute solely to the background chemical composition of each leaf's surface. Score plots for factors 2 and 4 (Figure 7.3(c,g)) mainly show the distinct patterns of crystal formation. Their respective loading plots (Figure 7.3(d,h)) show an increased contribution from C23-Alc, m/z 340, in loadings for factor 2 and C29-Alc, m/z 424 in factor 4. Figure C.4 in Appendix C clearly represents the two crystal patterns observed in a 2-color image overlay of score plots for factor 2 (pixels in red) and factor 4 (pixels in green). The loading plots corresponding to these factors have also been overlaid to identify the distinct chemical contribution. The score plot for factor 3 (Figure 7.3(e)) highlights the curved region at the upper right corner parallel to the lateral leaf vein; this curve can be attributed to the strong presence of C25-Alc (m/z 368) in the loadings plot (Figure 7.3(f)). These results from MCR analysis confirm the previously described results of PCA.

Cluster analysis

Clustering was performed on the same variables as those for PCA and MCR. The results for the cluster analysis are shown in Figure 7.4. First, HCA was applied to the 19 variables, in order to group them. The dendrogram and heat map representing the HCA results, constructed using the Wards linkage method, are shown in Figure 7.4(a). The dendrogram reveals the relationship between the ions of interest, based on the spectral differences/similarities amongst them. Each row in the heat map corresponds to an individual ion of interest, and each column corresponds to a spatial coordinate position (x, y) on the leaf's surface. The color of each cell in the heat map represents the Z-score which measures the distance, in standard deviations, between that cell and the mean of all cells in that column. As seen in the figure, the dendrogram consists of four major clusters. The first cluster is represented by a single ion forming the characteristic curved


(a) Hierarchical cluster analysis


(b) Divisive cluster analysis

(c) Spatially-aware *k*-means clustering (r =3)

Figure 7.4: Cluster analysis on TOF-SIMS imaging data. **(a)** Results for HCA showing three main cluster groups as seen in the dendrogram. Each row in the heat map represents a selected mass peak of interest and each column represents the deviation of the spectral information at each coordinate position (x, y). The heat map is color-coded based on the Z-score, where red represents a high positive deviation from the mean and blue represents a high negative deviation from the mean. White represents no deviation from the mean value. **(b)** Results for DCA using k-means algorithm performed on data obtained using TOF-SIMS imaging of the surface of *P. trichocarpa* leaves. **(I)** Clustering with k=3. **(II)** Clustering with k=4. **(III)** Clustering with k=5. **(IV)** Clustering with k=6. **(c)** Results for spatially-aware k-means clustering with pixel neighborhood radius r=3. **(I)** Clustering with k=2. **(II)** Clustering with k=3. **(III)** Clustering with k=4. **(IV)** Clustering with k=5.

pattern on each leafs surface (C25-Alc, m/z 368). The second cluster mainly consists of ions that belong to the Alc and Alk class, mainly dominated by those that exhibit a distinct crystal formation pattern. However, one can see that the ion intensity maps of m/z 452 and m/z 436 do not display any crystal formation pattern. Similarly, the third cluster also consists of ions that belong to the Alc and Alk class, but is dominated by ions that do not display crystal patterns. However, ion intensity maps of m/z 408 and m/z 401 show crystal formation patterns. One explanation for this inconsistent grouping of ions could be the distance measure used for HCA. It possibly performs grouping based on the intensities at individual pixels independently, and does not consider their spatial context. The last cluster in the dendrogram is formed of non-crystal forming ions, mainly comprising of WEs.

Results for DCA are shown in Figure 7.4(b) as a palette of pseudo-colored images (I-IV) representing the partitioning into 3, 4, 5 and 6 clusters. The three-class clustering (Figure 7.4(b.I)) broadly distinguishes the background structures of the leaf's surface without offering much information about the localization of the crystal. The four-class clustering (Figure 7.4(b.II)) distinguishes some of the crystal-forming regions, shown in red; the curved region on the leaf surface in the upper right corner, shown in turquoise; and the leaf background, shown in blue and green regions. The five-class clustering (Figure 7.4(b.III)) shows the crystal-forming regions more precisely in red, the leaf background mainly in blue, with irregular regions of green and turquoise, and the upper right curved region of the leaf in pink. The six-class clustering (Figure 7.4(b.IV)) shows the location of crystal-forming regions in red, the region of the leaf's vein as well as some regions of each leaf's background in gold, the curved region in the upper right corner in pink, and the background of each leaf's surface in green, blue and turquoise.

Since HCA as well as DCA using *k*-means treat intensities from pixels independently, we also applied spatially-aware k-means clustering. Results for spatially-aware *k*-means clustering using pixel neighborhood radius r=3 are shown in Figure 7.4(c). The clustering results are in the form of a palette of pseudo-colored images (I-IV) representing the spatially-aware partitioning into 2, 3, 4 and 5 clusters. As can be seen, the five-class clustering (Figure 7.4(c.IV)) broadly distinguishes localization of the crystal-forming regions, the region of the leaf vein, the background structures as well as the characteristic curved structure in the upper right corner of the leaf surface.

The results from DCA and spatially-aware clustering analysis also point to the localization of crystals with distinct chemical specificity.

As observed from MVA and clustering results, the chemical composition of the crystals differs clearly from the chemical composition of other areas. The crystals with smaller distances are formed mostly by C23-Alc and those with larger distances by C29-Alc, a finding that is fully congruent with previous findings on the uniform compositions of crystals in EWs [231]. The uniformity is likely related to molecular self-assembly during crystallization. The chemical composition of other EWs is much more diverse. The presence of a mixture of hydrocarbons (C30, C31) and aliphatic alcohols (C30, C31), and the virtual absence of WEs, are characteristic for the score and loading plot for MCR factor 1 in Figure 7.3(a,b). In contrast, the score plot for MCR factor 5 in Figure 7.3(i,j) shows the presence of alcohols (C21-C23) and WEs. This may further help in answering important biological questions and reveal unprecedented insights about the composition of the wax layer, how this layer is repaired after mechanical damage or insect feeding and which transport mechanisms are involved in deploying wax constituents to specific regions on the leaf surface.

Chapter 8

Conclusion

In this thesis, we have presented different computational approaches that can aid in the analysis and interpretation of MSI data in several ways. The three main contributions illustrated are mass recalibration, unsupervised exploration of MSI data and application of preprocessing and chemometric methods to analyze MSI data. In this chapter, we provide a summarized overview of the methods reported in this thesis and discuss the potential future directions for continuation of this work.

In Chapter 4, we presented a computational method to reduce mass shifts in MSI data. High mass accuracy is essential when performing mass spectrometry, to achieve compound identification with high confidence. However, mass accuracy could be severely compromised due to various factors, making recalibration important. A measured mass spectrum can be traditionally recalibrated by using an internal mass standard (typically known as the lock-mass) or by recalibrating it against a reference spectrum using computational methods like the one reported by Böcker and Mäkinen [6]. However, while working with MSI data, it is highly likely that the lock-mass peak is not present in all the spectra that have been acquired in a single experiment, making recalibration difficult for those spectra where this peak is absent. Secondly, for a MSI dataset there is no single reference mass spectrum against which all the spectra can be recalibrated. Our recalibration method exploits similarities amongst peaklist pairs and considers the spatial dimension of MSI data, to reduce the mass error, in an iterative manner. The last iteration generates a consensus (reference) spectrum against which all the peaklists are recalibrated. Since there are multiple peaklists in a single MSI dataset, it is important that all the peaklists are ordered in an efficient manner to minimize the growing mass error in the consensus spectrum. For ordering the peaklists within our method, we also presented three approaches namely: minimum spanning tree (MST), topological greedy (TG) and crystal growth (CG).

We evaluated our method against multiple datasets acquired using the MSI of intact *D. melanogaster* fly. These datasets had considerable mass shifts with variations in different regions of the fly body, mainly caused due to the sample topology. On applying our method to the preprocessed spectra from these datasets, we found that mass error in the data was strongly reduced, with CG ordering performing slightly better than MST and TG ordering. We also found that standard deviation of the mass error decreased significantly post-recalibration on using both CG and MST ordering. When the running times were compared, we observed that MST ordering took the longest for all the datasets whereas TG ordering was the fastest. Looking at the recalibration results and running times for the datasets used, we would suggest to use CG ordering, since it shows the best trade-off between quality of mass-shift correction and running-time performance. The main advantage of our approach is that, it is not dependent on the presence of a lock-mass peak nor does it require any reference peak list. This said, we would like to point out that our method can correct the statistical error present in the data. However, it cannot guess where the actual peak should be present in the mass spectrum. To get the correct position of the peaks, the user would have to provide standard reference masses to compare the mass spectra against it.

We also tested the impact of applying our recalibration method to raw mass spectra. However, doing this did not provide accurate recalibration results since the datasets selected for this study had numerous noisy peaks and weak analyte signals. Our tests confirmed that performing data preprocessing and peak-picking on raw mass spectra prior to applying recalibration steps yields best results for such data. We assume that data preprocessing steps would not be necessary in the case of high-resolution imaging datasets. To achieve best recalibration results, it is also important to select appropriate mass thresholds used at various steps in the recalibration method. The selection of these thresholds mainly depends on the quality of the acquired data in terms of mass resolution and accuracy. From the design of our recalibration method and from our experience with real data, we would argue that our method is rather robust against slight variations of preprocessing parameters. This is mainly due to outlier exclusion, which is performed by the method. It is also possible that the peaks combined in the consensus spectrum are not the same molecules but unresolved isobaric species; however, by the design of our method, which can detect and exclude outliers, it is unlikely that this will result in bad recalibration.

Although the method presented here has demonstrated its effectiveness, we believe that it can be further improved, for instance, by incorporating automated selection of data-dependent parameters to perform recalibration. This would minimize the current user-input requirements. Another possible approach could be inspired from the method proposed by McCombie *et al.* [116], that can objectively identify spatial correlations of mass spectra in a dataset to define regions. Mass-shift correction could be then performed by ordering spectra in these regions individually. It would be really interesting to examine and compare the performance of this approach with the existing CG ordering approach. It is highly likely that this approach may enhance mass-shift correction for datasets with varying mass error in different spatial regions of the imaged sample.

When larger MSI datasets are used for recalibration, we expect the total running time of our method to increase more than exponentially with the dataset size. A possible approach to decrease this could be by allowing users to select a narrower mass range of interest to perform recalibration.

Based on the implementation of our recalibration method, we also presented a Java-based tool called `MSICorrect`, in Chapter 5. The various functionalities in this tool include: visualizing distribution of different compounds in the form of ion intensity maps, visualizing the mean mass spectrum and all the mass spectra individually, mass shift correction, viewing the amount of correction performed by different ordering approaches as well as export of the recalibrated mass spectra.

In Chapter 6, we introduced our unsupervised pattern extraction and image ranking method. Since, MSI offers the unique advantage of obtaining the spatial and chemical information of many compounds simultaneously within in a single experiment, it can be extremely useful for untargeted analysis. However, spatial information in MSI data is still an under-utilized resource in the computational analysis of these data. An ideal approach would be the one that provides a list of mass values that exhibit a visibly significant ion distribution pattern from thousands of mass values measured, similar to what a human eye is able to perceive for a hand-full of images. However, this is not so easy to achieve computationally, since ion distribution patterns can be extremely diverse and their biological relevance cannot be defined, without expert judgement. In this chapter we presented our simplistic approach to extract ion species of possible importance from MSI data by ranking them based on their spatial pattern abundance. With this presented method, we took our first step in the direction towards automated and unsupervised characterization of biologically relevant patterns in MSI data. We demonstrated how our method could help to obtain a quick overview of a MSI dataset, by providing a ranked list of *m/z* values for a dataset, based on their spatial pattern abundance. From this list, the *m/z* values with a higher rank may act as candidates for further exploration. Our approach uses morphological transformations to extract the spatial pattern-related information present in an ion map and based on this provides the map a numerical score known as the *Image Content (IC) score*. Then it ranks all the ion intensity maps present in the dataset based on the IC score, allotting ion map with the highest IC score as rank=1. We also presented an extension to this approach, by grouping mass values that exhibit similar spatial distribution patterns. Our approach performs this by first calculating pair-wise spatial similarity between ion intensity maps using the *Structural Similarity (SSIM) index* to create a similarity matrix. Using this matrix as input, hierarchical clustering is performed to generate the groups.

We evaluated our image ranking approach on different levels. First we used a subset of 40 ion intensity maps, manually selected from a mouse urinary bladder MSI dataset. Our method was able to order the ion intensity maps, giving higher ranks to the ones with high spatial pattern abundance and lowest ranks to those with no spatial pattern on the imaged region. On performing spatial similarity-based grouping, we found that about 85% of the ion intensity maps were grouped correctly. Further, we used a full MSI dataset comprising of 300 ion intensity maps to evaluate our approach. We found that, even for this dataset, our approach was able to rank the mass values based on their spatial pattern abundance. We demonstrated this by highlighting the top 10 and top 50 ranked mass values in the mean mass spectrum of the dataset. We found that about 80% of the top 10 ranked mass values displayed visible spatial patterns and constituted of significant signals present in the mean mass spectrum. Spatial similarity-based grouping performed on the top 10 ranked mass values also provided desired groups based on the patterns visible in the ion maps. Our grouping approach was clearly able to distinguish between ion maps with less pattern and those with abundant spatial pattern, since these maps were placed in separate groups. For grouping the top 50 ranked mass values, about 74% of ion intensity maps were grouped correctly based on their visible spatial pattern. We showed that the IC score generation was very efficient in terms of running time, however this was not the case when performing spatial similarity-based grouping. In this case, the running time increased almost exponentially with increase in the number of input ion maps. We observed that a large fraction of the total running time for spatial similarity-based grouping stems from calculating pair-wise similarity indices between the ion maps. With the current approach, in order to decrease the running time for grouping the masses, one could select the top few ranked ion maps and perform grouping using these as input.

We also noticed that when the distribution of an ion species is highly localized, comprising only a few pixels of the overall imaged section (e.g. within a tissue sample where a drug has been injected in a specific region), our approach provides it a lower rank. This is mainly because the rank is provided based on the IC score which is the area fraction of the extracted pattern over the total imaged area. The IC score introduced in this chapter still holds immense scope for improvements, to make it conceptually sound in order to accommodate varying spatial patterns present in ion intensity maps. We hypothesize that using finer edge detection approaches and weighing local features within an ion intensity map can help improve the IC score. We need to remember that the proposed approach is just a first step towards automated pattern extraction and there is clearly much work to be done. There can be many possible directions to enhance our approach. For instance, in this setting we assume that our approach could highly benefit by the use of texture-based features for pattern classification and extraction [232] from an ion map. Another important aspect to consider for future evaluation would be to ensure that the method scales well with extremely high dimensional MSI datasets and allows for much larger data to be read into memory at a given point.

