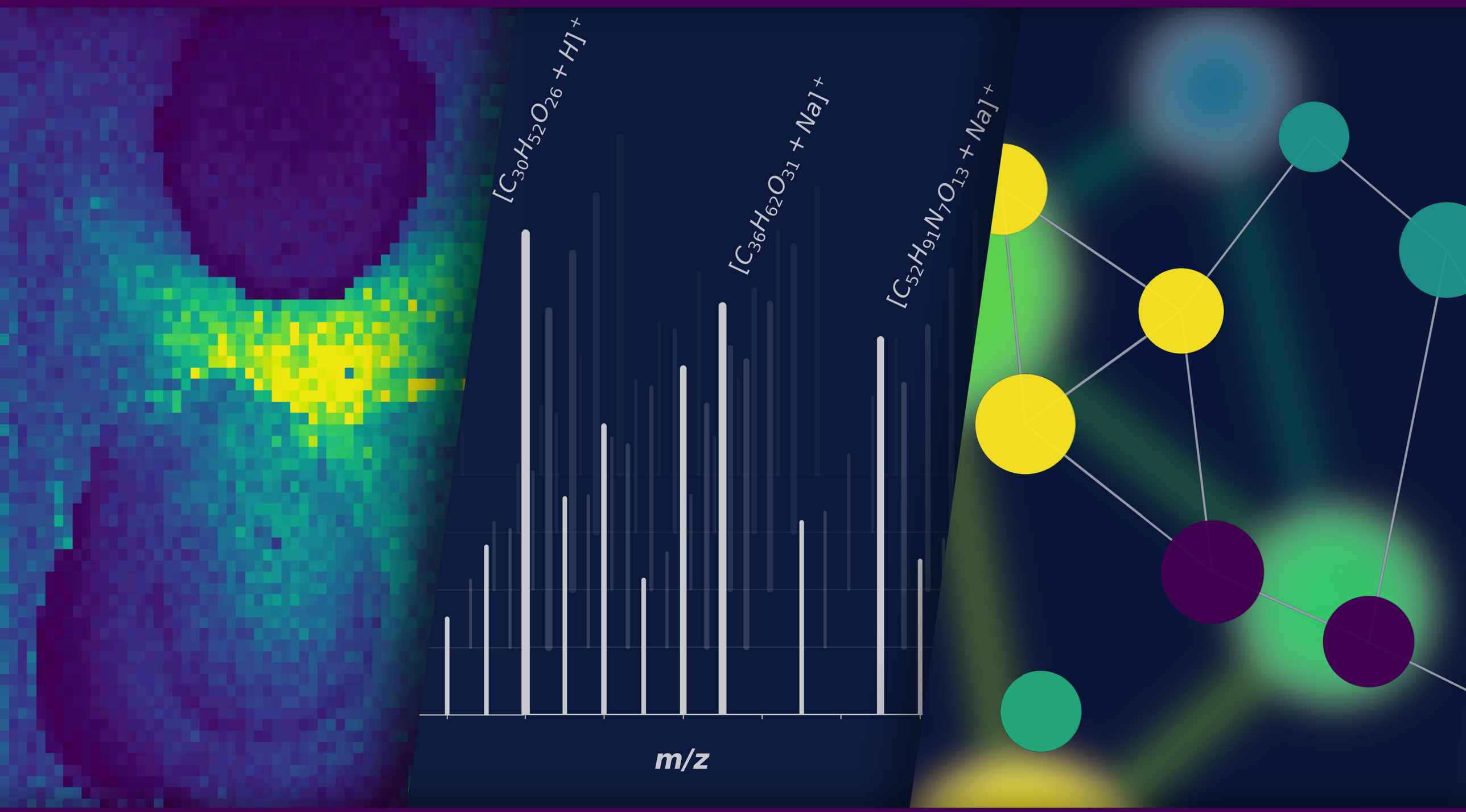
LIÈGE université
Sciences

DOCTORAL THESIS



Faculty of Sciences | Department of Chemistry
Academic Year 2025 - 2026

HIGH-RESOLUTION MASS SPECTROMETRY IMAGING: FROM MASS-SHIFT CORRECTION TO GRAPH-BASED MOLECULAR INTERPRETATION

RAPHAËL LA ROCCA

Mass Spectrometry Laboratory
MolSys Research Unit, ULiège
Under the supervision of Prof. Loïc Quinton
and co-supervision of Prof. Gauthier Eppe

MASS SPECTROMETRY LABORATORY

University of Liège

Faculty of Sciences

High-Resolution Mass Spectrometry Imaging: From Mass-Shift Correction to Graph-Based Molecular Interpretation

---

Supervisor

Prof. Loïc Quinton

Mass Spectrometry Laboratory
MolSys Research Unit, ULiège

Co-supervisor

Prof. Gauthier Eppe

Mass Spectrometry Laboratory
MolSys Research Unit, ULiège

Academic year 2025-2026

Dissertation submitted by

La Rocca Raphaël

for the fulfilment of the degree of
Doctor of Philosophy in Sciences

Jury members

Supervisor	Prof. Loïc QUINTON	University of Liège, Belgium
Co-supervisor	Prof. Gauthier EPPE	University of Liège, Belgium
President	Prof. Anne-Sophie DUWEZ	University of Liège, Belgium
Secretary	Prof. Sébastien RIGALI	University of Liège, Belgium
Member	Prof. Denis BAURAIN	University of Liège, Belgium
Member	Prof. Isabelle FOURNIER	University of Lille, France
Member	Prof. David TOUBOUL	Ecole Polytechnique of Palaiseau, France

Acknowledgements

First and foremost, I sincerely thank my promoter, Prof. Loïc Quinton, for the opportunity to conduct this thesis in his laboratory. I am truly grateful for the confidence he placed in me and for the freedom he gave me to pursue my own research ideas. I would also like to warmly thank my co-promoter, Prof. Gauthier Eppe, for his trust and support throughout this thesis. I thank them both for their feedback and for the opportunities they provided, which significantly contributed to my growth as a researcher.

I sincerely thank Prof. Edwin De Pauw, member of my thesis committee, for welcoming me into his research group and involving me in multiple projects. I deeply appreciate his continuous availability for discussions and meetings, as well as the time he dedicated to guiding me from the very beginning of my thesis. I gained a great deal from his insights, and I am grateful for his generous sharing of knowledge and experience.

I am also very grateful to Dr. Johann Far, member of my committee, for his valuable advice, both scientific and professional, and for sharing his expertise in mass spectrometry. I would also like to thank Dr. Cédric Malherbe for his thoughtful advice and support throughout this work. My sincere thanks also go to Dr. Virginie Bertrand for her help in the laboratory, particularly with microscopy imaging. I would also like to thank Dr. Gabriel Mazzucchelli for the different collaborations we shared and for the professional opportunities he offered during the course of this work.

I would also like to sincerely thank Prof. Theodore Alexandrov for his openness and continuous support, as well as the METASPACE team for their collaboration and for providing access to resources that contributed to key methods developed in this thesis.

I would like to thank the members of my thesis jury, namely Prof. Anne-Sophie Duwez, Prof. Sébastien Rigali, Prof. Denis Baurain, Prof. Isabelle Fournier, and Prof. David Touboul, for taking the time to carefully review this manuscript and for their valuable comments and suggestions, which contributed to improving this work. I also sincerely thank Prof. Anne-Sophie Duwez and Prof. Sébastien Rigali for accepting to serve respectively as President and Secretary of this thesis jury.

I would like to acknowledge the Mass Spectrometry Laboratory (MSLab) of the University of Liège for providing access to its extensive instrumentation and expertise in mass spectrometry. I gratefully acknowledge the financial support received from the Excellence of Science (EOS) program of the FWO/FNRS (Rhizoclip - EOS ID 30650620) and from the Interreg Euregio Meuse-Rhine (EMR) project EURLIPIDS (R8598), which made this work possible.

I would also like to thank all my co-authors and collaborators involved in the work presented in this thesis. Their contributions, discussions, and collaborative efforts were essential to the development of this research.

Of course, my thanks also go to all the current and former permanent staff, postdoctoral researchers, PhD students, and trainees of the MSLab, who shared part of this journey. I would like to give special thanks to Christopher Kune, Andrea McCann, Mathieu Tiquet, Johann Far, Thomas Tilmant, Pierre Burguet, and Cédric Delvaux. Thank you for the teamwork, the discussions, and the many moments shared in the lab. Without you, these years would not have been the same.

I would also like to sincerely thank Prof. Thomas Marichal for welcoming me back into academia after my time in industry, for including me in his team, and for encouraging me to bring this work to completion.

This PhD thesis would likely never have come to fruition without the support of my family and friends. I would like to thank my parents for giving me every opportunity to pursue what inspires me, and my brother Lionel, who helped me develop clean and reliable code for the deltamsi package. I also feel fortunate to have received the support and friendship of those around me. Special thanks to Anthony Cioppa for our many discussions and collaborations, and to Maxim Henry for the stimulating conversations we shared over the years.

Finally, I want to sincerely thank my partner, Kelly Cioppa, for her unwavering support, patience, and encouragement throughout this journey. Her invaluable help in reading and revising this thesis has been essential. Having you by my side, through both frustrations and successes, has given me the strength to persevere and grow, both as a researcher and as a person.