Towards the future, we believe that this approach could also be used for MSI combined with tandem MS (MS/MS) studies. Acquiring MS2 information for all major peaks acquired in MSI can be extremely time-consuming. Looking at the *m/z* lists and spatial similarity-based groups generated by our approach, one can be selective for the identification of compounds.

Finally, in Chapter 7 we demonstrated the ability of precise data analysis method selection to uncover the underlying chemistry of a biological MSI dataset. To mitigate the difficulties in handling elaborate MSI data and to quantitatively as well as qualitatively interpret them, many computational approaches have been developed. Preprocessing methods mainly help to reduce experimental variance within a dataset and help to clean the data of any noise, making it ready for multivariate analysis that helps in identifying significant features present in the data. In this chapter we presented the application of a protocol that combined bioinformatics and chemometric tools to analyze TOF-SIMS imaging data acquired from the leaf surface of *Populus trichocarpa*.

Based on our analysis, we were able to correlate the crystals observed in the ion intensity maps of the imaged leaf surface to the spheroidal grains seen on the scanning electron microscopy(SEM) image (Figure C.1(b)) obtained independently. While applying MVA approaches to this dataset we noticed that for better visualization and classification of regions with distinct chemical composition, the application of DCA and spatially-aware *k*-means clustering proved to be particularly useful. This study used tools that were able to provide insights into the chemical composition of areas showing distinct segregation/co-localization on the leaf surface. We would like to highlight that, the protocol presented in this chapter could be easily expanded to other MSI datasets as well.

Given a MSI dataset, acquired with an aim to perform untargeted analysis, we can apply the methods presented in this thesis one after the other to find significant masses of interest, that can help us discover more about the sample. Our first approach will take care of any mass shifts present in the data. Later to get an overview of what could be biologically significant in the dataset and the masses that exhibit similar spatial distribution, our pattern extraction and image ranking approach can be applied. Finally, the multivariate analysis workflow can be applied to extract significant chemical and correlation information from the dataset.

Nonetheless, the future looks bright for MSI. It is definitely a powerful tool that researchers can add to their analytical toolkit. The continuous increase in spatial and spectral resolution has further enhanced this technique along with increasing the need for robust computational analysis approaches. We hope that the methods presented in this thesis are significant steps in the right direction and can easily be a part of bigger analysis pipelines to provide insights into the complex metabolic processes at the spatial level.

Bibliography

[1] P. Kulkarni, F. Kaftan, P. Kynast, A. Svatoš, and S. Böcker. Correcting mass shifts: A lock-mass-free recalibration procedure for mass spectrometry imaging data. *Anal Bioanal Chem*, 405(25):7603–7613, Oct. 2015.

[2] F. Kaftan, V. Vrkoslav, P. Kynast, P. Kulkarni, et al. Mass spectrometry imaging of surface lipids on intact *Drosophila melanogaster* flies. *J Mass Spectrom*, 49(3): 223–232, 2014.

[3] P. Kulkarni, M. Dost, O. D. Bulut, A. Welle, S. Böcker, W. Boland, and A. Svatoš. Secondary ion mass spectrometry imaging and multivariate data analysis reveal co-aggregation patterns of populus trichocarpa leaf surface compounds on a micrometer scale. *The Plant Journal*, October 2017. Accepted for publication.

[4] B. Bartels, P. Kulkarni, N. Danz, S. Böcker, H. P. Saluz, and A. Svatoš. Mapping metabolites from rough terrain: laser ablation electrospray ionization on non-flat samples. *RSC Adv*, 7:9045–9050, 2017.

[5] P. Kynast. Segmentierung von Massenspektrometriebildern. Master's thesis, Friedrich Schiller University Jena, 2012.

[6] S. Böcker and V. Mäkinen. Combinatorial approaches for mass spectra recalibration. *IEEE/ACM Trans Comput Biology Bioinform*, 5(1):91–100, 2008.

[7] M. P. Sawicki, G. Samara, M. Hurwitz, and E. Passaro. Human genome project. *The American Journal of Surgery*, 165(2):258 – 264, 1993.

[8] K. Dettmer, P. A. Aronov, and B. D. Hammock. Mass spectrometry-based metabolomics. *Mass Spectrom Rev*, 26(1):51–78, 2007.

[9] Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[10] S. D. Patterson and R. H. Aebersold. Proteomics: the first decade and beyond. *Nature genetics*, 33(3s):311, 2003.

[11] R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in Biotechnology*, 22(5):245–252, 2004.

[12] A. Zhang, H. Sun, P. Wang, Y. Han, and X. Wang. Modern analytical techniques in metabolomics analysis. *Analyst*, 137(2):293–300, 2012.

[13] X. Feng, X. Liu, Q. Luo, and B.-F. Liu. Mass spectrometry in systems biology: An overview. *Mass Spectrometry Reviews*, 27(6):635–660, 2008.

[14] E. W. Deutsch, L. Mendoza, D. Shteynberg, J. Slagel, Z. Sun, and R. L. Moritz. Trans-proteomic pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *PROTEOMICS-Clinical Applications*, 9(7-8):745–754, 2015.

[15] M. Brown, W. B. Dunn, P. Dobson, Y. Patel, et al. Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst*, 134(7):1322–1332, 2009.

[16] M. Garnier, E. J. Dufourc, and B. Larijani. Characterisation of lipids in cell signalling and membrane dynamics by nuclear magnetic resonance spectroscopy and mass spectrometry. *Signal Transduction*, 6(2):133–143, 2006.

[17] C. Hu and G. Xu. Mass-spectrometry-based metabolomics analysis for foodomics. *TrAC Trends in Analytical Chemistry*, 52:36–46, 2013. Modern Food Analysis and Foodomics.

[18] M. L. Kraft, P. K. Weber, M. L. Longo, I. D. Hutcheon, and S. G. Boxer. Phase separation of lipid membranes analyzed with high-resolution secondary ion mass spectrometry. *Science*, 313(5795):1948–1951, 2006.

[19] R. M. Caprioli, T. B. Farmer, and J. Gile. Molecular imaging of biological samples: Localization of peptides and proteins using maldi-tof ms. *Analytical Chemistry*, 69(23):4751–4760, 1997.

[20] J. D. Watrous and P. C. Dorrestein. Imaging mass spectrometry in microbiology. *Nat Rev Microbiol*, 9(9):683–694, 2011.

[21] C. Bedia, R. Tauler, and J. Jaumot. Compression strategies for the chemometric analysis of mass spectrometry imaging data. *Journal of Chemometrics*, 30(10): 575–588, 2016.

[22] K. D. Bemis, A. Harry, L. S. Eberlin, C. R. Ferreira, et al. Probabilistic segmentation of mass spectrometry (ms) images helps select important ions and characterize confidence in the resulting segments. *Molecular & Cellular Proteomics*, 15 (5):1761–1772, 2016.

[23] J. M. Fonville, C. Carter, O. Cloarec, J. K. Nicholson, J. C. Lindon, J. Bunch, and E. Holmes. Robust data processing and normalization strategy for maldi mass spectrometric imaging. *Anal. Chem.*, 84(3):1310–1319, Feb. 2012.

[24] A. M. Race, R. T. Steven, A. D. Palmer, I. B. Styles, and J. Bunch. Memory efficient principal component analysis for the dimensionality reduction of large mass spectrometry imaging data sets. *Analytical chemistry*, 85(6):3071–3078, 2013.

[25] G. Siuzdak. *Mass Spectrometry for Biotechnology.* Academic Press, 1996.

[26] J. H. Gross. *Mass Spectrometry: A Textbook.* Springer, New York, 2nd edition, 2011.

[27] G. G. Hammes. *Spectroscopy for the biological sciences.* John Wiley & Sons, 2005.

[28] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422: 198–207, 2003.

[29] D. J. Harvey. Matrix-assisted laser desorption/ionization mass spectrometry of carbohydrates. *Mass Spectrometry Reviews*, 18(6):349–450, 1999.

[30] E. Nordhoff, F. Kirpekar, and P. Roepstorff. Mass spectrometry of nucleic acids. *Mass spectrometry reviews*, 15(2):67–138, 1996.

[31] R. C. Murphy. *Mass spectrometry of lipids.* Plenum Press, 1993.

[32] M. R. Meyer and H. H. Maurer. Current applications of high-resolution mass spectrometry in drug metabolism studies. *Anal Bioanal Chem*, 403(5):1221–1231, 2012.

[33] W. M. Niessen and D. Falck. Introduction to mass spectrometry, a tutorial. *Analyzing Biomolecular interaction by mass spectrometry*, pages 1–54, 2015.

[34] S. E. Van Bramer. An introduction to mass spectrometry. *Lecture Notes*, 1997.

[35] E. de Hoffmann and V. Stroobant. *Mass Spectrometry: Principles and Applications.* Wiley-Interscience, third edition, 2007.

[36] G. Hart-Smith and S. J. Blanksby. *Mass Analysis*, chapter 1, pages 5–32. Wiley-VCH Verlag GmbH & Co. KGaA, 2011. ISBN 9783527641826.

[37] K. K. Murray, R. K. Boyd, M. N. Eberlin, G. J. Langley, L. Li, and Y. Naito. Definitions of terms relating to mass spectrometry (iupac recommendations 2013). *Pure and Applied Chemistry*, 85(7):1515–1609, 2013.

[38] R. Thomas. *Practical guide to ICP-MS: a tutorial for beginners.* CRC press, 2013.

[39] S. A. McLuckey and J. M. Wells. Mass analysis at the advent of the 21st century. *Chemical Reviews*, 101(2):571–606, 2001.

[40] F. Gunzer and J. Grotemeyer. Chapter two - recent developments in time-of-flight mass spectrometry. In P. W. Hawkes, editor, *Advances in Imaging and Electron Physics*, volume 188 of *Advances in Imaging and Electron Physics*, pages 25 – 78. Elsevier, 2015.

[41] J. Greaves and J. Roboz. *Mass Spectrometry for the Novice.* CRC Press Taylor & Francis Group, Boca Raton, FL, 2013.

[42] S. S. Rubakhin, J. C. Jurchen, E. B. Monroe, and J. V. Sweedler. Imaging mass spectrometry: fundamentals and applications to drug discovery. *Drug Discov Today*, 10(12):823–837, 2005.

[43] T. Alexandrov and J. H. Kobarg. Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. *Bioinformatics*, 27 (13):i230–i238, 2011.

[44] E. Esquenazi, Y.-L. Yang, J. Watrous, W. H. Gerwick, and P. C. Dorrestein. Imaging mass spectrometry of natural products. *Natural product reports*, 26(12): 1521–1534, 2009.

[45] L. A. McDonnell and R. M. Heeren. Imaging mass spectrometry. *Mass Spectrometry Reviews*, 26(4):606–643, 2007.

[46] K. Chughtai and R. M. A. Heeren. Mass spectrometric imaging for biomedical tissue analysis. *Chemical Reviews*, 110(5):3237–3277, 2010.

[47] S. L. Luxembourg, T. H. Mize, L. A. McDonnell, and R. M. A. Heeren. High spatial resolution mass spectrometric imaging of peptide and protein distributions on a surface. *Analytical Chemistry*, 76(18):5339 – 5344, 2004.

[48] C. Wu, A. L. Dill, L. S. Eberlin, R. G. Cooks, and D. R. Ifa. Mass spectrometry imaging under ambient conditions. *Mass Spectrometry Reviews*, 32(3):218–243, 2013.

[49] P. Nemes and A. Vertes. Atmospheric-pressure molecular imaging of biological tissues and biofilms by laesi mass spectrometry. *JoVE (Journal of Visualized Experiments)*, pages 2097–2097, 2010.

[50] A. Bodzon-Kulakowska and P. Suder. Imaging mass spectrometry: Instrumentation, applications, and combination with other visualization techniques. *Mass Spectrometry Reviews*, 35(1):147–169, 2016.

[51] S. A. Schwartz and R. M. Caprioli. Imaging mass spectrometry: viewing the future. *Mass Spectrometry Imaging: Principles and Protocols*, pages 3–19, 2010.

[52] K. Schwamborn and R. M. Caprioli. Molecular imaging by mass spectrometry - looking beyond classical histology. *Nature Reviews Cancer*, 10(9):639–646, 2010.

[53] R. Zenobi and R. Knochenmuss. Ion formation in MALDI mass spectrometry. *Mass Spectrom Rev*, 17:337–366, 1998.

[54] J. D. Watrous, T. Alexandrov, and P. C. Dorrestein. The evolving field of imaging mass spectrometry and its impact on future biological research. *Journal of Mass Spectrometry*, 46(2):209–222, 2011.

[55] C. Simó, A. Cifuentes, and V. García-Cañas. *Fundamentals of advanced omics technologies: from genes to metabolites*, volume 63. Newnes, 2014.

[56] K. Shrivas, T. Hayasaka, N. Goto-Inoue, Y. Sugiura, N. Zaima, and M. Setou. Ionic matrix for enhanced maldi imaging mass spectrometry for identification of phospholipids in mouse liver and cerebellum tissue sections. *Analytical Chemistry*, 82(21):8800–8806, 2010.

[57] D. Sturtevant, Y.-J. Lee, and K. D. Chapman. Matrix assisted laser desorption/ionization-mass spectrometry imaging (maldi-msi) for direct visualization of plant metabolites in situ. *Current Opinion in Biotechnology*, 37:53 – 60, 2016. Food biotechnology Plant biotechnology.

[58] A. Benninghoven, F. Rudenauer, and H. W. Werner. *Secondary ion mass spectrometry: basic concepts, instrumental aspects, applications and trends.* John Wiley and Sons, New York, NY, 1987.

[59] B. K. Kaletaş, I. M. van der Wiel, J. Stauber, C. Güzel, J. M. Kros, T. M. Luider, and R. Heeren. Sample preparation issues for tissue imaging by imaging ms. *Proteomics*, 9(10):2622–2633, 2009.

[60] 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.

[61] J. M. Fonville, C. L. Carter, L. Pizarro, R. T. Steven, et al. Hyperspectral visualization of mass spectrometry imaging data. *Analytical Chemistry*, 85(3):1415–1423, 2013.

[62] B. A. Boughton, D. Thinagaran, D. Sarabia, A. Bacic, and U. Roessner. Mass spectrometry imaging for plant biology: a review. *Phytochemistry Reviews*, 15(3): 445–488, 2016.

[63] T. Alexandrov. Maldi imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC Bioinformatics*, 13(Suppl 16):S11–S11, Nov. 2012.

[64] C. W. Magee, R. E. Honig, and C. A. Evans. *Depth Profiling by SIMS: Depth Resolution, Dynamic Range and Sensitivity*, pages 172–185. Springer Berlin Heidelberg, Berlin, Heidelberg, 1982. ISBN 978-3-642-88152-7.

[65] S. Khatib-Shahidi, M. Andersson, J. L. Herman, T. A. Gillespie, and R. M. Caprioli. Direct molecular analysis of whole-body animal tissue sections by imaging maldi mass spectrometry. *Anal Chem*, 78(18):6448–6456, 2006.