Table of Contents

Chapter 1: General Introduction to Mass Spectrometry Imaging

1.1 Basic description of a mass spectrometer 4

1.1.1 Mass spectrometry principle 4

1.1.2 Ion source 5

1.1.3 Matrix-Assisted Laser Desorption Ionization (MALDI) 6

1.1.4 Mass analyzers 10

1.1.4.1 Time of Flight (TOF) mass analyzers 12

1.1.4.2 Fourier transform ion cyclotron resonance analyzer (FT-ICR) 15

1.1.5 Mass spectrometry imaging 17

1.2 Importance of spatial and spectral resolution 19

1.2.1 Spectral resolution and mass accuracy 19

1.2.2 Spatial resolution and MALDI imaging 24

1.3 Main applications of MSI 28

1.3.1 Example of applications 28

1.3.2 Data treatments 30

1.4 Objectives and thesis organization 34

1.5 Recent advances and the specific gaps addressed by this thesis 39

1.5.1 Software, preprocessing, and annotation 39

1.5.2 Spectrum representation, foundation models and graph-based classification.... 40

1.5.3 Single-cell analysis, super-resolution and microscopy integration 41

Chapter 2: Development of a General MSI Workflow for Feature Prioritization

2.1 Introduction: Detecting Condition-Specific Ions 51

2.2 MSI processing pipeline 53

2.2.1 Mass Spectrometry Image Handling for Data Processing 56

2.2.2 Average Spectrum and Peak Detection 59

2.2.3 Generation of Ion Images for Each Peak 62

2.2.4 Scoring ion image structure 67

2.3 Exploiting Differential Signal Filtering (DSF) and Image Structure Filtering (ISF) Methods for Untargeted Mass Spectrometry Imaging of Bacterial Metabolites 72

2.4 Generalization of the workflow in the deltamsi package 75

2.4.1 Software for MSI data analysis 75

2.4.2 General description of deltamsi 79

2.5 Conclusion and transition to other chapters 81

Chapter 3: Mass Shift in Mass Spectrometry Imaging, Data Interpretation and Recalibration

3.1 Problem of mass shift in mass spectrometry imaging 89

3.1.1 Impact of mass shift 89

3.1.2 Post-acquisition and pre-acquisition approaches to correct mass shift 90

3.1.3 Context and scientific production 91

3.2 Adaptive pixel mass recalibration for mass spectrometry imaging based on locally endogenous biological signals 97

3.2.1 Abstract 97

3.2.2 Introduction 98

3.2.3 Experimental Section 101

3.2.3.1 Materials 101
3.2.3.2 Methods 102
3.2.4 Results and Discussion 110
3.2.4.1 Recalibration Impact on Data Analysis 110
3.2.4.2 Impact of Recalibration on the Number of Molecular Annotations 112
3.2.4.3 FDR and Tolerance Selection 114
3.2.4.4 Evaluation on a Large Set of Public MSI Data Sets 115
3.2.5 Conclusions 120
3.2.6 Associated Content and Acknowledgments 122

Chapter 4: Mass, Mass defect, and Mass pattern as Complementary Information for Mass Spectrum Data Processing

4.1 Kendrick Mass Defect as a filter for MS signal 129
4.1.1 Lack of spectral information in MSI annotation engine 129
4.1.2 Kendrick mass defect graphical representation 130
4.1.3 Rapid Visualization of Chemically Related Compounds Using Kendrick Mass Defect as a Filter in Mass Spectrometry Imaging 133
4.1.4 Limitations of MSKF 137
4.2 Expand Kendrick by adding a graph of mass differences 139
4.2.1 Material 145
4.2.2 MS spectrum as a graph of mass differences 146
4.2.3 Application to bacterial confrontation MSI 150
4.2.4 Discussion and Conclusion 158
4.2.5 Introducing machine learning for processing graphs of mass differences 160

Chapter 5: Relational Graph Convolutional Network for Robust Mass Spectrum Classification

5.1 Preface 167

5.2 Relational Graph Convolutional Network for Robust Mass Spectrum Classification 168

5.2.1 Abstract 168

5.2.2 Introduction 170

5.2.3 Experimental section 173

5.2.4 Method 175

5.2.4.1 General approach 175

5.2.4.2 Graph representation of the spectrum 175

5.2.4.3 Relational Graph Convolutional Network 177

5.2.4.4 Class activation maps for network interpretability 177

5.2.4.5 Code availability 178

5.2.5 Results and discussion 179

5.2.5.1 Robustness toward signal variation 179

5.2.5.2 R-GCN signal interpretability 188

5.2.5.3 Relationship to Chapter 4 199

5.2.6 Conclusion and perspectives 202

Chapter 6: Interpreting the Spatial Dimension of Mass Spectrometry Images

6.1 Utilizing spatial dimensions in mass spectrometry imaging 209

6.2 Application of a Deep Learning Model for Image Fusion between Microscope Image and Mass Spectrometry Imaging 213

6.2.1 Abstract 213

6.2.2 Introduction 214

6.2.3 Materials and Methods 219

6.2.3.1 Cell cultures 219

6.2.3.2 MSI acquisition 220

6.2.3.3 Microscope image acquisition 221

6.2.3.4 Image registration 223

6.2.3.5 Manual annotation of cells in the microscope images 224

6.2.3.6 U-Net model adapted for image fusion 225

6.2.3.6.1 Model input and training 225

6.2.3.6.2 Mass signal normalization 226

6.2.3.6.3 U-Net architecture 228

6.2.3.6.4 Tensor cropping 230

6.2.3.6.5 Tensor to mass signal 232

6.2.3.6.6 Generating Fused Images with CAM 233

6.2.4 Results and Discussion 235

6.2.4.1 Data preparation 235

6.2.4.2 MSI pre-processing 236

6.2.4.2.1 Microscope image pre-processing 238

6.2.4.2.2 Ion delocalization on model performances 239

6.2.4.3 Ion signal prediction 243

6.2.4.4 Comparing predicted ion location at the microscope spatial resolution ... 246

6.2.4.5 Microscopy image segmentation based on fused images 251

6.2.5 Conclusion and perspectives 256

Conclusion and perspectives

1 Conclusion 265

1.1 Overall significance 265

1.2 A practical pipeline to prioritize biology-relevant ions (Chapter 2) 265

1.3 Robust mass accuracy for reliable identification (Chapter 3) 266

1.4 Structuring high-resolution spectra beyond databases (Chapter 4 and Chapter 5) 267

1.5 Bridging MSI and microscopy to interpret space (Chapter 6) 267

1.6 How these pieces fit together 268

2 Perspectives 269

2.1 Smarter acquisition 269

2.2 Molecular formula computation 270

2.3 Noise and low-abundance signal 272

2.4 Graph based on molecular formula 274

2.5 Homogeneous family segmentation 277

2.6 Multimodal image fusion 279

2.7 Scalable, standardized evaluation 280

2.8 Closing statement 281

Supporting information 285

Associated papers 347

CHAPTER ONE

***

General Introduction to Mass Spectrometry Imaging

Chapter 1: Content

1.1 Basic description of a mass spectrometer 4

1.1.1 Mass spectrometry principle 4

1.1.2 Ion source 5

1.1.3 Matrix-Assisted Laser Desorption Ionization (MALDI) 6

1.1.4 Mass analyzers 10

1.1.4.1 Time of Flight (TOF) mass analyzers 12

1.1.4.2 Fourier transform ion cyclotron resonance analyzer (FT-ICR) 15

1.1.5 Mass spectrometry imaging 17

1.2 Importance of spatial and spectral resolution 19

1.2.1 Spectral resolution and mass accuracy 19

1.2.2 Spatial resolution and MALDI imaging 24

1.3 Main applications of MSI 28

1.3.1 Example of applications 28

1.3.2 Data treatments 30

1.4 Objectives and thesis organization 34

1.5 Recent advances and the specific gaps addressed by this thesis 39

1.5.1 Software, preprocessing, and annotation 39

1.5.2 Spectrum representation, foundation models and graph-based classification 40

1.5.3 Single-cell analysis, super-resolution and microscopy integration 41

Mass Spectrometry Imaging (MSI) has emerged over the past decades as a versatile technique that couples mass spectrometry with spatially resolved analysis. By acquiring a full mass spectrum at each position on a surface, MSI produces molecular maps of materials, biomaterials, and biological tissues, and is now used from chemistry and microbiology to clinical research.

By preserving the spatial localization of molecules, MSI provides a holistic view of molecular distributions within complex systems. The first requirement rests on the production of molecular ions without disturbing their initial spatial distribution. Matrix-assisted laser desorption/ionization (MALDI) is the most used technique for ionizing molecules in MSI. MALDI-MSI uses a laser to desorb and ionize molecules from a solid sample, enabling the generation of molecular images from the spatial distribution of detected molecules/ions. MALDI excels at detecting a broad spectrum of biomolecules, including metabolites, peptides, and proteins, with spatial resolution down to ~5 µm. In biomedical research, MSI contributes to understanding disease mechanisms, drug distribution, and the spatial organization of metabolites and proteins in tissues. It has, for instance, enabled the identification of molecular signatures associated with various pathologies, including cancer and neurodegenerative disorders[1,2].

In recent years, MSI has witnessed remarkable advancements, revolutionizing analytical chemistry across material science, biology, and, again, medical applications. Recent advancements have improved spatial resolution, quantification capabilities, and throughput, thereby bringing MSI closer to its position as a valuable tool for life and medical sciences[3]. Given the abundance of information this technology provides, informatics processing becomes imperative for distilling biologically significant signals from these intricate datasets. This first chapter introduces MSI as a technique, outlines the basic principles of mass spectrometry and MSI instrumentation, summarizes key biological applications and data-analysis strategies, and finally explains the motivations behind the development presented in this thesis.

1.1 Basic description of a mass spectrometer

1.1.1 Mass spectrometry principle

The basic principle of mass spectrometry relies on three steps, usually performed under high-vacuum conditions: first, the generation of ions from a sample and their transfer into the gas phase, then the separation of these ions according to their mass-to-charge ratio (*m/z*), and finally, their qualitative (*m/z*) and pseudo-quantitative detection based on ion abundances. Therefore, the three common elements of a mass spectrometer are an ion source, which generates ions; a mass analyzer, which separates them according to their m/z; and a detector that converts the signals generated by the ions into a measurable output (Figure 1.1).

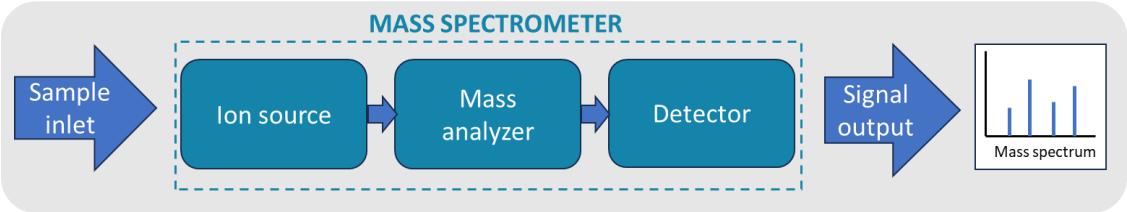


Figure 1.1. *Simplified scheme of a mass spectrometer composed of a sample inlet, an ion source, a mass analyzer, and a detector. The measured signals are ultimately transformed into a mass spectrum, a graph of signal intensity as a function of the mass-to-charge ratio of the detected ions.*

In the following section, we describe the general organization of a mass spectrometer and its three main constituents in the context of mass spectrometry imaging.