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

[67] Y. Schober, S. Guenther, B. Spengler, and A. Ro mpp. Single cell matrix-assisted laser desorption/ionization mass spectrometry imaging. *Anal Chem*, 84(15):6293–6297, 2012.

[68] K. J. Boggio, E. Obasuyi, K. Sugino, S. B. Nelson, N. Y. Agar, and J. N. Agar. Recent advances in single-cell maldi mass spectrometry imaging and potential clinical impact. *Expert Rev Proteomics*, 8(5):591–604, Oct. 2011.

[69] G. Marko-Varga, T. E. Fehniger, M. Rezeli, B. Dme, T. Laurell, and kos Vgvri. Drug localization in different lung cancer phenotypes by {MALDI} mass spectrometry imaging. *Journal of Proteomics*, 74(7):982–992, 2011.

[70] E. H. Seeley, P. S. Cantrell, C. M. Walsh, S. Wen, et al. Mass spectrometry imaging determines biomarkers of early adaptive precision drug resistance in lung cancer. *Cancer Research*, 76(14 Supplement):3874–3874, 2016.

[71] L. H. S. Mendis, A. C. Grey, R. L. M. Faull, and M. A. Curtis. Hippocampal lipid differences in alzheimer's disease: a human brain study using matrix-assisted laser desorption/ionization-imaging mass spectrometry. *Brain and Behavior*, 2016.

[72] J. Pierson, J. L. Norris, H.-R. Aerni, P. Svenningsson, R. M. Caprioli, and P. E. Andrén. Molecular profiling of experimental parkinson's disease: direct analysis of peptides and proteins on brain tissue sections by maldi mass spectrometry. *Journal of proteome research*, 3(2):289–295, 2004.

[73] D. Touboul, H. Piednoël, V. Voisin, S. De La Porte, A. Brunelle, F. Halgand, and O. Laprévote. Changes in phospholipid composition within the dystrophic muscle by matrix-assisted laser desorption/ionization mass spectrometry and mass spectrometry imaging. *European Journal of Mass Spectrometry*, 10(5):657–664, 2004.

[74] H. Meistermann, J. L. Norris, H.-R. Aerni, D. S. Cornett, et al. Biomarker discovery by imaging mass spectrometry transthyretin is a biomarker for gentamicin-induced nephrotoxicity in rat. *Molecular & Cellular Proteomics*, 5(10):1876–1886, 2006.

[75] E. R. Amstalden van Hove, D. F. Smith, and R. Heeren. A concise review of mass spectrometry imaging. *J Chromatogr A*, 1217(25):3946–3954, 2010.

[76] R. Van de Plas, J. Yang, J. Spraggins, and R. M. Caprioli. Fusion of mass spectrometry and microscopy: a multi-modality paradigm for molecular tissue mapping. *Nature methods*, 12(4):366–372, 2015.

[77] S. Chughtai, K. Chughtai, B. Cillero-Pastor, A. Kiss, P. Agrawal, L. MacAleese, and R. M. Heeren. A multimodal mass spectrometry imaging approach for the study of musculoskeletal tissues. *International Journal of Mass Spectrometry*, 325-327:150–160, 2012.

[78] C. D. Wijetunge, I. Saeed, S. K. Halgamuge, B. Boughton, and U. Roessner. Unsupervised learning for exploring maldi imaging mass spectrometry. In *7th International Conference on Information and Automation for Sustainability*, pages 1–6, Dec 2014.

[79] L. J. Winderbaum, I. Koch, O. J. Gustafsson, S. Meding, P. Hoffmann, et al. Feature extraction for proteomics imaging mass spectrometry data. *The Annals of Applied Statistics*, 9(4):1973–1996, 2015.

[80] A. D. Palmer. *Information processing for mass spectrometry imaging.* PhD thesis, University of Birmingham, December 2014.

[81] A. Römpp, T. Schramm, A. Hester, I. Klinkert, et al. imzml: imaging mass spectrometry markup language: a common data format for mass spectrometry imaging. *Data Mining in Proteomics: From Standards to Applications*, 696:205–224, October 2011.

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

[83] H. Shin, M. Mutlu, J. M. Koomen, and M. K. Markey. Parametric power spectral density analysis of noise from instrumentation in maldi tof mass spectrometry. *Cancer Informatics*, 3:219–30, 2007.

[84] I. Eidhammer, K. Flikka, L. Martens, and S.-O. Mikalsen. *Mass Spectrometry - MALDI-TOF*, pages 81–95. John Wiley & Sons, Ltd, 2007. ISBN 9780470724309.

[85] M. Cannataro, P. H. Guzzi, T. Mazza, G. Tradigo, and P. Veltri. *On the Preprocessing of Mass Spectrometry Proteomics Data*, pages 127–131. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-33184-1.

[86] R. Casadonte and R. M. Caprioli. Proteomic analysis of formalin-fixed paraffin-embedded tissue by maldi imaging mass spectrometry. *Nat. Protocols*, 6(11):1695–1709, Nov. 2011.

[87] J. Engel, J. Gerretzen, E. Szymanska, J. J. Jansen, G. Downey, L. Blanchet, and L. M. Buydens. Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry*, 50:96–106, 2013.

[88] C. Yang, Z. He, and W. Yu. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC Bioinf*, 10:4, 2009.

[89] A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem*, 36(8):1627–1639, 1964.

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

[91] P. Monchamp, L. Andrade-Cetto, J. Y. Zhang, and R. Henson. Signal processing methods for mass spectrometry. *Systems Bioinformatics: An Engineering Case-Based Approach, Artech House Publishers*, 2007.

[92] V. Barclay, R. Bonner, and I. Hamilton. Application of wavelet transforms to experimental spectra: smoothing, denoising, and data set compression. *Analytical Chemistry*, 69(1):78–90, 1997.

[93] A. V. Oppenheim and R. W. Schafer. *Discrete-time signal processing*. Pearson Higher Education, 2010.

[94] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

[95] A. C. Sauve and T. P. Speed. Normalization, baseline correction and alignment of high-throughput mass spectrometry data. In *Gensips*, 2004.

[96] H. Shin, M. P. Sampat, J. M. Koomen, and M. K. Markey. Wavelet-based adaptive denoising and baseline correction for MALDI TOF MS. *OMICS*, 14(3):283–295, 2010.

[97] B. Williams, S. Cornett, B. Dawant, A. Crecelius, B. Bodenheimer, and R. Caprioli. An algorithm for baseline correction of maldi mass spectra. In *Proceedings of the 43rd Annual Southeast Regional Conference - Volume 1*, ACM-SE 43, pages 137–142, New York, NY, USA, 2005. ISBN 1-59593-059-0.

[98] C. Ryan, E. Clayton, W. Griffin, S. Sie, and D. Cousens. Snip, a statistics-sensitive background treatment for the quantitative analysis of pixe spectra in geoscience applications. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 34(3):396 – 402, 1988.

[99] L. Andrade and E. S. Manolakos. Signal background estimation and baseline correction algorithms for accurate dna sequencing. *Journal of VLSI signal processing systems for signal, image and video technology*, 35(3):229–243, 2003.

[100] D. P. Enot, B. Haas, and K. M. Weinberger. *Bioinformatics for Mass Spectrometry-Based Metabolomics*, pages 351–375. Humana Press, Totowa, NJ, 2011. ISBN 978-1-61779-027-0.

[101] Y. Sugiura, I. Yao, and M. Setou. *Imaging Mass Spectrometry (IMS) for Biological Application*, pages 41–83. John Wiley & Sons, Inc., 2012. ISBN 9781118180730.

[102] S.-O. Deininger, D. S. Cornett, R. Paape, M. Becker, et al. Normalization in MALDI-TOF imaging datasets of proteins: practical considerations. *Analytical and Bioanalytical Chemistry*, 401(1):167–181, Mar. 2011.

[103] J. Inczedy, T. Lengyel, and A. M. Ure. *Compendium of analytical nomenclature*. Blackwell Science, 1998.

[104] O. Haglund. *Qualitative comparison of normalization approaches in MALDI-MS*. Datavetenskap och kommunikation, Kungliga Tekniska högskolan, 2008.

[105] E. A. Jones, S.-O. Deininger, P. C. Hogendoorn, A. M. Deelder, and L. A. McDonnell. Imaging mass spectrometry statistical analysis. *J Proteomics*, 75(16): 4962–4989, 2012.

[106] E. T. Fung and C. Enderwick. Proteinchip clinical proteomics: computational challenges and solutions. *Biotechniques*, 32(Suppl 1):34–41, 2002.

[107] J. L. Norris, D. S. Cornett, J. A. Mobley, M. Andersson, E. H. Seeley, P. Chaurand, and R. M. Caprioli. Processing maldi mass spectra to improve mass spectral direct tissue analysis. *Int J Mass Spectrom*, 260(2):212–221, 2007.

[108] K. A. Veselkov, L. K. Vingara, P. Masson, S. L. Robinette, et al. Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Anal. Chem.*, 83(15): 5864–5872, Aug. 2011.

[109] P. Rfols, D. Vilalta, J. Brezmes, N. Caellas, et al. Signal preprocessing, multivariate analysis and software tools for ma(ldi)-tof mass spectrometry imaging for biological applications. *Mass Spectrometry Reviews*, 2016.

[110] W. Yu, B. Wu, N. Lin, K. Stone, K. Williams, and H. Zhao. Detecting and aligning peaks in mass spectrometry data with applications to {MALDI}. *Computational Biology and Chemistry*, 30(1):27 – 38, 2006.

[111] T. Hayasaka, N. Goto-Inoue, M. Ushijima, I. Yao, et al. Development of imaging mass spectrometry (ims) dataset extractor software, ims convolution. *Analytical and Bioanalytical Chemistry*, 401(1):183–193, 2011.

[112] T. Alexandrov, M. Becker, S.-O. Deininger, G. n. Ernst, et al. Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering. *J Proteome Res*, 9(12):6535–6546, 2010.

[113] S. Har-Peled, P. Indyk, and R. Motwani. Approximate nearest neighbor: Towards removing the curse of dimensionality. *Theory of Computing*, 8(14):321–350, 2012.

[114] J. H. Friedman. On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997.

[115] R. Bellman. Curse of dimensionality. *Adaptive control processes: a guided tour. Princeton, NJ*, 1961.

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

[117] L. A. Klerk, A. Broersen, I. W. Fletcher, R. van Liere, and R. M. 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(2-3):222–236, 2007.

[118] M. Hanselmann, M. Kirchner, B. Y. Renard, E. R. Amstalden, K. Glunde, R. M. A. Heeren, and F. A. Hamprecht. Concise representation of mass spectrometry images by probabilistic latent semantic analysis. *Anal. Chem.*, 80(24):9649–9658, Dec. 2008.

[119] A. Malik, A. de Juan, and R. Tauler. Multivariate curve resolution: A different way to examine chemical data. In *ACS Symposium Series*, volume 1199, pages 95–128–. American Chemical Society, Jan. 2015.

[120] J. L. S. Lee and I. S. Gilmore. *The Application of Multivariate Data Analysis Techniques in Surface Analysis*, pages 563–612. John Wiley & Sons, Ltd, 2009. ISBN 9780470721582.

[121] M. Seah. Summary of iso/tc 201 standard: Iso 18115-1: 2010–surface chemical analysis–vocabulary–general terms and terms used in spectroscopy. *Surface and Interface Analysis*, 44(5):618–620, 2012.

[122] B. Vaezian, C. R. Anderton, and M. L. Kraft. Discriminating and imaging different phosphatidylcholine species within phase-separated model membranes by principal component analysis of tof-secondary ion mass spectrometry images. *Anal. Chem.*, 82(24):10006–10014, Dec. 2010.

[123] M. El Ayed, D. Bonnel, R. Longuespée, C. Castelier, et al. Maldi imaging mass spectrometry in ovarian cancer for tracking, identifying, and validating biomarkers. *Medical Science Monitor*, 16(8):BR233–BR245, 2010.

[124] A. L. Dill, L. S. Eberlin, A. B. Costa, C. Zheng, et al. Multivariate statistical identification of human bladder carcinomas using ambient ionization imaging mass spectrometry. *Chemistry-A European Journal*, 17(10):2897–2902, 2011.

[125] K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2 (11):559–572, 1901.

[126] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[127] W. R. Zwick and W. F. Velicer. Factors influencing four rules for determining the number of components to retain. *Multivariate behavioral research*, 17(2):253–269, 1982.

[128] W. H. Lawton and E. A. Sylvestre. Self modeling curve resolution. *Technometrics*, 13(3):617–633, 1971.

[129] J. L. S. Lee, I. S. Gilmore, and M. P. Seah. Quantification and methodology issues in multivariate analysis of tof-sims data for mixed organic systems. *Surface and Interface Analysis*, 40(1):1–14, 2008.

[130] N. B. Gallagher, J. M. Shaver, E. B. Martin, J. Morris, B. M. Wise, and W. Windig. Curve resolution for multivariate images with applications to tof-sims and raman. *Chemometrics and Intelligent Laboratory Systems*, 73(1):105 – 117, 2004. 8th Scandinavian Symposium on Chemometrics (SSC8), Mariehamn, Aland, Finland 14-18 June 2003.

[131] J. L. S. Lee, I. S. Gilmore, I. W. Fletcher, and M. P. Seah. Multivariate image analysis strategies for tof-sims images with topography. *Surface and Interface Analysis*, 41(8):653–665, 2009.

[132] J. Jaumot and R. Tauler. Potential use of multivariate curve resolution for the analysis of mass spectrometry images. *Analyst*, 140(3):837–846, 2015.

[133] R. Wehrens. *Chemometrics with R: multivariate data analysis in the natural sciences and life sciences*. Springer Science & Business Media, 2011.

[134] R. Tauler, A. Smilde, and B. Kowalski. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *Journal of Chemometrics*, 9(1):31–58, 1995.

[135] C. Rodrguez-Rodrguez, J. M. Amigo, J. Coello, and S. Maspoch. An introduction to multivariate curve resolution-alternating least squares: Spectrophotometric study of the acidbase equilibria of 8-hydroxyquinoline-5-sulfonic acid. *J. Chem. Educ.*, 84(7):1190–, July 2007.

[136] J. Mendieta, M. Daz-Cruz, M. Esteban, and R. Tauler. Multivariate curve resolution: A possible tool in the detection of intermediate structures in protein folding. *Biophysical Journal*, 74(6):2876–2888, June 1998.

[137] H. Motegi, Y. Tsuboi, A. Saga, T. Kagami, et al. Identification of reliable components in multivariate curve resolution-alternating least squares (mcr-als): a data-driven approach across metabolic processes. *Scientific Reports*, 5:15710–, Nov. 2015.

[138] W. Windig and J. Guilment. Interactive self-modeling mixture analysis. *Anal. Chem.*, 63(14):1425–1432, July 1991.

[139] B. O. Budevska, S. T. Sum, and T. J. Jones. Application of multivariate curve resolution for analysis of ft-ir microspectroscopic images of in situ plant tissue. *Applied Spectroscopy*, 57(2):124–131, 2003.

[140] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[141] Y. Zhao and G. Karypis. Clustering in life sciences. *Methods in molecular biology (Clifton, N.J.)*, 224:183–218, jan 2003.

[142] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput Surv*, 31(3):264–323, 1999.

[143] R. Xu and D. Wunsch. *Clustering*, volume 10. John Wiley & Sons, 2008.

[144] I. Koch. *Analysis of multivariate and high-dimensional data*, volume 32. Cambridge University Press, 2013.

[145] B. Mirkin. Choosing the number of clusters. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):252–260, 2011.

[146] L. S. Wolfgang Hrdle. *Cluster Analysis*, pages 271–288. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-72244-1.

[147] H. Joe and J. Ward. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc*, 58:236–244, 1963.

[148] S. Yu, S. Van Vooren, L.-C. Tranchevent, B. De Moor, and Y. Moreau. Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining. *Bioinformatics (Oxford, England)*, 24(16):i119–25, aug 2008.

[149] R. Remesan and J. Mathew. *Model Data Selection and Data Pre-processing Approaches*, pages 41–70. Springer International Publishing, Cham, 2015. ISBN 978-3-319-09235-5.

[150] F. De la Torre and T. Kanade. Discriminative cluster analysis. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 241–248, New York, NY, USA, 2006. ISBN 1-59593-383-2.

[151] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[152] N. Bratchell. Cluster analysis. *Chemometrics and Intelligent Laboratory Systems*, 6(2):105–125, 1989.

[153] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. of Berkeley Symposium on Math. Statist. and Prob.*, pages 281–297, 1967.

[154] B. Spengler. Mass spectrometry imaging of biomolecular information. *Analytical Chemistry*, 87(1):64–82, 2015.

[155] B. Li, D. R. Bhandari, C. Janfelt, A. Rmpp, and B. Spengler. Natural products in glycyrrhiza glabra (licorice) rhizome imaged at the cellular level by atmospheric pressure matrix-assisted laser desorption/ionization tandem mass spectrometry imaging. *The Plant Journal*, 80(1):161–171, 2014.

[156] Y. J. Lee, D. C. Perdian, Z. Song, E. S. Yeung, and B. J. Nikolau. Use of mass spectrometry for imaging metabolites in plants. *The Plant Journal*, 70(1):81–95, 2012.

[157] A. Palmer, P. Phapale, I. Chernyavsky, R. Lavigne, et al. Fdr-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nat Meth*, 14(1):57–60, 2016.

[158] M. Lagarrigue, T. Alexandrov, G. Dieuset, A. Perrin, et al. New analysis workflow for maldi imaging mass spectrometry: Application to the discovery and identification of potential markers of childhood absence epilepsy. *Journal of Proteome Research*, 11(11):5453–5463, 2012.

[159] K. A. Veselkov, R. Mirnezami, N. Strittmatter, R. D. Goldin, et al. Chemoinformatic strategy for imaging mass spectrometry-based hyperspectral profiling of lipid signatures in colorectal cancer. *Proceedings of the National Academy of Sciences*, 111(3):1216–1221, 2014.

[160] L. S. Eberlin, I. Norton, A. L. Dill, A. J. Golby, et al. Classifying human brain tumors by lipid imaging with mass spectrometry. *Cancer Research*, 72(3):645–654, 2012.

[161] C. Affonso, R. J. Sassi, and R. M. Barreiros. Biological image classification using rough-fuzzy artificial neural network. *Expert Systems with Applications*, 42(24): 9482 – 9488, 2015.

[162] Y. Fujimura and D. Miura. Maldi mass spectrometry imaging for visualizing in situ metabolism of endogenous metabolites and dietary phytochemicals. *Metabolites*, 4(2):319, 2014.

[163] L. F. Marvin, M. A. Roberts, and L. B. Fay. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry in clinical chemistry. *Clinica Chimica Acta*, 337(1-2):11–21, 2003.

[164] B. Heijs, S. Holst, I. H. Briaire-de Bruijn, G. W. van Pelt, et al. Multimodal mass spectrometry imaging of n-glycans and proteins from the same tissue section. *Analytical Chemistry*, 88(15):7745–7753, 2016.

[165] J. Hanrieder, O. Karlsson, E. B. Brittebo, P. Malmberg, and A. G. Ewing. Probing the lipid chemistry of neurotoxin-induced hippocampal lesions using multimodal imaging mass spectrometry. *Surface and Interface Analysis*, 46(S1):375–378, 2014.

[166] E. R. Amstalden van Hove, T. R. Blackwell, I. Klinkert, G. B. Eijkel, R. M. Heeren, and K. Glunde. Multimodal mass spectrometric imaging of small molecules reveals distinct spatio-molecular signatures in differentially metastatic breast tumor models. *Cancer Research*, 70(22):9012–9021, 2010.

[167] I. V. Chernushevich, A. V. Loboda, and B. A. Thomson. An introduction to quadrupole-time-of-flight mass spectrometry. *J Mass Spectrom*, 36:849–865, 2001.

[168] W. R. Plass, H. Li, and R. Cooks. Theory, simulation and measurement of chemical mass shifts in {RF} quadrupole ion traps. *International Journal of Mass Spectrometry*, 228:237–267, 2003.

[169] R. E. March and J. F. J. Todd. *Dynamics of Ion Trapping*, chapter 3, pages 73–132. John Wiley & Sons, Inc., 2005. ISBN 9780471717980.

[170] J. Zhang, J. Ma, W. Zhang, C. Xu, Y. Zhu, and H. Xie. Ftdr 2.0: A tool to achieve sub-ppm level recalibrated accuracy in routine LC–ms analysis. *J Proteome Res*, 12(9):3857–3864, 2013.

[171] D. C. Muddiman and A. L. Oberg. Statistical evaluation of internal and external mass calibration laws utilized in Fourier transform ion cyclotron resonance mass spectrometry. *Anal Chem*, 77(8):2406–2414, 2005.

[172] A. Staes, J. Vandenbussche, H. Demol, M. Goethals, et al. Asn3, a reliable, robust, and universal lock mass for improved accuracy in LC–ms and LC–ms/ms. *Anal Chem*, 85(22):11054–11060, 2013.

[173] J. R. Chapman. *Practical organic mass spectrometry: A guide for chemical and biological analysis.* John Wiley & Sons Ltd, West Sussex, England, second edition, 1985.

[174] A. Römpp, S. Guenther, Y. Schober, O. Schulz, Z. Takats, W. Kummer, and B. Spengler. Histology by mass spectrometry: Label-free tissue characterization obtained from high-accuracy bioanalytical imaging. *Angew Chem Int Ed*, 49(22): 3834–3838, 2010.

[175] V. A. Petyuk, N. Jaitly, R. J. Moore, J. Ding, et al. Elimination of systematic mass measurement errors in liquid chromatography-mass spectrometry based proteomics using regression models and a priori partial knowledge of the sample content. *Anal Chem*, 80(3):693–706, 2008.

[176] A. N. Kozhinov, K. O. Zhurov, and Y. O. Tsybin. Iterative method for mass spectra recalibration via empirical estimation of the mass calibration function for Fourier transform mass spectrometry-based petroleomics. *Anal Chem*, 85(13): 6437–6445, 2013.

[177] J. A. Barry, G. Robichaud, and D. C. Muddiman. Mass recalibration of ft-ICr mass spectrometry imaging data using the average frequency shift of ambient ions. *J Am Soc Mass Spectrom*, 24(7):1137–1145, July 2013.

[178] A. V. Tolmachev, M. E. Monroe, N. Jaitly, V. A. Petyuk, J. N. Adkins, and R. D. Smith. Mass measurement accuracy in analyses of highly complex mixtures based upon multidimensional recalibration. *Analytical Chemistry*, 78(24):8374–8385, 2006.

[179] J. Gobom, M. Mueller, V. Egelhofer, D. Theiss, H. Lehrach, and E. Nordhoff. A calibration method that simplifies and improves accurate determination of peptide molecular masses by MALDI-TOF MS. *Anal Chem*, 74(15):3915–3923, 2002.

[180] W. E. Wolski, M. Lalowski, P. Jungblut, and K. Reinert. Calibration of mass spectrometric peptide mass fingerprint data without specific external or internal calibrants. *BMC Bioinf*, 6:203, 2005.

[181] M. A. Stravs, E. L. Schymanski, H. P. Singer, and J. Hollender. Automatic recalibration and processing of tandem mass spectra using formula annotation. *J Mass Spectrom*, 48(1):89–99, 2013.

[182] H. Horai, M. Arita, S. Kanaya, Y. Nihei, et al. MassBank: A public repository for sharing mass spectral data for life sciences. *J Mass Spectrom*, 45(7):703–714, 2010.

[183] E. Dijkstra. A note on two problems in connexion with graphs. *Numer Math*, 1: 269–271, 1959.

[184] J. Y. Yew, K. Dreisewerd, H. Luftmann, J. Müthing, G. Pohlentz, and E. A. Kravitz. A new male sex pheromone and novel cuticular cues for chemical communication in drosophila. *Curr Biol*, 19(15):1245–1254, 2009.

[185] R. J. Bartelt, A. M. Schaner, and L. L. Jackson. cis-vaccenyl acetate as an aggregation pheromone indrosophila melanogaster. *J Chem Ecol*, 11(12):1747–1756, 1985.

[186] L. A. Hammad, B. S. Cooper, N. P. Fisher, K. L. Montooth, and J. A. Karty. Profiling and quantification of drosophila melanogaster lipids using liquid chromatography/mass spectrometry. *Rapid Commun Mass Spectrom*, 25(19):2959–2968, 2011.

[187] S. Urbanek. A fast way to provide r functionality to applications. In *Proceedings of DSC*, volume 2. Citeseer, 2003.

[188] F. Y. Shih. *Image processing and pattern recognition: fundamentals and techniques*. John Wiley & Sons, 2010.

[189] L. A. McDonnell, A. van Remoortere, N. de Velde, R. J. van Zeijl, and A. 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.

[190] T. Alexandrov and A. Bartels. Testing for presence of known and unknown molecules in imaging mass spectrometry. *Bioinformatics*, 29(18):2335–2342, 2013.

[191] T. Ridler and S. Calvard. Picture thresholding using an iterative selection method. *IEEE Transactions on Systems, Man, and Cybernetics*, 8(8):630–632, 1978.

[192] I. Pitas. *Digital image processing algorithms and applications.* John Wiley & Sons, 2000.

[193] J. S. Weszka, R. N. Nagel, and A. Rosenfeld. A threshold selection technique. *IEEE Trans. Comput.*, 23(12):1322–1326, Dec. 1974.

[194] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):62–66, 1979.

[195] M. Sezgin and B. Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, 13(1):146–168, 2004.

[196] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[197] B. Wang and K. J. Ciuffreda. Depth-of-focus of the human eye: Theory and clinical implications. *Survey of Ophthalmology*, 51(1):75 – 85, 2006.

[198] K. Matković, L. Neumann, A. Neumann, T. Psik, and W. Purgathofer. Global contrast factor - a new approach to image contrast. In *Proceedings of the First Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*, Computational Aesthetics'05, pages 159–167, Aire-la-Ville, Switzerland, Switzerland, 2005. ISBN 3-905673-27-4.

[199] J. A. Vizcano, R. G. Ct, A. Csordas, J. A. Dianes, et al. The proteomics identifications (pride) database and associated tools: status in 2013. *Nucleic Acids Research*, 41(D1):D1063–D1069, 2013.

[200] L. Samuels, L. Kunst, and R. Jetter. Sealing plant surfaces: cuticular wax formation by epidermal cells. *Annual review of plant biology*, 59, 2008.

[201] K. Koch, B. Bhushan, and W. Barthlott. Diversity of structure, morphology and wetting of plant surfaces. *Soft Matter*, 4:1943–1963, 2008.

[202] K. Koch and W. Barthlott. Superhydrophobic and superhydrophilic plant surfaces: an inspiration for biomimetic materials. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1893): 1487–1509, 2009.

[203] K. Haas and I. Rentschler. Discrimination between epicuticular and intracuticular wax in blackberry leaves: Ultrastructural and chemical evidence. *Plant Science Letters*, 36(2):143 – 147, 1984.

[204] S. D. Wullschleger, G. A. Tuskan, and S. P. DiFazio. Genomics and the tree physiologist. *Tree physiology*, 22(18):1273–6, dec 2002.

[205] H. Bradshaw, R. Ceulemans, J. Davis, and R. Stettler. Emerging model systems in plant biology: Poplar (populus) as a model forest tree. *Journal of Plant Growth Regulation*, 19(3):306–313, sep 2000.

[206] A. M. Brunner, V. B. Busov, and S. H. Strauss. Poplar genome sequence: functional genomics in an ecologically dominant plant species. *Trends in plant science*, 9(1):49–56, jan 2004.

[207] G. Taylor. Populus: arabidopsis for forestry. do we need a model tree? *Annals of botany*, 90(6):681–9, dec 2002.

[208] D. Post-Beittenmiller. Biochemistry and molecular biology of wax production in plants. *Annual Review of Plant Physiology and Plant Molecular Biology*, 47(1): 405–430, 1996. PMID: 15012295.

[209] C. E. Jeffree. Structure and ontogeny of plant cuticles. *Plant cuticles: an integrated functional approach. BIOS Scientific Publishers Ltd.: Oxford, UK*, pages 33–82, 1996.

[210] R. Jetter and S. Schffer. Chemical composition of the prunus laurocerasus leaf surface. dynamic changes of the epicuticular wax film during leaf development. *Plant Physiology*, 126(4):1725–1737, 2001.

[211] A. Alfaro-Tapia, J. A. Verdugo, L. A. Astudillo, and C. C. Ramrez. Effect of epicuticular waxes of poplar hybrids on the aphid chaitophorus leucomelas (hemiptera: Aphididae). *Journal of Applied Entomology*, 131(7):486–492, 2007.

[212] J.-W. Park, H. Min, Y.-P. Kim, H. Kyong Shon, J. Kim, D. W. Moon, and T. G. Lee. Multivariate analysis of tof-sims data for biological applications. *Surface and Interface Analysis*, 41(8):694–703, 2009.

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

[214] D. J. Graham, M. S. Wagner, and D. G. Castner. Information from complexity: Challenges of tof-sims data interpretation. *Appl Surf Sci*, 252(19):6860–6868, May 2006.

[215] B. Tyler. Interpretation of tof-sims images: multivariate and univariate approaches to image de-noising, image segmentation and compound identification. *Appl Surf Sci*, 203:825–831, November 2003.

[216] D. J. Graham and D. G. Castner. Multivariate analysis of tof-sims data from multicomponent systems: The why, when, and how. *Biointerphases*, 7(1):49, 2012.

[217] S. G. Boxer, M. L. Kraft, and P. K. Weber. Advances in imaging secondary ion mass spectrometry for biological samples. *Annual Review of Biophysics*, 38(1): 53–74, 2009.