1.1.2 Ion source

The main role of the ion source is to extract or desorb molecules from the sample, ionize them, and transfer the resulting ions into the gas phase. Positively or negatively charging analytes is a prerequisite for mass spectrometry, as the technique relies on measuring the behavior of ions under electrical and/or magnetic fields to determine their mass-to-charge ratio (*m/z*). Ion sources are generally categorized into two main types: soft and hard ionization techniques[4]. Soft ionization methods limit analyte fragmentation by minimizing energy uptake during desorption/ionization, thereby favoring the formation of intact ions that remain sufficiently stable during analysis, whereas hard ionization methods involve higher energy input, leading to more extensive fragmentation. Such fragmentation can provide useful structural information, although it may also reduce the abundance of intact molecular ions.

In MSI, SIMS (secondary ion mass spectrometry) represents an example of an ionization method that can induce substantial fragmentation, especially under high primary-ion-dose conditions. In SIMS, the sample surface is bombarded with primary ions, leading to the generation of secondary ions[5]. In contrast, the most commonly used soft ionization methods in MSI are desorption electrospray ionization (DESI) and matrix-assisted laser desorption/ionization (MALDI). DESI uses accelerated charged solvent microdroplets to capture and ionize analytes from the sample surface[6]. MALDI, in contrast, uses a laser beam to desorb and ionize molecules from a matrix-coated sample surface[7]. Each ionization method has its own characteristics (Table 1.1), but due to its versatility, MALDI has become the predominant ionization method for MSI.

Property	MALDI	SIMS	DESI
Ionization method	UV laser	Primary charged particles (Cs^+, SF_5^+, Au and Sb clusters, C_{60}^+) eject and ionize material from the surface	Modified ESI source spraying solvent using high-pressure gas flow onto the sample surface
Sample preparation	Fresh-frozen sections (cryomicrotome) or FFPE sections; matrix application needed	Fresh-frozen sections, then directly analyzed	Directly analyzed (fresh-frozen or FFPE)
Matrix	Required	None	None
Environment	Vacuum (AP-MALDI also available)	Vacuum / low pressure	Atmospheric
Spatial resolution	Down to ~5 µm on commercial instruments; sub-micron reported in research settings	100 nm – 1 µm	Down to ~20 µm on the latest Waters DESI XS source; nano-DESI provides ~10 µm
Sample damage	Low — ionization proceeds via matrix co-crystallization / extraction rather than direct surface ablation; the matrix layer can be removed afterwards, allowing histological staining or IHC on the same section	Low	Low
Detected analytes	Wide range of biomolecules (lipids, metabolites, drugs, peptides, glycans, small proteins); intact proteins from tissue typically limited to ~30–40 kDa	Hydrophobic compounds with molecular weight not exceeding 1,000 Da	Compounds with molecular weight not exceeding 2,000 Da

Table 1.1. *Differences between matrix-assisted laser desorption/ionization (MALDI), secondary ion mass spectrometry (SIMS), and desorption electrospray ionization (DESI)*[7–14]

1.1.3 Matrix-Assisted Laser Desorption Ionization (MALDI)

MALDI is part of the broader family of laser desorption/ionization (LDI) techniques, in which a pulsed laser desorbs and ionizes analytes from a solid surface. In “classical” LDI, the laser interacts directly with the analyte or with a solid support, whereas in MALDI an excess of organic matrix mediates the absorption of laser energy and its transfer to embedded analytes. This matrix-assisted variant greatly improves ionization efficiency for labile and high-mass species and has become the dominant laser-based ion source for MSI.

MALDI was developed in the mid-1980s by Karas, Bachmann, and Hillenkamp[15], and shortly afterward, Tanaka and co-workers demonstrated its ability to ionize large biomolecules such as proteins[16]. Early MALDI ion imaging experiments had already been presented by Spengler and colleagues in an ASMS abstract[17]. A few years later, Caprioli and co-workers published the first peer-reviewed application of MALDI imaging to biological tissues, adapting the methodology to map peptides and proteins in tissue sections[7]. Since then, MALDI-MSI has gained popularity because it combines several key

advantages over other MS-based imaging approaches: it is a soft ionization method, is compatible with a wide range of biomolecules, and can be implemented on instruments with high mass resolving power and accuracy.

As a soft ionization technique, MALDI typically produces singly charged molecular ions with limited fragmentation during ionization[18,19]. This reduced fragmentation is crucial for maintaining the integrity of molecules, especially large biomolecules such as proteins and peptides. It allows the molecular mass of the entire molecule to be measured, ensuring a high specificity in molecule detection. MALDI is also relatively fast and compatible with minimal sample preparation for many solid biological specimens, including tissue sections, cell cultures, microbial colonies, and clinical biopsies.

The sample-preparation workflow is dictated by the class of molecule to be imaged, because the steps that make one class efficiently ionizable can suppress another[3,20]. The common starting point is minimal: tissue is snap-frozen, cryosectioned, and thaw-mounted onto a slide. For metabolites and lipids, the priority is minimal perturbation; as these species are small and often reasonably soluble, sections are analyzed with little or no washing, since the solvent washes used for proteins would extract or delocalize them[21]. Protein and peptide imaging follows the opposite logic: because abundant lipids and salts suppress their ionization, sections are first washed, typically in graded ethanol, to remove these species and enhance signal[22], and because intact high-mass proteins ionize poorly, an on-tissue tryptic digestion is commonly used to convert them into far more readily ionized peptides[23]. A more targeted, optional step, mainly for small molecules, is on-tissue chemical derivatization, in which a reagent tags a specific functional group to improve the ionization of otherwise poorly detected compounds such as steroids or neurotransmitters[20,24].

MALDI experiments start with the application of a matrix to the sample surface. MALDI matrices are low-molecular-weight organic compounds that (i) are soluble in solvents compatible with the sample, (ii) co-crystallize with the analytes, (iii) strongly absorb the laser wavelength (typically UV), (iv) efficiently transfer energy to analytes and promote their ionization, and (v) are stable under high vacuum[25]. The choice of matrix molecule primarily depends on the targeted molecules and the ionization mode. For example, α-Cyano-4-hydroxycinnamic acid (CHCA) is adapted to the detection of peptides (< 6kDa), small molecules such as lipids, and metabolites in positive mode, whereas 1,5-Diaminonaphthalene (DAN) can be used both in positive and negative ionization modes, for high spatial resolution imaging of lipids, nucleotides, peptides, and proteins. 2,5-Dihydroxybenzoic acid (DHB) is a third well-known matrix suitable for imaging proteins, peptides, and lipids in positive mode.

The matrix plays a central role in both desorption and ionization. It first aids in extracting analytes from the sample surface by co-crystallizing them. The resulting matrix crystals can exhibit diverse morphologies, ranging from small spots to sharp needle-like structures. The variation in crystal shapes is influenced not only by the matrix's chemical composition but also by the specific matrix-deposition strategy. After this step, a short ultraviolet (UV) laser pulse (nanosecond) is directed onto the sample's surface. Upon irradiation, the matrix, typically an aromatic organic compound, absorbs most of the laser energy, triggering a phase explosion and subsequent supersonic jet expansion that gives rise to the so-called "MALDI plume". This plume behaves like a dense plasma in which numerous ion-molecule reactions, such as proton transfers, generate ionized analyte molecules while preserving their molecular integrity. Although analyte molecules are efficiently transferred into the gas phase, only a small fraction is ultimately ionized and detected, with typical ion-to-neutral ratios on the order of $1:10^3$-10^4, giving rise to the "lucky survivor" description of MALDI analyte ions[25–27].

In the context of imaging, matrix application must be carefully controlled because it has a significant impact on ionization efficiency, spatial resolution, and analyte delocalization. The impact on spatial resolution is discussed in more detail in Section 1.2.2. Common strategies include dry sieving, spraying, or sublimation[28,29]. The sieving method is a straightforward approach that does not require expensive materials. A dry matrix is applied simply with a stainless-steel sieve. This approach, however, is progressively forsaken because it does not allow acceptable spatial resolutions to be reached.

Nowadays, spraying methods deposit a matrix solution as fine droplets (typically ~10–50 µm in diameter) that extract analytes from the surface and co-crystallize upon drying. Multiple spray–drying cycles are used to build up matrix coverage. Parameters such as solvent composition, flow rate, nozzle temperature, spray velocity, and number of passes strongly influence crystal size, homogeneity, and analyte extraction. Excessively “wet” conditions promote analyte diffusion and delocalization, whereas overly “dry” conditions can lead to poor extraction and reduced sensitivity. As too much solvent on the surface leads to molecule delocalization and a loss of spatial resolution, solvent-free matrix application appears attractive.

Sublimation is a phase transition in which a solid directly converts to a gas without passing through the liquid state. Sublimation used in MSI provides a more uniform matrix deposition and a better spatial resolution. In MSI, sample preparation/coating particularly impacts the image quality. The crystal formed by the matrix should be small and regular (100–200 nm) to avoid the formation of hot spots and intense delocalization[30]. The formation of hot spots is caused by uneven matrix coating and co-crystallization, leading to higher punctual ionization efficiency and, therefore, higher signal in those regions. This process must be controlled to ensure consistent signal comparison between samples and improve reproducibility. To that end, the spraying process must be automated by systems such as the TM-Sprayer from HTX Technologies. This system robotically controls factors such as temperature, solvent

flow rate, spray velocity, and the number of passes, thereby increasing reproducibility[29].

It is finally important to notice that matrix molecules also become ionized during the MALDI process. When the matrix is applied in excess relative to the sample, the resulting mass spectra primarily contain low *m/z* ions and matrix aggregates. Since matrix molecules are more abundant than the analytes, they can overwhelm analyte signals, making it challenging to identify the target analytes. As a result, a significant chemical background is established in the low *m/z* range, which can significantly affect the detection of small molecules[19].

1.1.4 Mass analyzers

The main objective of a mass analyzer is to separate the generated ions based on their mass-to-charge ratios. Mass analyzers differ in how they separate ions[31]. For instance, quadrupole analyzers employ radiofrequency and direct current voltages to selectively stabilize ions with specific mass-to-charge ratios through a quadrupolar field. Ion trap analyzers use radiofrequency and static electric fields to trap ions in a 3D space and analyze their mass based on their oscillation frequency. On the other hand, Time-of-flight (TOF) analyzers rely on the acceleration of ions and their separation in a free-field flight tube, where they are separated according to their velocities, and the time they need to travel the TOF is precisely measured to determine their mass-to-charge ratios. Fourier-transform ion cyclotron resonance (FT-ICR) analyzers exploit the cyclotron frequency of ions orbiting in a magnetic field, enabling ultra-high-resolution mass analysis. Orbitrap analyzers use a combination of electric fields to trap and orbit ions, measuring their axial frequencies to determine mass-to-charge ratios.

The effectiveness of a spectrometer is determined by its mass resolving power, speed, and mass accuracy. These factors must be carefully considered based on the specific requirements of an experiment. The mass resolving power is a fundamental parameter that indicates the ability of a mass analyzer to separate ions with similar masses. The acquisition speed is the number of spectra generated per second (pixels per second in MSI). This parameter is critical in high-throughput applications where rapid data acquisition is essential. Time-of-flight (TOF) analyzers are popular for their exceptional speed, making them suitable for scenarios where quick results are imperative, such as in drug discovery or environmental monitoring. On the other hand, mass accuracy measures how closely the measured mass of an ion corresponds to its actual mass. Section 1.2.1 provides more details on mass accuracy and mass resolving power.

Mass Analyzer	Mass Resolving Power ($m/\Delta m$)	m/z Range	Throughput (pixels/s)
Linear TOF	~10^3	0 – 10^6	2 – 50
Reflectron TOF	10^3 – 10^4	0 – 10^5	2 – 50
TOF/TOF	10^3 – 10^4	0 – 10^4	2 – 10
Q-TOF	10^4 – 10^5	50 – 20k	5 – 20
Linear Ion Trap	10^3 – 10^4	50 – 4k	< 1
Triple Quadrupole	~10^3	5 – 3k	> 100 (MRM)
FT-Orbitrap	10^4 – 10^6	40 – 80k	< 1 – 5
FT-ICR	10^5 – 10^6	100 – 10k	1 – 3

Table 1.2. *Characteristics of Mass analyzers used with a MALDI source for MSI*[32].

As this thesis's MSI experiments were primarily performed using MALDI-FT-ICR and MALDI-TOF mass spectrometers, the chapter will focus on FT-ICR analyzers, with some attention to TOF. Therefore, the two types of mass analyzers are described.

1.1.4.1 Time of Flight (TOF) mass analyzers

The operating principle of time-of-flight (TOF) mass spectrometry is based on a straightforward principle as ion's *m/z* is computed according to the time they need to cross a field-free drift path of a predetermined length up to reaching the detector after acceleration[4]. Indeed, after the ionization, ions (mass m_i, charge q) are extracted from the source and accelerated under a constant electric field of voltage U. The potential energy E_p of the ions in the electric field is then given by Equation 1.

$$E_p = qU \ (1)$$

When an ion is accelerated, its potential energy is converted into kinetic energy E_k (Equation 2), which depends on its mass m_i and its velocity v_i (Equation 3). This equation allows us to easily isolate the expression of the ion velocities (Equation 4).

$$E_p = E_k \ (2)$$

$$qU = \frac{1}{2} m_i {v_i}^2 \ (3)$$

$$v_i = \sqrt{\frac{2qU}{m_i}} \ (4)$$

The velocity of the ions appears then directly proportional to the square root of the charge, of the voltage, and to the inverse of the mass. After the acceleration region, the ions enter a free-field region at a constant velocity. As the field-free region, called drift tube or time-of-flight tube (TOF), has a fixed given length of l, it is possible to express the velocity in the function of this length l and of the time taken by the ions to cross the TOF and reach the detector (Equation 5).

$$v_i = \frac{l}{t} = \sqrt{\frac{2qU}{m_i}} \Leftrightarrow t = \frac{l}{\sqrt{\frac{2qU}{m_i}}} \Leftrightarrow t = \frac{l}{\sqrt{2U}} \sqrt{\frac{m_i}{q}} \ (5)$$

The experimental value t directly depends on the square root of m_i/q, classically called m_i/z in mass spectrometry, as $\frac{l}{\sqrt{2U}}$ is constant for the mass analyzer. If two ions have

the same charge and are accelerated from the same initial position, the lighter one will reach the detector before the heavier one. This equation also suggests that measuring an ion of any mass is theoretically possible, as there is no limit to ion mass. However, in practice, issues can be met with the detector's efficiency at higher masses, particularly its insensitivity to slow ions.

Advancements in the time-of-flight analyzer technology have led to different geometries, each designed to offer unique and valuable properties. The most basic TOF analyzer, the linear TOF, has a simple design, making it relatively easy to implement and suitable for a wide range of applications. Linear TOF lacks mass resolution due to significant variations/distributions of the initial kinetic energy of ions sharing the same *m/z*[33]. Correcting this phenomenon to increase the mass resolution has been made possible by the emergence of pulsed delayed extraction and reflectron technologies.

Pulsed delay extraction addresses this issue by applying the extraction field only after a controlled delay following ion formation. During this delay, ions of the same *m/z* separate in space according to their initial velocities, with faster ions drifting further from the back electrode than slower ones. When the extraction pulse is then applied, the slower ions, being further from the exit, traverse a larger potential difference and gain more kinetic energy than the faster ions. As a consequence, the initially slower ions catch up to the initially faster ones at a well-defined space–time focal plane[4]. By tuning the delay time and the extraction voltage, this focal plane is positioned onto the detector surface, minimizing the arrival-time spread of isobaric ions and thereby enhancing the resolution of the TOF analyzer.

Reflectron also addresses this issue. In such mass spectrometers, a reflecting electrostatic mirror is incorporated at the flight tube end. This element forces ions with higher kinetic energy to travel a longer path[4] before being reflected to the detector (Figure 1.2). Indeed, as the electrostatic mirror is made of a succession of electrodes held at increasing repulsive potentials, ions with lower kinetic energy are reflected after

a shorter penetration depth, whereas faster ions penetrate deeper into the mirror and therefore travel a longer path before being reflected. The reflector effectively time-refocuses the ions, causing ions with different initial kinetic energies but the same *m/z* to reach the detector simultaneously. Reflectron TOF analyzers are often chosen for high-resolution applications where precise mass determination is crucial.

However, because all ions travel a longer path in the reflectron than in the linear mode, some are lost along the way, resulting in reduced sensitivity. Simply, the linear mode is less resolutive but more sensitive than the reflectron mode.

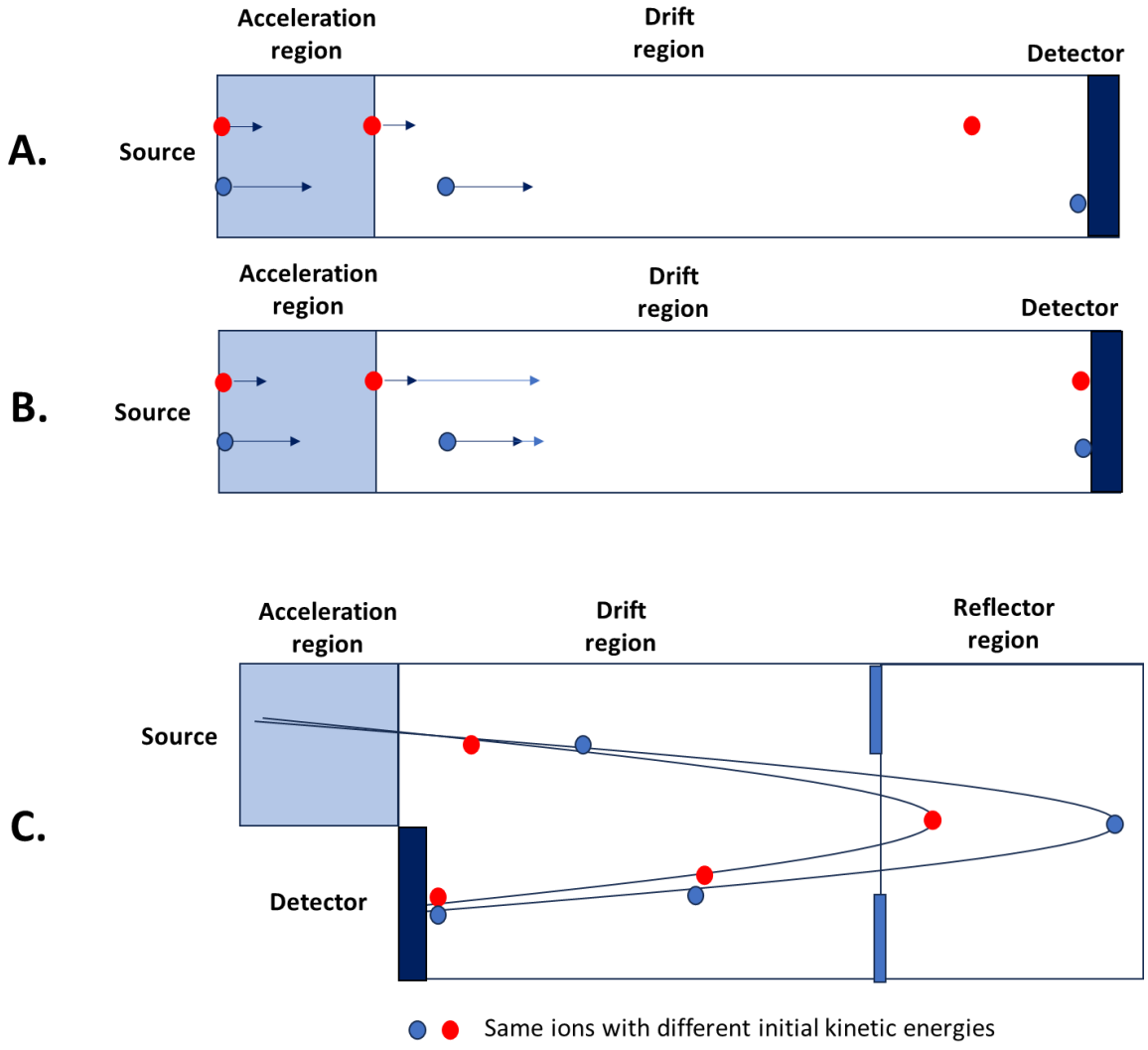


Figure 1.2. *Schematic illustration of a TOF analyzer in A. linear continuous extraction mode, B. linear with a delayed pulsed extraction mode, C. Reflectron TOF.*

MALDI produces ions in a pulsed manner, which matches the operational mode of TOF analyzers[4]. Each laser pulse in MALDI generates a burst of ions, which can be analyzed in a single TOF acceleration pulse. This combination is highly efficient for capturing transient ion populations and makes TOF perfect for MALDI imaging applications.