[218] S. Aoyagi, M. Hayama, U. Hasegawa, K. Sakai, M. Tozu, T. Hoshi, and M. Kudo. Estimation of protein adsorption on dialysis membrane by means of tof-sims imaging. *Journal of Membrane Science*, 236(1-2):91–99, 2004.

[219] M. Brulet, A. Seyer, A. Edelman, A. Brunelle, J. Fritsch, M. Ollero, and O. Laprévote. Lipid mapping of colonic mucosa by cluster tof-sims imaging and multivariate analysis in cftr knockout mice. *J Lipid Res*, 51(10):3034–3045, October 2010.

[220] M. C. Biesinger, D. J. Miller, R. R. Harbottle, F. Possmayer, N. S. McIntyre, and N. O. Petersen. Imaging lipid distributions in model monolayers by tof-sims with selectively deuterated components and principal components analysis. *Applied Surface Science*, 252(19):6957 – 6965, 2006.

[221] T. L. Colliver, C. L. Brummel, M. L. Pacholski, F. D. Swanek, A. G. Ewing, and N. Winograd. Atomic and molecular imaging at the single-cell level with tof-sims. *Analytical Chemistry*, 69(13):2225–2231, 1997.

[222] W. B. Dunn. Mass spectrometry in systems biology an introduction. *Methods Enzymol*, 500:15–35, 2011.

[223] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[224] M. R. Keenan and P. G. Kotula. Optimal scaling of tof-sims spectrum-images prior to multivariate statistical analysis. *Appl Surf Sci*, 231:240–244, June 2004.

[225] R. N. Cochran and F. H. Horne. Statistically weighted principal component analysis of rapid scanning wavelength kinetics experiments. *Analytical Chemistry*, 49 (6):846–853, may 1977.

[226] A. Henderson, J. S. Fletcher, and J. C. Vickerman. A comparison of pca and maf for tof-sims image interpretation. *Surf Interface Anal*, 41(8):666–674, July 2009.

[227] Q. Teng. *Structural Biology.* Springer US, Boston, MA, 2013. ISBN 978-1-4614-3963-9.

[228] K. D. Bemis, A. Harry, L. S. Eberlin, C. Ferreira, et al. Cardinal: an r package for statistical analysis of mass spectrometry-based imaging experiments. *Bioinformatics*, 31(14):2418–2420, 2015.

[229] J. Cvačka and A. Svatoš. Matrix-assisted laser desorption/ionization analysis of lipids and high molecular weight hydrocarbons with lithium 2, 5-dihydroxybenzoate matrix. *Rapid Commun Mass Spectrom*, 17(19):2203–2207, 2003.

[230] V. Vrkoslav, A. Muck, J. Cvacka, and A. Svatos. Maldi imaging of neutral cuticular lipids in insects and plants. *Journal of the American Society for Mass Spectrometry*, 21(2):220–31, feb 2010.

[231] H. J. Ensikat, M. Boese, W. Mader, W. Barthlott, and K. Koch. Crystallinity of plant epicuticular waxes: electron and x-ray diffraction studies. *Chemistry and physics of lipids*, 144(1):45–59, oct 2006.

[232] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6): 610–621, Nov 1973.

[233] R. Jetter, S. Schaffer, and M. Riederer. Leaf cuticular waxes are arranged in chemically and mechanically distinct layers: evidence from prunus laurocerasus l. *Plant, Cell and Environment*, 23(6):619–628, jun 2000.

Appendix A

Supplementary information for Chapter 4

A.1 Experimental details

Insect material

Drosophila melanogaster (Meigen, 1830) flies (Diptera, Drosophilidae) were obtained from a colony (strain Canton S) maintained in the Max Planck Institute for Chemical Ecology, Jena, Germany. All flies (males and females) used for MSI experiments were 6-day-old virgins. Each fly was bred separately in an Eppendorf tube for 4 days on a normal diet and for the final 2 days on sugar solution only. Insect samples were immobilized by freezing at 18°C before they were prepared. Sample preparation consisted of slowly warming flies (sealed in plastic Eppendorf tubes) up to room temperature, then fixing them on a special MALDI target.

Chemicals used

2,5-Dihydroxybenzoic acid (DHB) for MALDI-MS with purity above 98% was purchased from Sigma-Aldrich (Germany). Lithium 2,5-dihydroxybenzoate (LiDHB) matrix was synthesized as described by Cvačka *et al.* [229]. HPLC grade solvents (acetone, chloroform, and dichloromethane) and high-purity MS grade solvent (methanol) were also supplied by Sigma-Aldrich. Analytic grade poly(ethylene glycol) oligomers (PEG with average masses of 200, 300, 600, and 1000 Da) for calibration of the mass spectrometer were purchased from Sigma-Aldrich.

Sample preparation

Samples of *D. melanogaster* flies were fixed on dedicated aluminum MALDI plates [2] by epoxy glue (Hardman, www.royaladhesives.com, Belleville, NJ, USA). The target, with flies on it, was then placed into a desiccator with phosphorus pentoxide (Sicapent) at ambient temperature and pressure for 6 hours. Later, the flies were completely dried and prepared to be subjected to imaging experiments. LiDHB matrix solutions were prepared in acetone:dichloromethane (9:1, v/v) at a concentration of 30 mg/mL. The matrix was sprayed on the samples by a commercial airbrush (Harder&Steenbeck, www.airbrushuniverse.com, Norderstedt, Germany) with a 0.15 mm diameter nozzle from a distance of 160 mm. For one sample, 2 mL of LiDHB matrix solution was used to form approximately 100 layers. The waiting time between two consecutive sprays was 4 seconds.

A.2 Dataset 1 - *m/z* 327.15 ([M+K]$^+$)

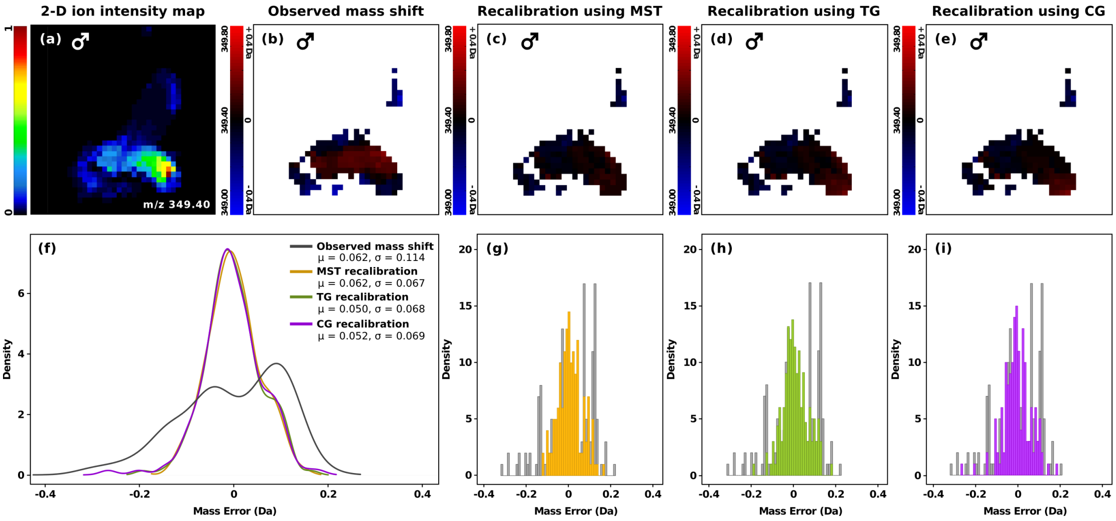


Figure A.1: Demonstration of mass shift and its correction using data obtained during LDI-MSI of six-day-old virgin male *D. melanogaster* fly in lateral position **(a)** 2-D ion intensity map of cVA *m/z* 349.40 ([M+K]$^+$), signal intensity in voxel view with rainbow-color scale, in a mass window size of ± 0.4 Da. **(b)** 2-D pseudo color plot of the observed mass shift with PlusMinus 1 color scale (*red* represents a positive mass shift and *blue* represents a negative mass shift, the pixels in *black* signify no mass shift, whereas the pixels in *white* correspond to no signal from that specific pixel in the selected mass range). Mass shift correction by the recalibration method **(c)** using MST ordering, **(d)** using TG ordering, **(e)** using CG ordering. **(f)** Pseudo density plot of the mass shift observed in preprocessed data (*gray*) and its correction (MST ordering - *yellow*, TG ordering - *green*, CG - *violet*). **(g - i)** Mass error distribution histograms of the preprocessed data before and after recalibration using MST, TG and CG ordering respectively (Observed mass error - *gray*, correction using MST ordering - *yellow*, correction using TG ordering - *green*, correction using CG ordering - *violet*).

Figure A.1 shows the performance of the recalibration method for the distribution of cVA, in a six-day-old virgin male *D. melanogaster* fly. Figure A.1(a) shows a 2-D ion intensity map representing the distribution of cVA at *m/z* 349.40 [(M+K)$^+$], which is localized with the highest signal intensities mainly in the abdomen region, and shows the abundance distribution of the specific ion. The observed mass shift without recalibration for *m/z* 349.40 in a mass window of ± 0.4 Da is shown in Figure A.1(b). The red pixels in the central part of the abdomen, correspond to the maximum observed mass shift of about 0.25 Da (represented by few pixels). The rest of the abdomen region shows a positive mass shift of about 0.1 - 0.2 Da. A part of the lower region of the abdomen, the short stub leg and few pixels in the wing region of the fly show a negative mass shift indicated as blue pixels of about 0.35 Da. In general, the strongest signals of cVA on intact virgin male fruit fly are observed in the genital area [184]. The compound could be spread from this region by diffusion in cuticular lipids layer to the abdominal tip and middle abdominal region [2]. This is evident from Figure A.1(a). In addition, the male fly can spread the cVA to the whole body using its legs. Due to this, a weak signal of cVA is observed also on the wings. The black pixels in the last abdominal tergites represent no mass shift. The effect of applying the recalibration method on the distribution and magnitude of the observed mass shift is shown in the form of 2-D pseudo color plots of the observed and corrected mass shift in Figures A.1(c-e). Figure A.1(c) represents the 2-D pseudo color plot of the mass shift correction using the MST approach of ordering the peaklists in the recalibration method. After recalibration, the positive mass shift in the central abdominal region is strongly reduced and most of the pixels are now black, representing almost no mass shift. Figure A.1(d) represents the 2-D pseudo color plot obtained after recalibration, using the TG ordering of peaklists. Similar to the recalibration performed using MST ordering, the 2-D pseudo color plot obtained using TG ordering greatly reduces the positive mass shift in the central abdomen region. However, a slight positive mass shift of about 0.2 Da is still observed at the tip of the abdomen. This is also observed for the correction performed using CG ordering, as shown in Figure A.1(e). There is a significant difference observed in the 2-D pseudo color plots generated before and after recalibration and the range of mass error distribution is seen to be reduced. This is evident from the pseudo-density plot shown in Figure A.1(f), which compares the performance of the three peaklist ordering approaches. The gray line represents the initial observed mass error distribution for the preprocessed data, yellow line represents the recalibration performed using MST ordering, green line represents the recalibration performed using TG ordering and violet line represents recalibration using CG ordering. Figures A.1(g-i) represent the mass error distribution histogram for the preprocessed data before and after recalibration, using MST, TG and CG ordering, respectively. The histograms show that, prior to recalibration, many of the mass peaks are shifted by about 0.25 Da in the positive direction and by about 0.35 Da in the negative direction, relative to the theoretical mass. In contrast, post-recalibration it is observed that the mass error distribution histograms for MST, TG and CG ordering, are nearly similar.

A.3 Dataset 2 - *m/z* 193.15 [(M+K)$^+$] and *m/z* 327.15 ([M+K]$^+$)

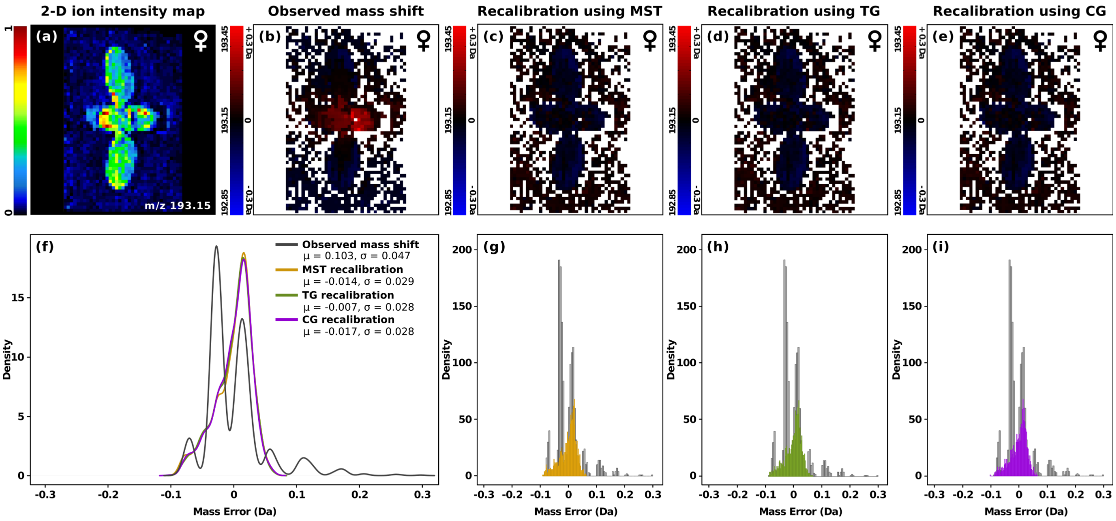


Figure A.2: Demonstration of mass shift and its correction using data obtained during MALDI-MSI of one *D. melanogaster* virgin female fly in dorsal position (a) 2-D ion intensity map of DHB, the matrix ion, at *m/z* 193.15 ([M+K]$^+$), signal intensity in voxel view with rainbow-color scale, in a mass window size of ± 0.3 Da. **(b)** 2-D pseudo color plot of the observed mass shift with PlusMinus 1 color scale (*red* represents a positive mass shift and *blue* represents a negative mass shift, the pixels in *black* signify no mass shift, whereas the pixels in *white* correspond to no signal from that specific pixel in the selected mass range). Mass shift correction by the recalibration method **(c)** using MST ordering, **(d)** using TG ordering, **(e)** using CG ordering. **(f)** Pseudo density plot of the mass shift observed in preprocessed data (*gray*) and its correction (MST ordering - *yellow*, TG ordering - *green*, CG - *violet*). **(g - i)** Mass error distribution histograms of the preprocessed data before and after recalibration using MST, TG and CG ordering respectively (Observed mass error - *gray*, correction using MST ordering - *yellow*, correction using TG ordering - *green*, correction using CG ordering - *violet*).