1.1.4.2 Fourier transform ion cyclotron resonance analyzer (FT-ICR)

In Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FT-ICR MS), ions are separated based on their motion in a cell (ion trap) installed in the heart of a strong magnetic field (Figure 1.3).

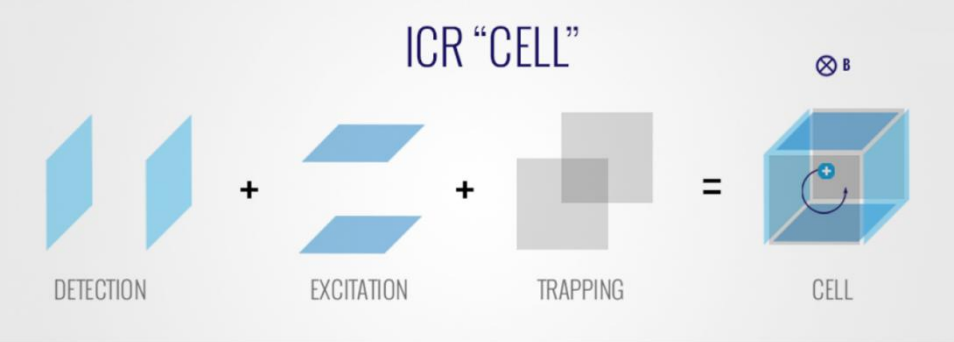


Figure 1.3. *Schematic view of an ICR cell. Reproduced from Mass Spec Pro, www.massspecpro.com*

When ions are injected and trapped in the cell by the action of two trapping plates, the ions adopt a circular motion, perpendicular to the magnetic field direction, at a frequency that depends on their *m/z*[4]. This circular motion is caused by the Lorentz force, which is the force experienced by a charged particle q, trapped in an electromagnetic field B. The force is perpendicular to the velocity v of the charged particle and the direction of the magnetic field. If $\vec{v}$ and $\vec{B}$ are vectors that are at a right angle to each other, then the force can be computed as a scalar product (Equation 6).

The charged particle under this force will be forced to move in a circular trajectory of radius r_m and, considering a charged particle with a mass m, the related force can be computed according to Equation 7. The latter permits the isolation of the radius r_m of the movement, which directly appears as a function of the m/q ratio (Equation 8). Knowing that the velocity v is the product of the radius r_m by the angular frequency, ω_c , the latter can be expressed (Equation 9), then converted into the cyclotron frequency (Equation 10).

$$F_L = qvB \quad (6)$$

$$F_L = qvB = \frac{m_i v^2}{r_m} \quad (7)$$

$$r_m = \frac{m_i v}{qB} \quad (8)$$

$$\omega_C = \frac{qB}{m_i} \quad (9)$$

$$f_C = \frac{\omega_C}{2\pi} = \frac{B}{2\pi} \cdot \frac{q}{m_i} \quad (10)$$

This equation clearly demonstrates that a precise measure of the frequency resonance of each ion trapped in the cell can lead to a precise computation of its m/q, the ratio $\frac{B}{2\pi}$ being a constant (assuming that the magnetic field is constant over time). To do so, a pair of excitation and detection plates is placed on opposite sides of the ion's orbit (Figure 1.3). As the ions initially oscillate in the middle of the cell, far from the detection plates, no induced current can be measured. The ions are then excited by applying a transverse electric signal that oscillates at the frequency of the desired ion (f_C). This excitation increases their kinetic energy and therefore their orbital radius, while conserving their cyclotron frequency. In other words, the additional energy is encoded in the radius of the orbit, not in a change of frequency. As the radius increases, the ion packet approaches the detection plates and induces an image current whose temporal oscillation directly reflects the cyclotron frequency. This readback is referred to as a

free induction decay (FID). Once all the ions have been excited, the final FID is decomposed into individual frequencies using Fourier transformation, which can be converted to *m/z* (Equation 10), to yield a mass spectrum. Thanks to the high vacuum in the cell, FT-ICR can maintain stable ion trajectories, enabling long FID measurements and very accurate frequency measurements. The related mass spectra are consequently made of very high peak resolution, associated with very accurate ("exact") masses.

1.1.5 Mass spectrometry imaging

In MSI, mass spectra are acquired at many discrete locations across the sample surface. Practically, the sample is virtually divided into a grid of pixels, with a typical lateral size ranging from ~20 to 200 µm depending on the application and instrument. The laser (for MALDI) or primary ion/solvent beam (for SIMS or DESI) is directed sequentially to the center of each pixel, and the material within the irradiated area is desorbed and ionized. A full mass spectrum is recorded at every pixel, so that each position on the grid is represented by a spectrum describing the local molecular composition.

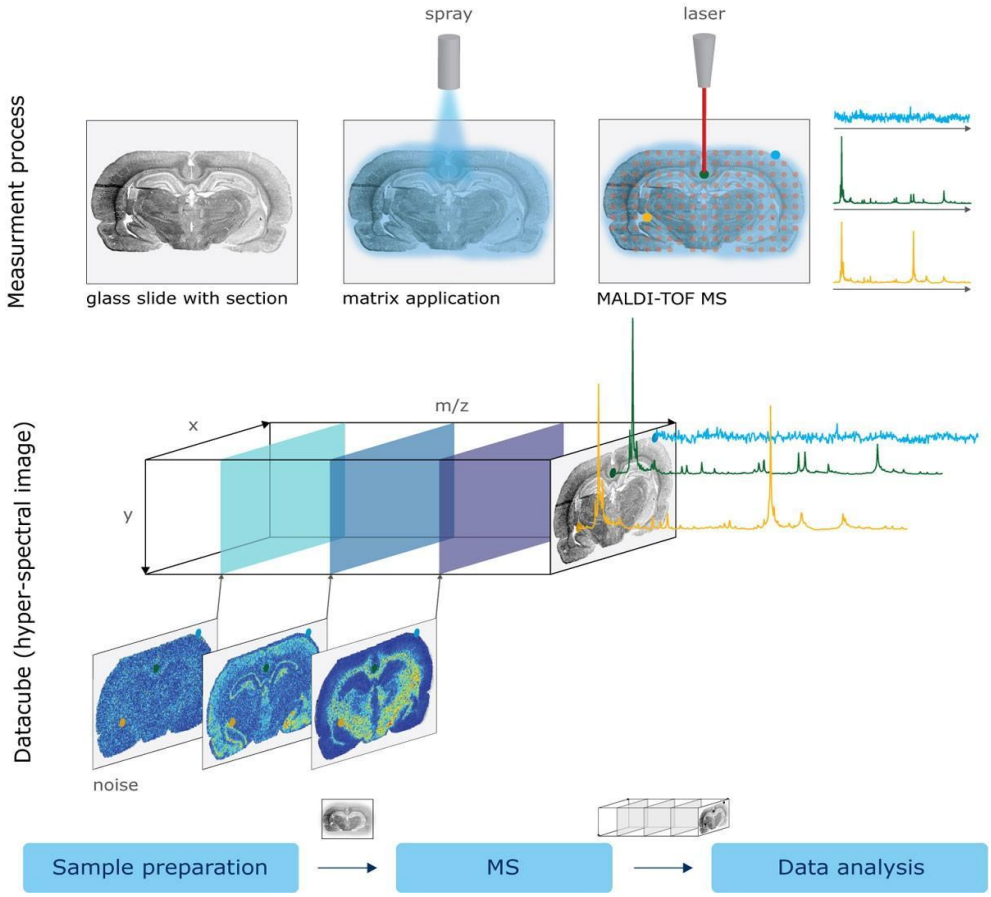


Figure 1.4. *MALDI-imaging data acquisition workflow from*[34]*. MALDI-imaging data acquisition workflow and data representation as a data cube or a hyperspectral image represented by a 3D data cube, where the three dimensions correspond to 2 spatial dimensions (X and Y) and the m/z dimension. Here, ion images are represented as heat maps by a jet colormap from blue (lowest intensity) to yellow to red (highest intensity). Ion images with a similar colormap are colocalized on the sample, while ion images with a dissimilar colormap are anti-localized.*

Because of this acquisition strategy, MSI data are often described as hyperspectral images. In a conventional RGB image, each pixel is characterized by three intensity values (red, green, and blue). In MSI, each pixel is instead characterized by a full mass spectrum containing intensities over many *m/z* values. Therefore, MSI is often represented by a 3D data cube, where the three dimensions correspond to the two spatial dimensions (X and Y) and the *m/z* dimension[34]. This data cube can be used to extract the spatial and spectral information of the detected molecules in a sample. A

spectrum containing all the detected *m/z* values can also be obtained by summing all the mass spectra of the image. Such a spectrum is generally called a total ion spectrum and is sometimes called an average spectrum. It helps gain insights into the average molecular composition of the entire sample or even specific regions of interest. It can help to identify common molecular species across the sample and provide valuable information about their chemical composition and heterogeneity. Finally, using a data cube, one can create an ion image that shows the distribution of an ion's measured abundances across the sample. This ion image is a 2D heat map that shows the intensity of a specific *m/z* across all acquisitions (Figure 1.4). You can also think of a data cube as a collection of ion images, each corresponding to a different *m/z*.

1.2 Importance of spatial and spectral resolution

1.2.1 Spectral resolution and mass accuracy

In mass spectrometry imaging, mass resolving power (or resolution, often considered a synonym[35]) and mass accuracy are key parameters, especially in untargeted analyses where the goal is to detect and identify as many molecular species as possible without predefining targets. High resolving power helps separate closely spaced signals, while high mass accuracy improves the confidence of molecular formula assignment.

Mass resolving power and mass accuracy depend strongly on the type of mass analyzer used. Mass resolving power can have various definitions, but we decided to use the proposed definition of K. Murray of Mass resolving power, which is the ability to distinguish between peaks differing in the *m/z* by an increment Δ(*m/z*) given by (*m/z*)/Δ(*m/z*). The value of Δ(*m/z*) is obtained from the peak width at a specified fraction of the peak height or from the valley between two peaks of specified height[35].

Here, mass resolving power is calculated by dividing the m, the *m/z* of the peak maximum, by the peak width 50% ($\Delta m_{50\%}$) of peak height[36] (Equation 11).

$$Resolving\ power = \frac{m}{\Delta m_{50\%}} \quad (11)$$

For example, in Figure 1.5, the peak has a resolution of $\frac{1058.6746}{1058.6692-1058.6803} = 95\,376$ at *m/z* 1058. It's important to specify the mass-to-charge ratio (*m/z*) when calculating mass resolving power because it's dependent on the mass. The mass resolving power differs based on the mass analyzer and generally decreases as *m/z* increases. For example, in an FT-ICR, the mass resolving power is inversely proportional to the mass (*m/z*), meaning that as the mass of the ion increases, the mass resolving power decreases.

Different mass resolving powers can be achieved according to the mass analyzer (Table 1.2). At *m/z* 400, a linear geometry time-of-flight (TOF) analyzer can achieve an approximate mass resolution of 1,000. On the other hand, a reflector geometry TOF analyzer can achieve a mass resolution of 50,000. Furthermore, FT-ICR can provide even higher resolutions, up to 10,000,000.

In untargeted mass spectrometry, high resolution is particularly important because it improves the separation of isobaric and near-isobaric species, molecules that share the same nominal mass but differ in elemental composition or structure. When resolution is insufficient, their signals overlap, leading to ambiguous or incorrect peak assignments and biased quantification. This problem is exacerbated in complex biological samples (tissues, microbial communities), where many endogenous and exogenous compounds contribute to dense, crowded spectra. High resolving power also facilitates the detection and characterization of low-abundance analytes that may carry important biological or diagnostic information, but whose peaks are easily masked by more intense neighbors.

Alongside resolving power, mass accuracy is equally critical for untargeted MSI. According to the IUPAC definition, mass accuracy is "the ability of a measuring instrument to give responses close to a true value." In our context, the instrument response is the measured *m/z* of a peak (often at its apex), and the "true value" is the theoretical *m/z* calculated from the ion's molecular formula and charge state. The mass error can be expressed as an absolute difference in Daltons (Equation 12), but is most commonly reported in parts per million (ppm) (Equation 13), which normalizes the error by the ion mass.

$$Mass\ Difference\ (Da) = Measured\ Mass - Theoretical\ Mass \quad (12)$$

$$Mass\ Accuracy\ (ppm) = \left(\frac{Measured\ Mass - Theoretical\ Mass}{Theoretical\ Mass}\right) \times 10^6 \quad (13)$$

Figure 1.5 illustrates this point for the ion $[C_{53}H_{93}N_7O_{13} + Na]^+$, for which the experimentally measured mass differs from the theoretical one. The mass accuracy can be measured at $\left(\frac{1058.67481-1058.67236}{1058.67236}\right) \times 10^6$= 2.3 ppm. Different analyzers achieve different levels of mass accuracy. Linear TOF instruments often provide accuracies in the tens of ppm range, whereas reflectron TOF instruments can reach low single-digit ppm under good calibration conditions. FT-ICR and Orbitrap analyzers can achieve sub-ppm mass accuracy at typical MSI *m/z* values (e.g., around *m/z* 400) when properly calibrated and operated.

In MSI, molecular annotation can in practice be approached at three distinct levels of evidence, each with its own trade-off between spatial coverage and identification confidence. MS1-only annotation from MSI data, in which the assignment relies solely on the accurate mass and, when available, the isotopic pattern measured at each pixel or over all pixels. This is the default annotation mode in MSI, the only one that retains the full spatial information of the image, and the most demanding in terms of mass accuracy and resolving power, since no fragmentation evidence is available to

disambiguate isobaric candidates. High mass accuracy directly narrows the list of plausible molecular formulas for each peak, reduces false-positive assignments, and increases confidence in metabolite or lipid identifications[37], which is why sub-ppm accuracy is so critical when MS1 is the only level of evidence. (ii) On-tissue (in situ) MS/MS, in which fragmentation spectra are acquired directly on selected pixels or regions of interest using MS/MS-capable MSI instruments (Q-TOF, TOF/TOF, FT-ICR, Orbitrap, timsTOF). Because it consumes additional sample material and acquisition time, this strategy is typically restricted to a small number of target ions and is rarely performed at every pixel of an entire image; however, the MS/MS spectra obtained for those targeted ions provide direct structural evidence, fragment patterns, neutral losses, and diagnostic ions that cannot be inferred from accurate mass alone. (iii) Orthogonal LC–MS/MS follow-up, in which regions of interest highlighted by the MSI experiment are isolated (typically by laser microdissection), extracted, and re-analyzed by LC–MS/MS. This provides the most complete structural characterization, combining chromatographic separation, retention-time information, and high-quality fragmentation spectra, but at the cost of losing the direct spatial registration with the MSI image. The three strategies are complementary rather than exclusive: MS1 annotation across the whole image, on-tissue MS/MS on a few selected target ions, and LC–MS/MS on dissected regions of interest. Because per-pixel MS/MS is rarely feasible on a complete image due to limited material and acquisition time, the mass accuracy and resolving power of the MS1 measurement remain central to reliable molecular identification and interpretation in MSI datasets. Consistent with this, the methodological developments presented in this thesis rely exclusively on MS1-only annotation.

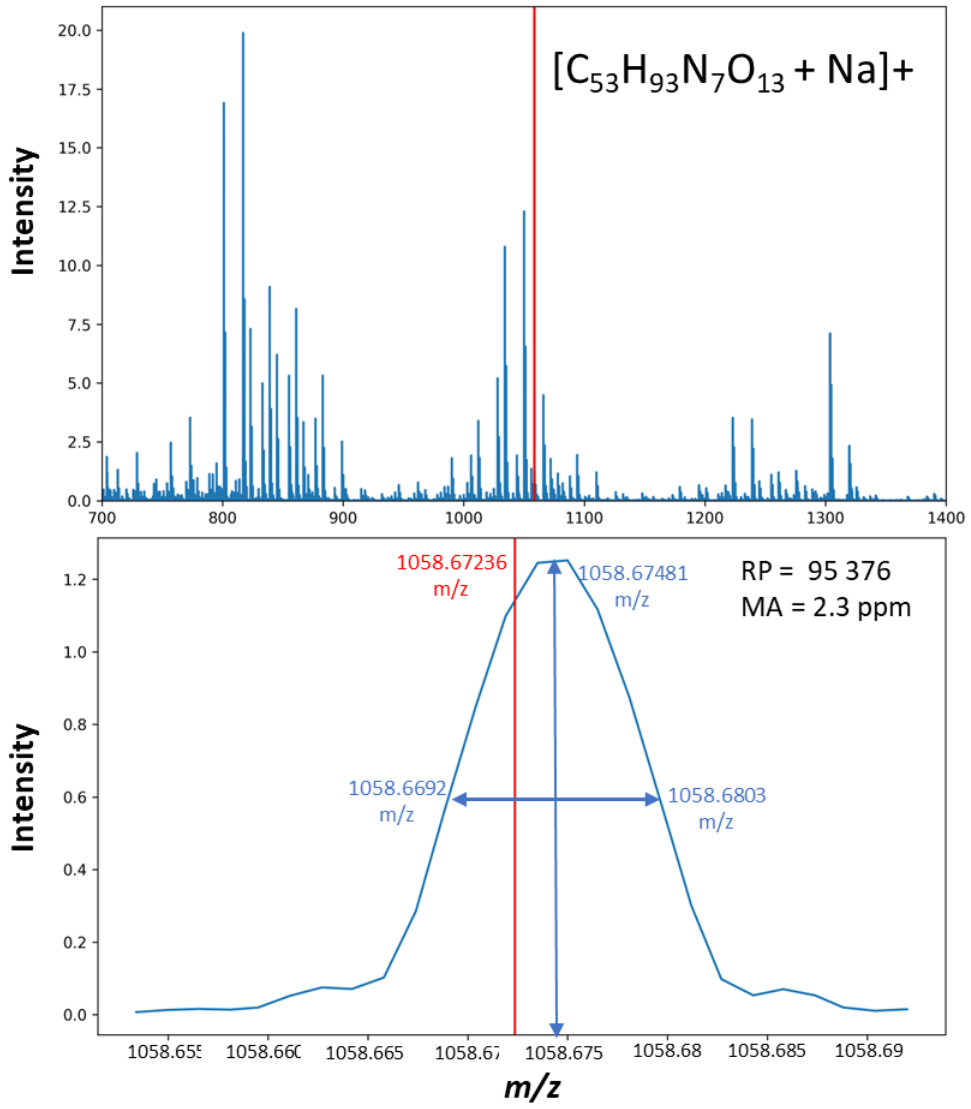


Figure 1.5. *Example of the mass spectrum acquired from bacterial colonies, with an FT-ICR mass spectrometer in MALDI mass spectrometry with CHCA as a matrix. The vertical blue line corresponds to the maximum of the peak. The horizontal blue line is the peak width at 50% of peak height. The horizontal red arrow corresponds to the mass difference between the theoretical and experimental values. RP is the peak resolution, and MA is the mass error of the peak maximum compared to the theoretical mass of* $[C_{53}H_{93}N_7O_{13} + Na]^+$ *represented by a vertical red line.*

1.2.2 Spatial resolution and MALDI imaging

In MSI, spatial resolution determines the smallest structures that can be distinguished in the resulting images. For biological samples, these structures can range from gross anatomical regions (hundreds of micrometers) to individual cells and even subcellular compartments (down to the sub-micrometer scale), depending on the ion source and acquisition mode (Table 1.1)[38–42].

For a fixed field of view, higher spatial resolution corresponds to smaller pixels and a larger number of spectra. The "appropriate" resolution depends on the biological question: for example, a pixel size of ~100 µm is typically sufficient to resolve bacterial colonies and their interaction zones[38,39]. This resolution is also sufficient to map the main anatomical zones in a particular organ, such as the different sections of a murine brain[40]. To distinguish between different cell types, a resolution of around 1 µm is required. In comparison, subcellular structures such as organelles or single bacteria require a resolution below 1 µm.