Figure A.2(a) represents the distribution of the matrix ion on an embedded female virgin fruit fly covered with LiDHB matrix. As can be seen, the potassium adduct of DHB matrix ion is evenly distributed all over the fly body with comparable intensity signals on all the regions of the fly. Figure A.2(b) shows a high positive mass shift in the abdomen region of the fly, a maximum of 0.3 Da and a negative mass shift of about 0.1 Da, on the edges of both the wings. The small islands of mostly black and some red and blue pixels, in Figure A.2(b), outside the fly body, represent areas of deposited matrix and epoxy glue, which may have been unevenly applied on the MALDI target plate. The white pixels represent no signal from the specific position in the selected mass window. These white pixels are visible in the pseudo color plots (see Figures A.2(b - e)), because, the raw imaging dataset had multiple low intensity regions outside the fly body, which got eliminated when we applied the data preprocessing steps. After applying our recalibration method, we observed that the high positive mass shift is strongly reduced using all the three ordering approaches. However, we can still observe a negative mass shift of about 0.1 Da around the edge of both the wings. This can be seen in Figure A.2(c) and A.2(d). All the three ordering approaches perform comparable recalibration for the mass shift observed for matrix ion distribution.

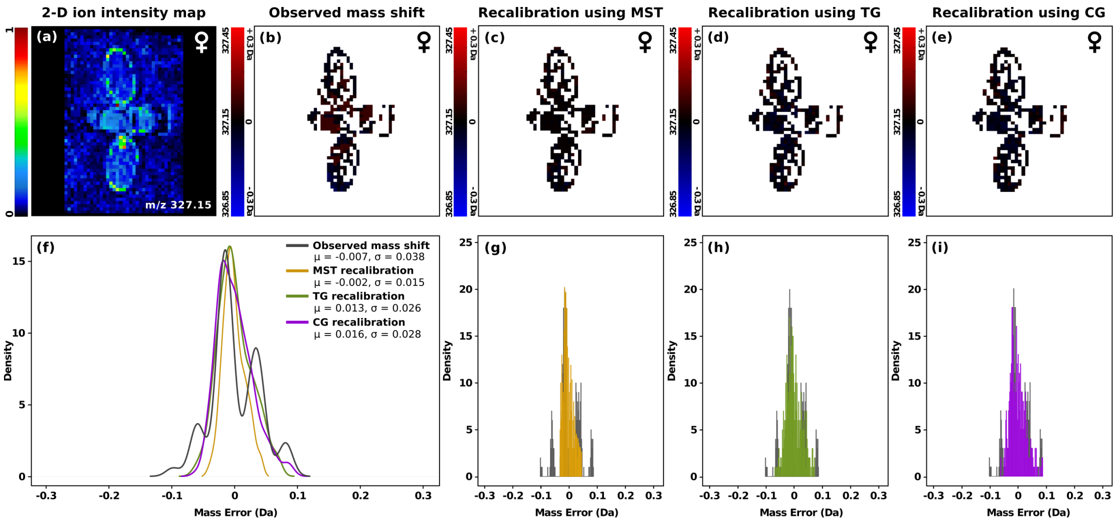


Figure A.3: Demonstration of mass shift and its correction using data obtained during MALDI-MSI of one *D. melanogaster* virgin female fly in dorsal position **(a)** 2-D ion intensity map of tricosadiene *m/z* 327.15 ($[M+Li]^+$), signal intensity in voxel view with rainbow-color scale, in a mass window size of ± 0.3 Da. **(b)** 2-D pseudo color plot of the observed mass shift with PlusMinus 1 color scale (*red* represents a positive mass shift and *blue* represents a negative mass shift, the pixels in *black* signify no mass shift, whereas the pixels in *white* correspond to no signal from that specific pixel in the selected mass range). Mass shift correction by the recalibration method **(c)** using MST ordering, **(d)** using TG ordering, **(e)** using CG ordering. **(f)** Pseudo density plot of the mass shift observed in preprocessed data (*gray*) and its correction (MST ordering - *yellow*, TG ordering - *green*, CG - *violet*). **(g - i)** Mass error distribution histograms of the preprocessed data before and after recalibration using MST, TG and CG ordering respectively (Observed mass error - *gray*, correction using MST ordering - *yellow*, correction using TG ordering - *green*, correction using CG ordering - *violet*).

Figure A.3 shows the distribution of tricosadiene at *m/z* 327.15 [$(M+Li)^+$] and recalibration of the observed mass shift. Figure A.3(b) shows a positive mass shift of about 0.15 Da in the abdominal region and some parts of the base of the wing. A negative mass shift of about 0.1 Da is observed at the edge of the lower wing. The visible signals outside the body of the fly in Figure A.3(a) probably originate from epoxy glue or rather hardener, because of poor polymerization in this specific experiment. Most of these signals are not visible in Figures A.3(b) - A.3(e), since these 2-D pseudo color plots are generated after applying preprocessing steps to the raw imaging dataset. After peak picking, the low intensity signals outside the fly body are automatically removed. The only residue of these signals originating from epoxy glue can be observed in the area around the edge of the cell imprinted into metal MALDI plate (see Figures A.3(b) - A.3(e)). All the three ordering approaches by our recalibration method were able to strongly reduce the mass shift, with MST ordering performing slightly better than TG and CG ordering.

A.4 Dataset 3 - *m/z* 327.15 ($[M+K]^+$) and *m/z* 193.15 ($[M+K]^+$)

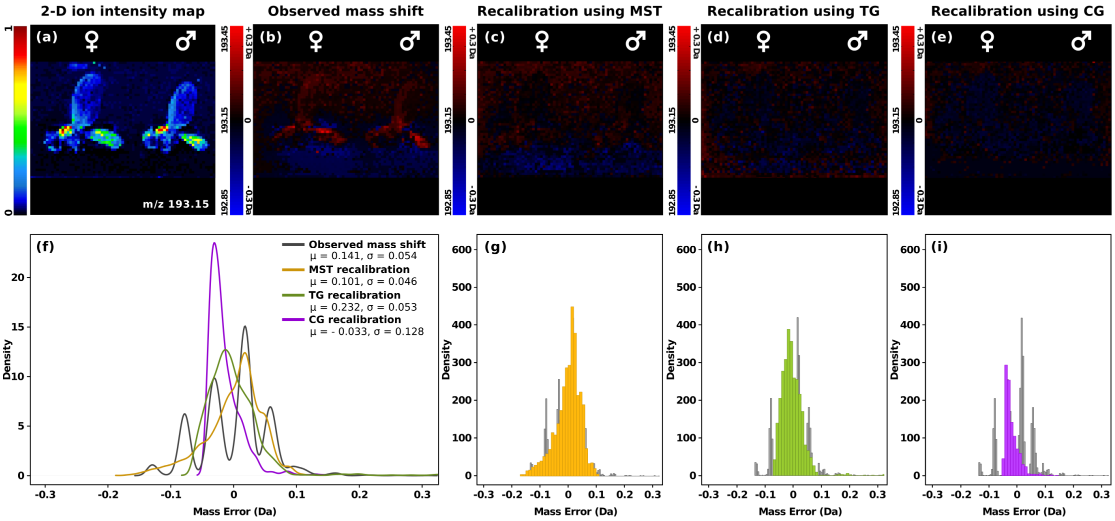


Figure A.4: Demonstration of mass shift and its correction using data obtained during MALDI-MSI of six-day-old virgin two pairs of (female/male) *D. melanogaster* flies imaged in lateral position (a) 2-D ion intensity map of DHB, the matrix ion, at *m/z* 193.15 ($[M+K]^+$), signal intensity in voxel view with rainbow-color scale, in a mass window size of ± 0.3 Da. **(b)** 2-D pseudo color plot of the observed mass shift with PlusMinus 1 color scale (*red* represents a positive mass shift and *blue* represents a negative mass shift, the pixels in *black* signify no mass shift, whereas the pixels in *white* correspond to no signal from that specific pixel in the selected mass range). Mass shift correction by the recalibration method **(c)** using MST ordering, **(d)** using TG ordering, **(e)** using CG ordering. **(f)** Pseudo density plot of the mass shift observed in preprocessed data (*gray*) and its correction (MST ordering - *yellow*, TG ordering - *green*, CG - *violet*). **(g - i)** Mass error distribution histograms of the preprocessed data before and after recalibration using MST, TG and CG ordering respectively (Observed mass error - *gray*, correction using MST ordering - *yellow*, correction using TG ordering - *green*, correction using CG ordering - *violet*).

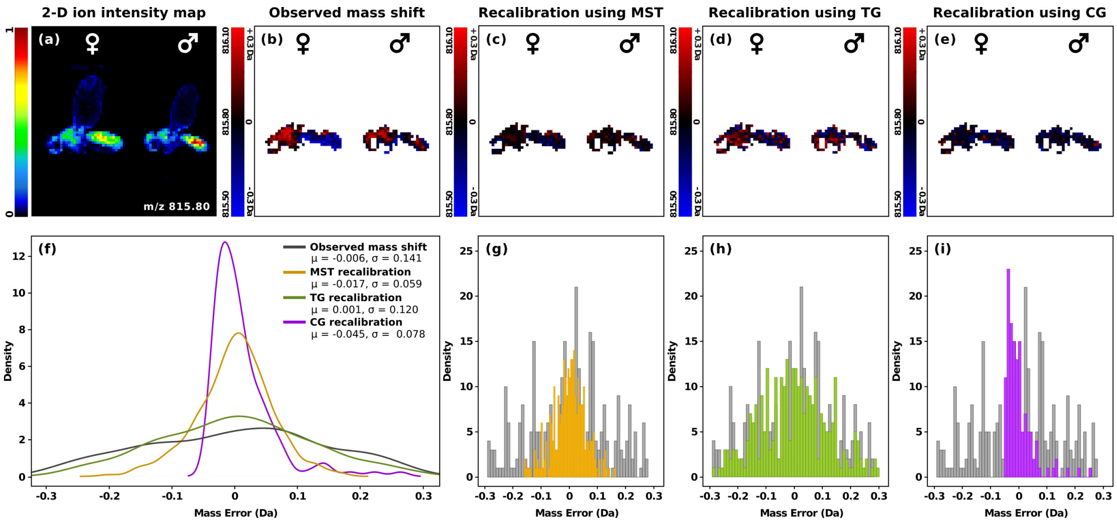


Figure A.5: Demonstration of mass shift and its correction using data obtained during MALDI-MSI of six-day-old virgin two pairs of (female/male) *D. melanogaster* flies in lateral position (a) 2-D ion intensity map of triacylglycerol *m/z* 815.80 ($[M+K]^+$), signal intensity in voxel view with rainbow-color scale, in a mass window size of ± 0.3 Da. **(b)** 2-D pseudo color plot of the observed mass shift with PlusMinus 1 color scale (*red* represents a positive mass shift and *blue* represents a negative mass shift, the pixels in *black* signify no mass shift, whereas the pixels in *white* correspond to no signal from that specific pixel in the selected mass range). Mass shift correction by the recalibration method **(c)** using MST ordering, **(d)** using TG ordering, **(e)** using CG ordering. **(f)** Pseudo density plot of the mass shift observed in preprocessed data (*gray*) and its correction (MST ordering - *yellow*, TG ordering - *green*, CG - *violet*). **(g - i)** Mass error distribution histograms of the preprocessed data before and after recalibration using MST, TG and CG ordering respectively (Observed mass error - *gray*, correction using MST ordering - *yellow*, correction using TG ordering - *green*, correction using CG ordering - *violet*).

Appendix B

Supplementary information for Chapter 6

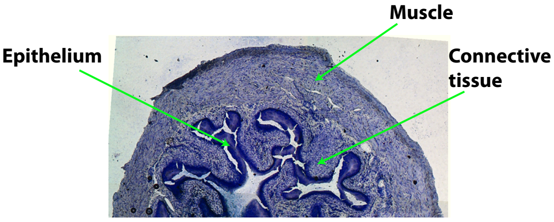


Figure B.1: Optical image of the mouse urinary bladder tissue used for MSI. The labeled optical image represents different regions of the measured mouse urinary bladder tissue section. This optical image is available for download along with the high-resolution MSI dataset (dataset identifier: PXD001283) deposited to the ProteomeXchange Consortium (`http://proteomecentral.proteomexchange.org`) via the PRIDE repository (`http://www.ebi.ac.uk/pride`) [199]

.

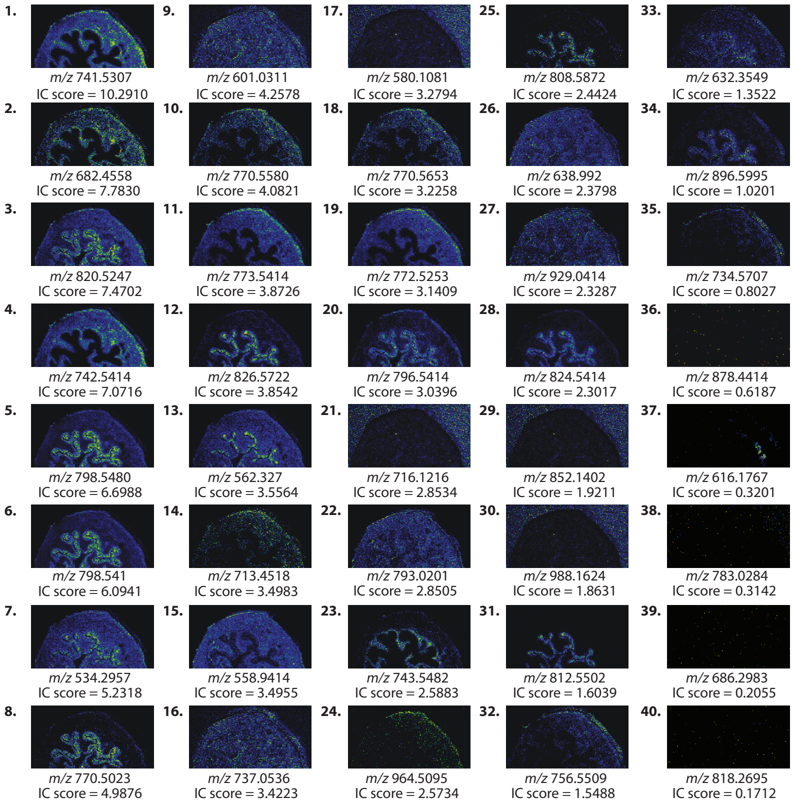


Figure B.2: Ion intensity maps corresponding to the 40 selected mass values from the tissue section of the mouse urinary bladder The ion maps shown from 1 to 40 are ranked based on their IC scores, where rank=1 represents the highest IC score and rank=40 represents the lowest IC score.

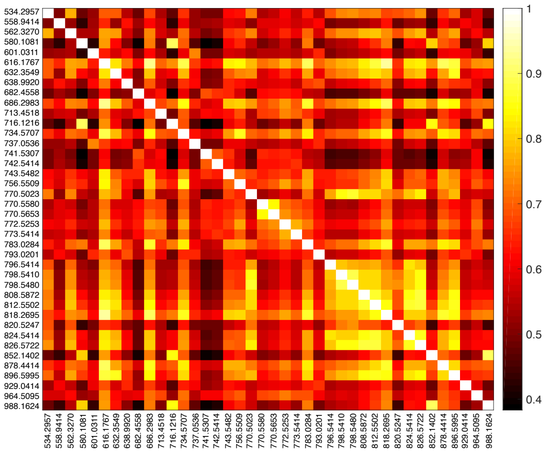


Figure B.3: Heat map of the calculated similarity matrix for dataset 1. Each cell in the heat map represents the amount of structural similarity between a pair of ion intensity maps corresponding to two mass values. This is calculated using the structural similarity (SSIM) index which lies between 0 (least similar) and 1 (identical). The colors in the heat map represent the amount of similarity: The lighter the color, the more similar the two images are.