The achievable spatial resolution is constrained both by the ion source and by practical considerations such as sensitivity, acquisition time, and data volume. Unlike MALDI, DESI and SIMS do not require a matrix-coating step (see Section 1.1.2 and Table 1.1), their spatial resolution is governed by the geometry of the solvent spray or primary-ion beam rather than by matrix crystal size and the matrix-driven delocalization discussed below for MALDI. For DESI, typical pixel sizes are now in the range of ~20–200 µm on the latest commercial sources (e.g. the Waters DESI XS)[14], with nano-DESI implementations reaching down to ~10 µm[13]. SIMS spans a much wider range depending on the acquisition regime. In static SIMS, the primary-ion dose is kept below the so-called static limit (typically $<10^{13}$ ions/cm^2) so that each surface site is sampled at most once; this regime preserves intact molecular and low-mass fragment ions and typically achieves spatial resolutions of ~1 µm[5]. In dynamic SIMS, including nanoSIMS, much higher primary-ion doses are used and the surface is progressively eroded, which

restricts the detection to atomic and small fragment ions but pushes spatial resolution down to ~50–100 nm, making this regime particularly suited to elemental and isotopic subcellular imaging[5,42]. Routine MALDI-MSI typically operates at 10–50 µm, but recent commercial instruments now reach down to ~5 µm at the pixel level, with sub-micron resolutions reported in dedicated research settings[11].

A MALDI acquisition of a sample containing cells (grey areas) is schematized (Figure 1.6). Before image acquisition, a pixel grid (or raster) must be defined over the sample surface (black squares). The laser irradiates the center of each pixel. The laser shots generate ablation marks (blue traces) that correspond to the part of the sample being ionized, and the ablation diameter depends on the MS spectrometer acquisition settings. The spacing between pixel centers corresponds to the previously defined spatial resolution of the image. Remarkably, the surface covered by a pixel appears higher than the ablation spot, which indicates that the signal extracted from a pixel does not correspond fully to the part of the sample covered by that pixel. A difference between the pixel surface and the ionized surface is common in MALDI imaging. Thus, precautions must be taken when comparing an MSI to its optical image (microscope), as not all biological features within a pixel may be ablated and detected. Increasing the spatial resolution by reducing the spot size and/or the spacing between adjacent spots can help limit this risk. However, there is a trade-off between spatial resolution and sensitivity[3] since small ablation leads to less material ionization. Moreover, increasing the spatial resolution will also increase the acquisition time.

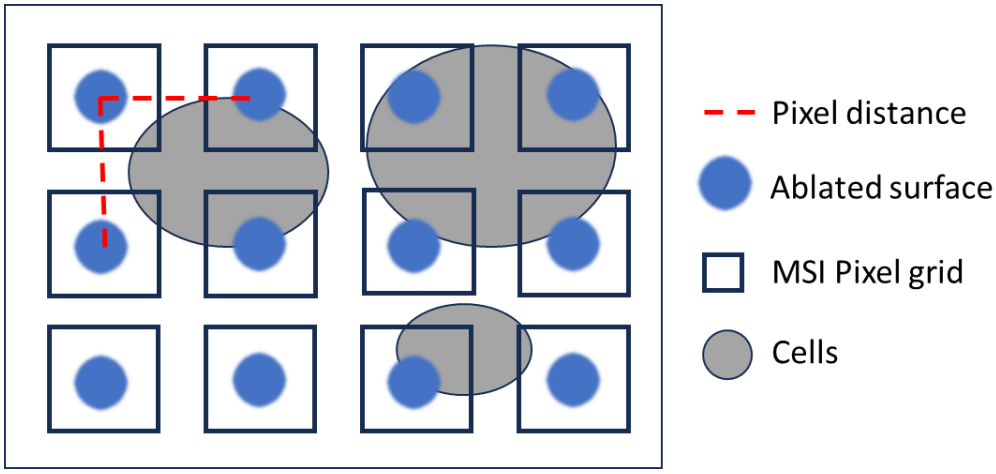


Figure 1.6. *Representation of a MALDI-MSI experiment of a sample containing cells (a grey area). The blue spots correspond to laser-ablated regions of the sample. The black square delimits the MSI's pixel grid (raster). The red lines correspond to the distance between the acquisitions and define the MSI pixel size and spatial resolution.*

A complementary approach, implemented on instruments such as the Bruker rapifleX MALDI Tissuetyper, is to actively scan the laser beam within each pixel rather than relying on a single static shot at the pixel center. In this configuration, the Smartbeam 3D laser is focused to a spot of ~5 µm and is rapidly deflected across the pixel area during the acquisition of that pixel's spectrum, so that the ablated region effectively fills a square pattern matching the user-defined pixel size[43]. Compared with a conventional circular Gaussian ablation spot smaller than the pixel, this strategy produces nearly contiguous, square-shaped sampling: more of the surface contained within each pixel is actually ablated and contributes to the spectrum, the gap between adjacent pixels is minimized. This significantly reduces the discrepancy between the pixel surface and the ionized surface highlighted in Figure 1.6 and, in practice, allows pixel sizes down to ~5 µm without resorting to oversampling.

In addition to spot size and raster spacing, oversampling strategies can be used to increase the apparent spatial sampling. Oversampling refers to acquisition strategies where the distance between successive laser shots (the raster step) is smaller than the laser ablation spot diameter. In that case, neighboring pixels partially probe overlapping regions of the sample. Oversampling can improve apparent spatial

sampling beyond the nominal laser spot size, but it also has drawbacks: adjacent pixels are no longer independent measurements, and quantitative interpretation or spatial statistics may be biased if this redundancy is not accounted for.

Analyte delocalization is another major factor that degrades spatial resolution. Delocalization refers to the diffusion or displacement of molecules away from their original position at various stages of sample preparation and matrix deposition[39]. At the first stage, for tissue MSI, sections are typically cut to a thickness of 2–7 µm for formalin-fixed paraffin-embedded (FFPE) samples and 7–20 µm for fresh-frozen sections, mounted, and dried. These steps can already introduce some molecular and morphological distortion.

However, the most important source of delocalization is often the matrix application stage. Analyte delocalization caused by matrix application to the sample arises from two phenomena: analyte diffusion due to excessive solvent deposition and crystal size. To reduce analyte diffusion, the matrix must be sprayed onto the sample within an optimal range of “wetness”. The spray must remain sufficiently “wet” to allow the matrix solution to penetrate the tissue and crystallize with the analytes before the solvent evaporates. However, it must also be “dry” enough to prevent the matrix from overwhelming the analytes of interest, possibly leading to matrix suppression[44,45].

Finally, the crystal size clearly impacts the spatial resolution. This phenomenon can be linked to a delocalization process that happens during the crystallization of matrix molecules. During crystallization, the molecules of the samples diffuse to form a crystal. During acquisition, when this crystal is ablated, all molecules within it are analyzed simultaneously. Suppose the crystal is large enough to extend beyond the pixel area. In that case, the compound outside the area will be misassociated with this pixel, reducing the MSI's resolution.

Careful control of matrix deposition is therefore essential to preserve spatial resolution. The main matrix-application strategies, dry sieving, spraying, and sublimation,

together with the role of automated sprayers such as the TM-Sprayer in producing small, regular, and reproducible crystals, have already been described in Section 1.1.3.

Overall, the achievable spatial resolution in MALDI-MSI reflects the combined contributions of the laser spot size, the raster step (including possible oversampling), and the matrix-crystal size and homogeneity. Optimizing matrix deposition is consequently as critical as optimizing the laser and acquisition settings when targeting cellular or subcellular resolution.

1.3 Main applications of MSI

Most MSI experiments follow the same three-step logic. The first step is to identify the mass signals (peaks) whose intensity is concentrated in specific regions of the sample, regions that often correspond to a biological feature of interest, such as the interaction zone between two microbial colonies or a cancerous area within a tissue. The second step is to annotate these region-specific peaks, that is, to assign each one a molecular formula from its exact mass and, where available, complementary MS/MS information. The third step is to compare the abundance and spatial distribution of these annotated signals across experimental conditions, so that the differences can be interpreted in biological terms.

1.3.1 Example of applications

In microbiology and natural product discovery, MSI enables in situ detection of specialized metabolites[46]. In a typical experiment, bacterial strains are grown alone or in co-culture with other microorganisms (e.g., different strains or species) on agar-based media. Each strain initially develops in a defined region, forming colonies that communicate with their surroundings via secreted chemicals. In response to neighboring organisms (bacteria, fungi, etc.), they activate otherwise cryptic biosynthetic gene clusters (BGCs), that is, clusters present in the genome but silent under standard monoculture conditions, leading to the production of specialized

metabolites that are not observed when the strain is grown alone. The metabolites produced depend on the particular strains placed in competition and represent a rich source of potential bioactive compounds (e.g., antibacterial or antifungal agents). MSI is especially useful here because it can detect mass signals that co-localize with specific colonies or interaction zones, thereby highlighting compounds produced during microbial confrontation[47,48].

A second application of MSI is linked to the advances in "single-cell technologies," which have shown the importance of cellular metabolism within tissues by unveiling cell-cell heterogeneity[49]. Recently, the SpaceM approach was developed to approximate single-cell metabolomics by combining MSI with microscopy and computational analysis[50]. In SpaceM, images of the same cell culture are acquired by MALDI-MSI and optical microscopy. An image-processing pipeline then identifies individual cells on the microscopy image and estimates, for each cell, the associated mass spectrum by aggregating the MALDI laser shots overlapping the cell footprint. This enables the characterization of metabolic states and cell-to-cell heterogeneity, for example, by detecting and identifying lipid signals associated with specific cell populations in differentiated human hepatocytes[50].

MSI can also be combined with LC–MS/MS to improve compound identification and structural characterization as introduced in Section 1.2.1. MSI provides spatial maps of molecular distributions, whereas LC–MS/MS offers complementary physical and chemical information through chromatographic separation (retention times) and fragmentation spectra. A common strategy is to use MALDI-MSI to identify regions with distinct molecular profiles, then isolate these regions by laser microdissection, extract their contents, and analyze the extracts by LC–MS/MS[51,52]. This hybrid pipeline preserves the spatial targeting power of MSI while adding the retention-time and high-quality fragmentation information needed to assign molecular formulas and structures with higher confidence than MS1 alone.