Table B.1: Spatial similarity-based grouping of the ion intensity maps corresponding to the 40 selected mass values from dataset 1.

m/z	IC score	m/z	IC score	m/z	IC score	m/z	IC score	m/z	IC score
Group 1		Group 2		Group 3		Group 4		Group 5	
580.1081	3.2794	743.5482	2.5883	820.5247	7.4702	741.5307	10.291	601.0311	4.2578
716.1216	2.8534	756.5509	1.5488	798.5480	6.6988	682.4558	7.7830	558.9414	3.4955
852.1402	1.9211	734.5707	0.8027	798.5410	6.0941	742.5414	7.0716	737.0536	3.4223
988.1624	1.8631	878.4414	0.6187	534.2957	5.2318	770.5580	4.0821	793.0201	2.8505
		616.1767	0.3201	770.5023	4.9876	773.5414	3.8726	638.9920	2.3798
		783.0284	0.3142	826.5722	3.8542	713.4518	3.4983	929.0414	2.3287
		686.2983	0.2055	562.3270	3.5564	770.5653	3.2258		
		818.2695	0.1712	796.5414	3.0396	772.5253	3.1409		
				808.5872	2.4424	964.5095	2.5734		
				824.5414	2.3017				
				812.5502	1.6039				
				632.3549	1.3522				
				896.5995	1.0201				

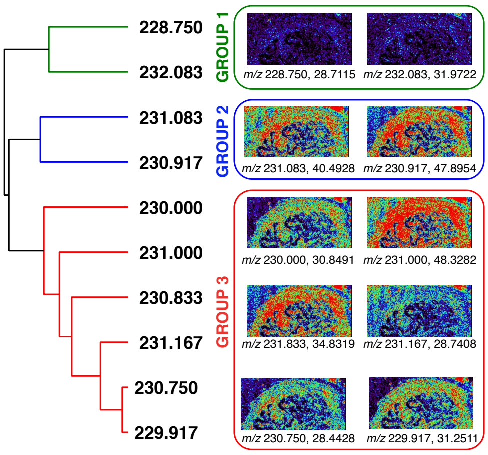


Figure B.4: Dendrogram representing the result of hierarchical clustering performed for the top 10 masses in dataset 2. Each nested group formed within the dendrogram contains ion maps that exhibit similar spatial distribution on the imaged tissue section. The top 10 masses have been selected based on their IC scores.

Table B.2: List of mass values corresponding to dataset 2 ranked in descending order of their IC scores.

Rank	*m/z* value	IC score
1	231	48.3282
2	230.917	47.8954
3	231.083	40.4928
4	230.833	34.8319
5	232.083	31.9722
6	229.917	31.2511
7	230	30.8491
8	231.167	28.7408
9	228.75	28.7115
10	230.75	28.4428
11	232.833	27.9242
12	229.833	27.4424
13	229.667	26.0017
14	231.833	25.4084
15	230.083	24.7494
16	230.25	22.6309
17	240	22.0618
18	230.583	20.8312
19	232.167	20.4686
20	239.917	18.6668
21	232.75	18.5229
22	246	17.7613
23	231.75	16.6236
24	245.917	16.5166
25	229.75	16.0282
26	244.083	15.8267
27	246.083	15.4536
28	244.167	14.4451
29	225.083	13.8926
30	226.167	13.7967
31	228.667	13.6997
32	226.083	13.6300
33	244	13.5568
34	243.083	13.3245
35	229.583	13.2473
36	231.333	12.5903
37	243	12.4736
38	225.167	12.4150
39	243.167	12.2519
40	225	12.1530
41	239.833	12.1126
42	246.833	12.1040
43	246.917	12.0171
44	249.917	11.8050
45	228.083	11.6692
46	242.083	11.6672
47	249	11.6121
48	244.25	11.3859
49	249.083	11.2824
50	237.083	11.1480
51	237	11.0844
52	242.167	11.0768
53	244.75	11.0122
54	228	10.9945
55	247	10.9673
56	226.25	10.9289
57	248.917	10.8955
58	246.75	10.7289
59	242.917	10.7254
60	232.25	10.3628
61	234	10.3603
62	245.833	10.2719
63	228.167	10.1542
64	226	10.0128
65	249.167	10.0042
66	246.25	9.8790
67	229	9.8376
68	243.25	9.8184
69	230.667	9.8017
70	245.083	9.6669
71	238.917	9.5962
72	234.083	9.4770
73	245.167	9.3922
74	235.083	9.3806
75	228.917	9.3053
76	225.25	9.2478
77	227.167	9.2326
78	242	9.1806
79	235	9.1518
80	247.083	9.1215
81	245	9.0392
82	248.667	8.9392
83	237.167	8.8645
84	226.917	8.8301
85	248.833	8.7584
86	227.25	8.6484
87	246.5	8.6413
88	244.333	8.6276
89	243.917	8.5751
90	248	8.5448
91	246.333	8.4772
92	236.917	8.4262
93	238.833	8.3661
94	242.25	8.3242
95	248.083	8.3216
96	233.917	8.2555
97	234.167	8.2252
98	226.333	8.1883
99	246.667	8.1636
100	245.25	8.1328
101	235.167	8.1302
102	227.917	8.0797
103	246.417	8.0797
104	240.417	8.0196
105	244.667	7.9156
106	249.25	7.8798
107	232.667	7.8373
108	238.083	7.8328
109	228.25	7.7884
110	248.167	7.6848
111	249.833	7.6419
112	226.833	7.4864
113	244.917	7.4485
114	230.5	7.3697
115	238.167	7.1531
116	247.167	7.1460
117	230.333	7.1107
118	238	7.0758
119	230.167	7.0470
120	229.083	6.9703
121	233	6.9395
122	244.417	6.9006
123	234.917	6.8728
124	231.667	6.7597
125	227.333	6.7506
126	236.083	6.6466
127	247.333	6.4239
128	241.917	6.4224
129	233.083	6.3860
130	225.917	6.3724
131	245.333	6.3512
132	234.25	6.3456
133	242.833	6.3310
134	247.917	6.3178
135	237.25	6.2901
136	225.333	6.2057
137	241.083	6.1446
138	240.167	6.1118
139	240.083	6.1022
140	239.75	5.9805
141	231.25	5.8901
142	236.167	5.8881
143	235.25	5.8674
144	243.333	5.8169
145	232.333	5.7911
146	232.917	5.7528
147	236	5.7164
148	226.417	5.7149
149	242.333	5.7018
150	233.833	5.6937
151	238.75	5.6669
152	240.75	5.6331
153	226.75	5.5821
154	245.75	5.5775
155	241	5.5371
156	248.25	5.5068
157	241.167	5.4735
158	244.583	5.4402
159	238.25	5.3765
160	241.417	5.2755
161	244.5	5.2256
162	229.5	5.1786
163	228.833	5.0741
164	228.333	5.0104
165	239.167	4.9796
166	228.583	4.9579
167	236.833	4.9529
168	231.417	4.9428
169	234.833	4.9397
170	236.25	4.9115
171	233.417	4.8488
172	240.5	4.8312
173	248.75	4.8251
174	227.833	4.7887
175	249.333	4.7862
176	248.583	4.7559
177	243.833	4.7554
178	247.417	4.7362
179	240.25	4.6978
180	230.417	4.6282
181	237.917	4.6206
182	233.167	4.5913
183	239.083	4.5494
184	229.417	4.5312
185	227.417	4.4999
186	235.917	4.4762
187	229.167	4.4716
188	231.917	4.4373
189	247.25	4.4352
190	247.833	4.4221
191	232	4.3913
192	234.333	4.3691
193	226.5	4.3176
194	249.75	4.2060
195	239.25	4.0439
196	226.667	3.9813
197	240.667	3.9404
198	237.333	3.9293
199	241.25	3.9217
200	235.333	3.9091
201	240.917	3.8636
202	244.833	3.7980
203	242.417	3.7712
204	225.833	3.7318
205	225.417	3.6833
206	233.75	3.6763
207	248.333	3.6702
208	228.417	3.6182
209	241.833	3.5833
210	242.75	3.5455
211	239.667	3.5323
212	238.667	3.5136
213	238.333	3.5116
214	246.583	3.5096
215	245.417	3.5086
216	232.583	3.3940
217	232.417	3.3874
218	239.5	3.3712
219	226.583	3.3667
220	236.333	3.3243
221	245.667	3.2834
222	239	3.2829
223	240.583	3.2551
224	233.5	3.2480
225	243.417	3.2470
226	228.5	3.2283
227	231.583	3.1859
228	237.833	3.1814
229	235.833	3.1637
230	249.5	3.1576
231	246.167	3.1324
232	227.75	3.1288
233	247.75	3.0854
234	234.75	2.9324
235	236.75	2.9046
236	227.5	2.8910
237	249.667	2.8733
238	243.75	2.8117
239	245.583	2.7920
240	239.583	2.7890
241	229.25	2.7880
242	227.083	2.7829
243	231.5	2.7425
244	233.25	2.7334
245	233.667	2.7284
246	241.5	2.6996
247	240.333	2.5991
248	248.417	2.5653
249	248.5	2.5592
250	234.417	2.5335
251	237.417	2.5208
252	227	2.5057
253	249.583	2.4512
254	238.417	2.4193
255	247.5	2.4153
256	235.417	2.3941
257	239.333	2.3784
258	233.583	2.3699
259	232.5	2.3693
260	242.5	2.3552
261	241.75	2.3406
262	238.583	2.2830
263	235.75	2.2487
264	225.75	2.2300
265	242.667	2.1775
266	227.667	2.1229
267	227.583	2.1199
268	236.667	2.0754
269	237.75	2.0497
270	234.667	2.0310
271	247.667	2.0275
272	236.417	1.9775
273	238.5	1.9194
274	243.667	1.9088
275	225.5	1.8937
276	240.833	1.8770
277	242.583	1.8669
278	243.5	1.8325
279	235.667	1.7714
280	247.583	1.7235
281	241.333	1.6831
282	229.333	1.6467
283	234.5	1.6255
284	241.583	1.5634
285	241.667	1.5422
286	235.5	1.5154
287	233.333	1.5129
288	245.5	1.4685
289	235.583	1.3765
290	234.583	1.3725
291	236.583	1.3478
292	237.5	1.3442
293	243.583	1.3407
294	225.667	1.3326
295	236.5	1.3321
296	249.417	1.2942
297	237.667	1.1998
298	239.417	1.1604
299	225.583	1.1033
300	237.583	0.9907

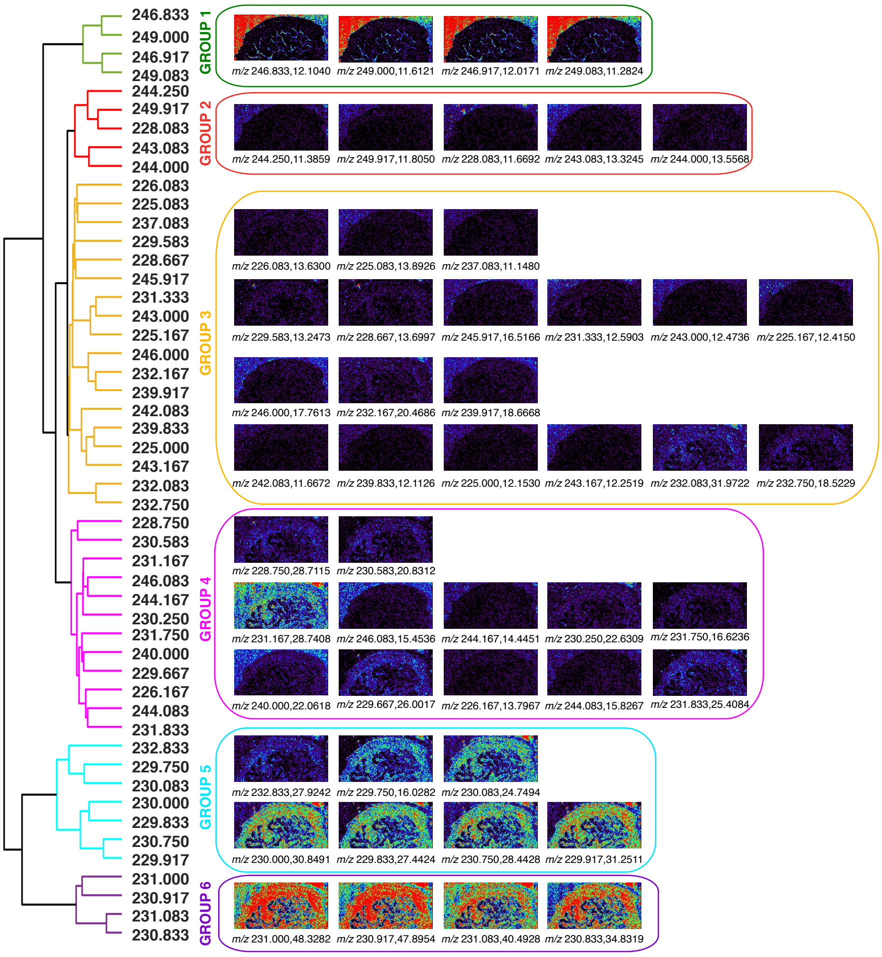


Figure B.5: Dendrogram representing the result of hierarchical clustering performed for the top 50 masses in dataset 2. Each nested group formed within the dendrogram contains ion maps that exhibit similar spatial distribution on the imaged tissue section. The top 50 masses have been selected based on their IC scores.

Appendix C

Supplementary information for Chapter 7

B.1 Experimental details

Plant material and growth conditions

Populus trichocarpa Torr. & Gray (clone 600-25) was cultivated in a growth chamber at 23°C day and 18°C night with 50% relative humidity, light intensity of 30% (Osram De Luxe 36W Natura) and a diurnal cycle of 12 h light and 12 h dark. Cuttings were kept at 12°C, and after 8 weeks when they reached 15-20 cm in length, plants were transferred to 2 L pots. Standard substrate (Klasmann-Deilmann GmbH, Postfach D-49744 Geeste, Germany, `www.klasmann-deilmann.com`) was used as soil, and beginning 4 weeks after transfer, plants were fertilized with 1% of 10:15:10 N:P:K supplement (Ferty, PLANTA Düngemittel GmbH, Regenstauf D-93128, Germany, `www.plantafert.com`) every second week. Plants were grown under controlled condition throughout the cultivation and not treated with any pesticides to prevent the potential chemical modification of the leafs surface. Plants used for all experiments were 4-5 months old.

Blotting of leaf cuticle

EWs were isolated using the cryo-adhesive method [233] with water as a transfer medium. *P. trichocarpa* leaves were rinsed briefly with distilled water to remove any contaminants. Leaf tissues were cut into rectangle-shaped pieces with dimensions slightly larger than the solid substrate support (a metallic MALDI plate for MALDI-TOF MS, and a p-type (100) orientation silicon wafer for TOF-SIMS). Each leaf cutting was placed onto the cleaned substrate holding 4 to 5 evenly distributed droplets (double distilled water 3 μl each). Afterwards, a glass slide was placed on top of the substrate/leaf stack and gently pressed, creating a sandwich. The leaf was homogeneously moistened with a very thin water film. Tweezers were used to dip the whole leaf in liquid nitrogen for 30 s. Water droplets served as a medium to transfer the wax layer from the leaf to the substrate, and the second glass slide enabled slight force to be put on the tissue without direct contact and allowed the EWs to be isolated, without disturbing the other cuticle layers. Metal plate or silicon wafer was detached from the plant tissue and was kept in a desiccator prior to analysis. This protocol results in a very flat EW transfer layer on conductive substrates, representing an ideal sample for TOF-SIMS imaging, to achieve best mass and lateral resolution.

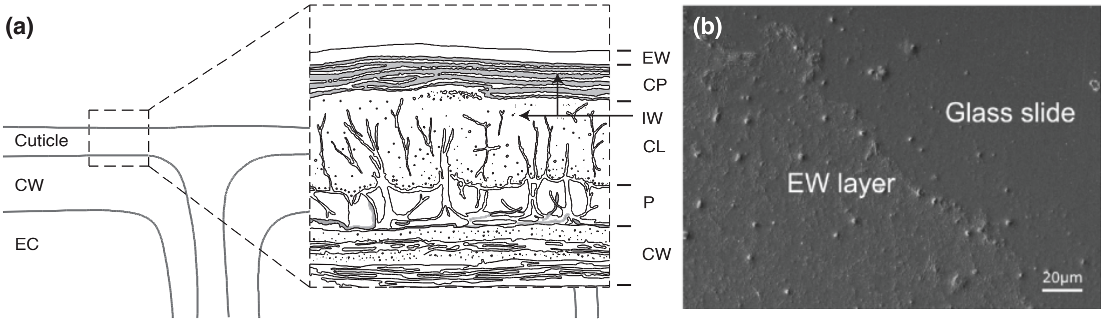


Figure C.1: Representation of the cuticle layers on the leaf surface (a) Cross-section of cuticle layers and upper epidermis (EW epicuticular waxes, IW- intracuticular waxes, CP cuticle proper, CL cuticle layer, P pectinaceous layer and middle lamella, CW cell wall, EC epidermal cell) **(b)** SEM micrograph of EWs from the surface of adaxial leaves of P. trichocarpa were isolated using cryo-adhesive tape embedding method.

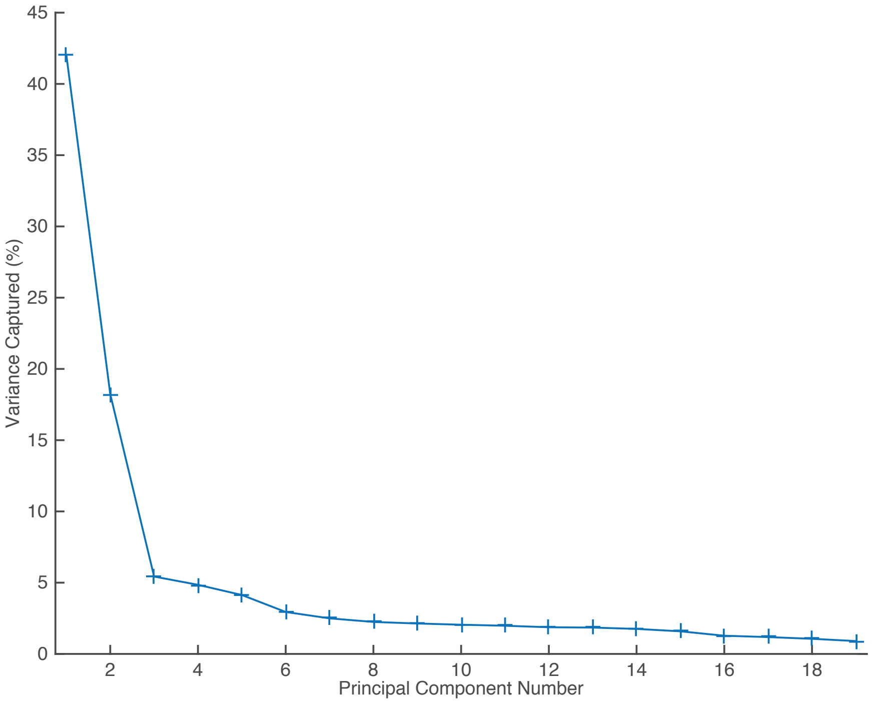
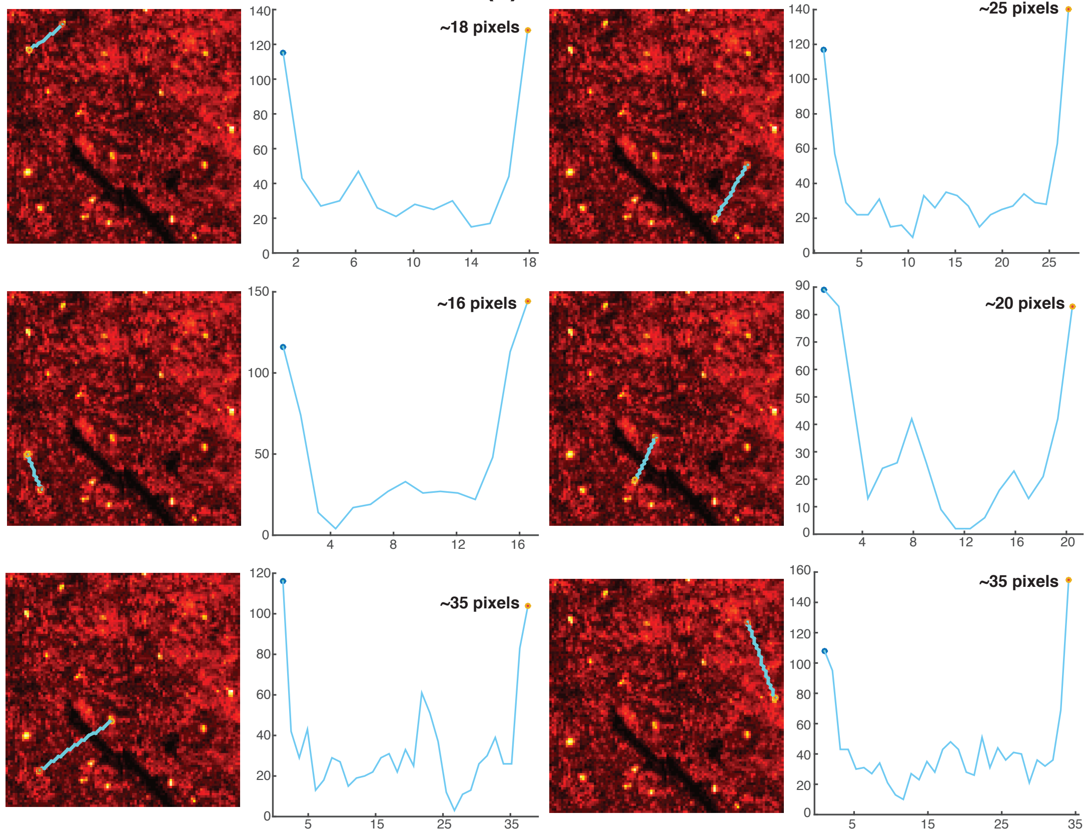


Figure C.2: Plot representing percentage of variance captured by each principal component Based on this plot we selected an optimal number of six principal components for the PCA of TOF-SIMS imaging data

(a) *m/z* 340.32



(b) *m/z* 424.24

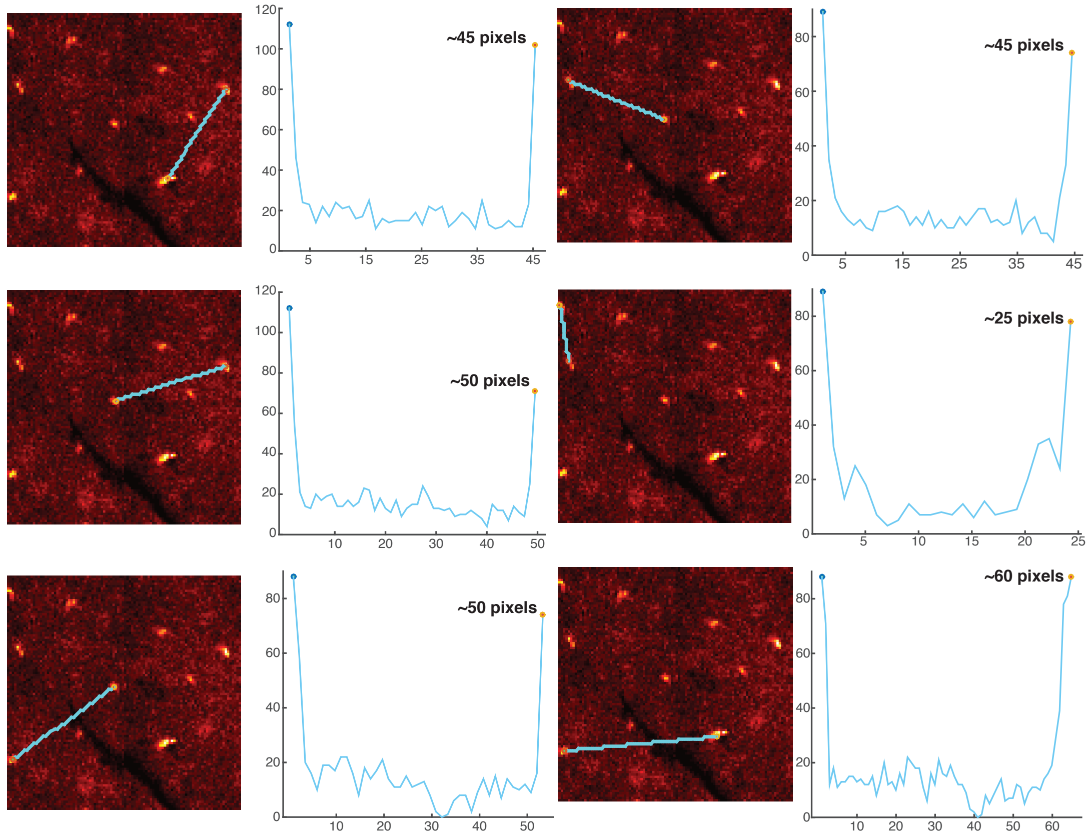


Figure C.3: Representation of distance measured among spatially localized crystals
An example of distance measured (in pixels) among different spatially localized crystals observed on pseudo-colored 2-D ion intensity maps of the leaf surface at **(a)** *m/z* 340.32 and **(b)** *m/z* 424.24. The pixels size is 1 μm.

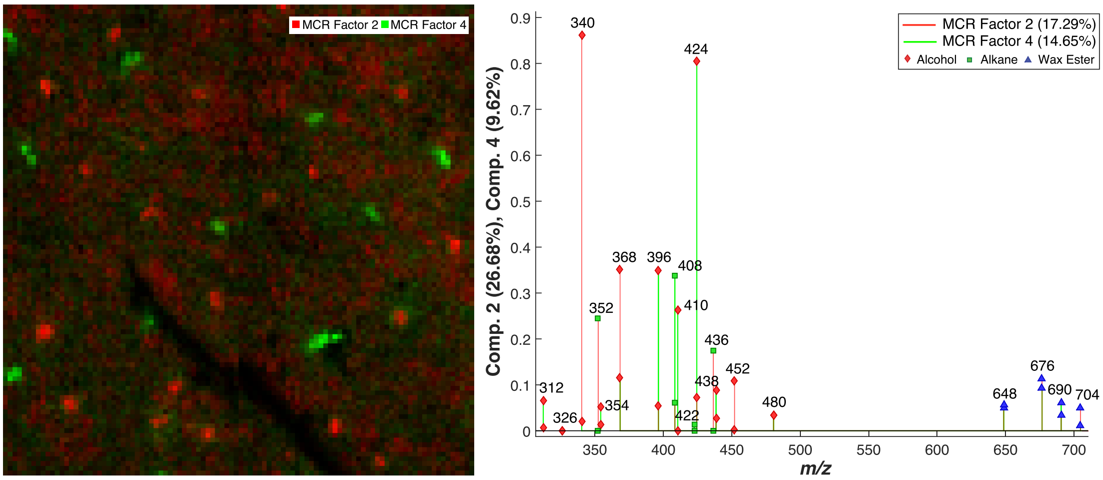


Figure C.4: Representation of an overlay of two crystal patterns on the leaf surface of *P. trichocarpa* Score plot (left) and loading plot (right) representing the spatial location of crystals on the leaf surface and the compounds responsible for their formation, by overlaying factor 2 and factor 4 obtained after MCR analysis. The slice for factor 2 is represented in red and the slice for factor 4 is represented in green.

Table C.1: Comparison of masses detected by MALDI-TOF MS with their respective theoretical masses

$[M+{}^7Li^+]$ monoisotopic	$[M+{}^7Li^+]$ detected	Mass difference [mDa]	Assigned compound
361.4022	361.4570	54.85	$C_{24}H_{50}O$
389.4335	389.4753	41.88	$C_{26}H_{54}O$
415.4855	415.5356	50.17	$C_{29}H_{60}O$
417.4648	417.5088	44.08	$C_{28}H_{58}O$
429.5012	429.5090	7.85	$C_{30}H_{62}O$
443.5168	443.5261	9.37	$C_{31}H_{64}O$
445.4961	445.5371	41.12	$C_{30}H_{62}O$
459.5117	459.5370	25.3	$C_{31}H_{64}O$
473.5274	473.5636	36.25	$C_{32}H_{66}O$
599.6318	599.6818	50.03	$C_{40}H_{80}O_2$
627.6631	627.715	51.9	$C_{42}H_{84}O_2$
655.6944	655.7461	51.77	$C_{44}H_{88}O_2$
669.7101	669.7516	41.62	$C_{45}H_{90}O_2$
683.7257	683.7775	51.8	$C_{46}H_{92}O_2$
697.7414	697.7780	36.65	$C_{47}H_{94}O_2$
711.757	711.8121	55.17	$C_{48}H_{96}O_2$
725.7727	725.8095	36.85	$C_{49}H_{98}O_2$
739.7883	739.8455	57.2	$C_{50}H_{100}O_2$
767.8196	767.8746	55.07	$C_{52}H_{104}O_2$
795.8509	795.9095	58.6	$C_{54}H_{108}O_2$