In clinical research, MSI is increasingly used to study human tissues, particularly in oncology. When combined with histological and morphological data, MSI allows spatially resolved molecular information to be integrated with classical pathology readouts[53]. This integration provides a more comprehensive description of tissue morphology and molecular state, opening up applications such as tumour classification, identification of prognostic markers, and prediction of treatment response. We refer the reader to Vaysse *et al.* for a detailed overview of MSI applications in clinical research[54].

An interesting application of mass spectrometry imaging is in vivo imaging. Nina Ogrinc *et al*. developed a robotically assisted topography MSI coupled with water-assisted laser desorption ionization technology (SpiderMass) to achieve real-time in vivo imaging. The method allows the analysis of Metabo-lipidomic molecular profiles of living cells in tissue without alteration by making two main developments. First, the method uses an ionization process based on the resonant excitation of the endogenous water molecules. Secondly, they developed a manually movable probe that doesn't make contact with the sample. The goal of this method is to be applied in clinical settings, such as in oncology surgery, to delineate better cancer tissues[55]. To achieve this goal, it's crucial to integrate efficient data processing and mass spectrum classification programs to achieve automatic real-time diagnostics.

1.3.2 Data treatments

One of the main bottlenecks in MSI is the time-consuming extraction of biologically meaningful information from complex datasets. Multiple computational steps are required to transform raw signals into interpretable molecular features. This processing can be grouped into three broad levels of analysis[56] (Figure 1.7).

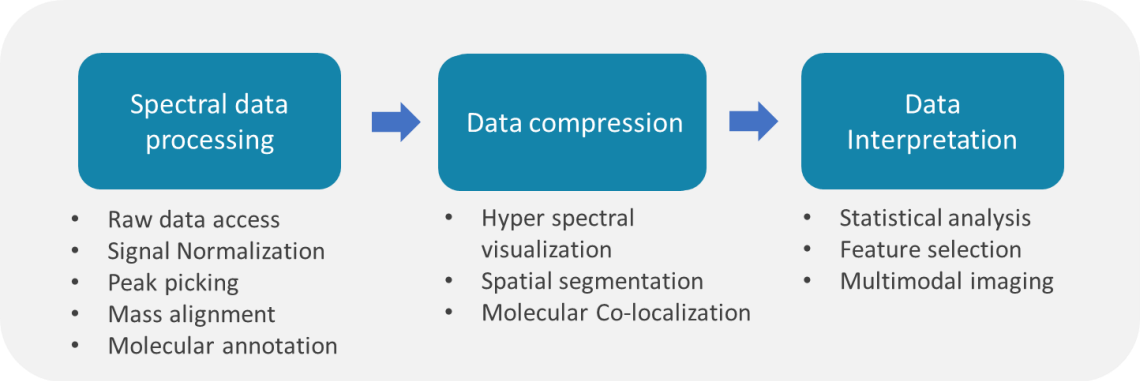


Figure 1.7. *Main data processing flow for mass spectrometry imaging.*

Spectral processing focuses primarily on the mass spectra themselves, ignoring spatial relationships at this stage. It usually begins by accessing each spectrum in the raw data file and proceeds through a series of steps: peak detection, selection of peaks associated with the sample rather than background, calibration and/or alignment of peak positions, signal normalization, and finally molecular formula assignment for detected peaks. At each step, the operator must choose suitable parameters that retain as many valid signals as possible while limiting the inclusion of noise or background peaks. Visual inspection of ion images can help refine peak selection by assessing whether a given signal exhibits a plausible spatial pattern linked to regions of interest.

After spectral processing, an MSI dataset is represented as a data cube (x, y, *m/z*, intensity) containing large numbers of peaks, often well over 1,000 per spectrum. Data compression aims to reduce this high dimensionality and complexity, making the data more tractable for visual exploration and downstream machine learning, thereby mitigating the "curse of dimensionality". In this context, "data compression" encompasses methods that reduce the number of variables by grouping peaks with similar spatial distributions and/or grouping spectra with similar molecular compositions.

Data compression methods can be used in both unsupervised and supervised settings[57,58]. Unsupervised approaches are often applied for exploratory analysis, visualization, and the identification of regions of interest by clustering pixels with

similar molecular content. Transform-based methods, such as principal component analysis (PCA) or related techniques, describe the set of detected peaks in terms of a limited number of latent components. Mapping these components back onto the sample surface yields heat maps that highlight regions sharing similar ion-expression patterns. Co-localization methods instead cluster peaks according to the similarity of their ion images, focusing on spatial correlations between ions. These methods help interpret the relationships between molecules and their spatial distributions and can also reveal how the localization of the same ion varies across different samples[59,60].

In contrast, supervised methods require training on annotated data[61]. For spatial segmentation, models are first trained on MSI pixels that have been assigned to a defined set of classes (e.g. different tissue types). After training, the model can recognize spectral signatures characteristic of each class and assign a class label to new pixels, thereby producing segmentation maps for entire MSI datasets. This capability is particularly useful in digital pathology: models trained on MSI of tissues containing healthy and pathological regions can be applied to new, unlabeled tissues to predict those regions[57,62], Figure 1.8. Supervised methods are also used at the ion-image level; for example, classification models can distinguish “off-sample” ion images (originating from matrix, background, or artifacts) from on-tissue ion images, enabling automated pre-filtering before annotation and downstream analysis[63].

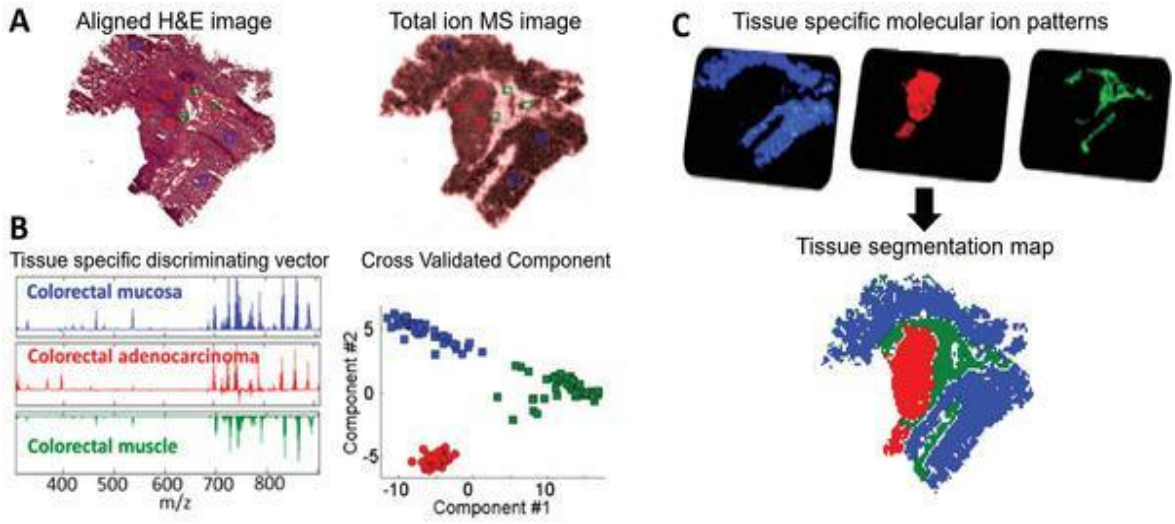


Figure 1.8. *Supervised MS multivariate analysis and classification for colorectal cancer tissues. A) Spatially aligned H&E image and total ion MS image. The locations where mass spectra were selected for RMMC modeling are marked with colors. B) RMMC multivariate analysis reveals tissue-specific discriminating vectors and generates components for efficient transformation of MSI data. C) In the spatial domain, RMMC analysis provides tissue-specific molecular ion patterns. Tissue class annotation is also obtained by k-nearest-neighbor classification of RMMC features*[62]*.*

Finally, data interpretation focuses on linking ions and spatial patterns to underlying biological mechanisms. Most MSI experiments that aim at biological insight compare at least two conditions (e.g., different phenotypes or treatments). Statistical univariate and multivariate analyses, together with machine-learning approaches, are used to identify ions whose detection levels or spatial distributions differ between conditions. However, isolating pixels that represent a single, well-defined phenotype is often difficult. For instance, in tumor imaging, a single biopsy may contain both healthy and cancerous regions within the same section, and each region can itself comprise multiple cell subtypes. This intrinsic heterogeneity complicates spatial segmentation and biological interpretation.

To mitigate these issues, MSI is frequently combined with histology by registering MSI data with corresponding microscopy images of the same sample. High-resolution histological images provide detailed morphological information and enable annotation of tissue structures and cell types based on features that are not visible, or

only coarsely resolved, in MSI. Once the MSI and microscopy images are co-registered, these annotations can be transferred to the MSI pixels, providing a biological context for interpreting changes in *m/z* detection. This integrated view of molecular and morphological information is a key step towards robust, biologically grounded interpretation of MSI data.

1.4 Objectives and thesis organization

Mass spectrometry imaging (MSI) generates complex datasets that link molecular composition to spatial organization. However, transforming these raw signals into biologically interpretable information remains challenging due to limitations in data preprocessing, mass accuracy, spectral interpretation, and spatial resolution. The overall objective of this thesis is to develop computational strategies that improve the reliability, interpretability, and spatial contextualization of MSI data. By addressing key analytical bottlenecks, this work aims to facilitate the extraction of biologically meaningful molecular patterns from high-dimensional MSI datasets and to support reproducible and scalable analysis workflows.

To achieve this general objective, several complementary objectives were defined, each corresponding to a methodological development presented in the following chapters, Figure 1.9. First, a general workflow is developed to organize MSI data into structured representations and to prioritize biologically relevant ions based on both signal characteristics and spatial organization (Chapter 2). Second, strategies are introduced to correct pixel-to-pixel mass shifts occurring during acquisition, thereby improving mass accuracy and strengthening confidence in downstream molecular annotation (Chapter 3). Third, new approaches are proposed to interpret high-resolution spectral information beyond traditional database matching, using mass-defect–based and graph-based representations to capture relationships among chemically related ions (Chapters 4 and 5). Finally, the spatial dimension of MSI data is addressed through the integration of molecular information with microscopy, enabling

improved interpretation of molecular distributions within their spatial context (Chapter 6).

Together, these developments are organized to progressively address major analytical challenges in MSI data processing, from data structuring and spectral reliability to molecular interpretation and spatial integration. The thesis is therefore structured to guide MSI data from raw acquisition toward biologically meaningful interpretation through successive methodological advances.

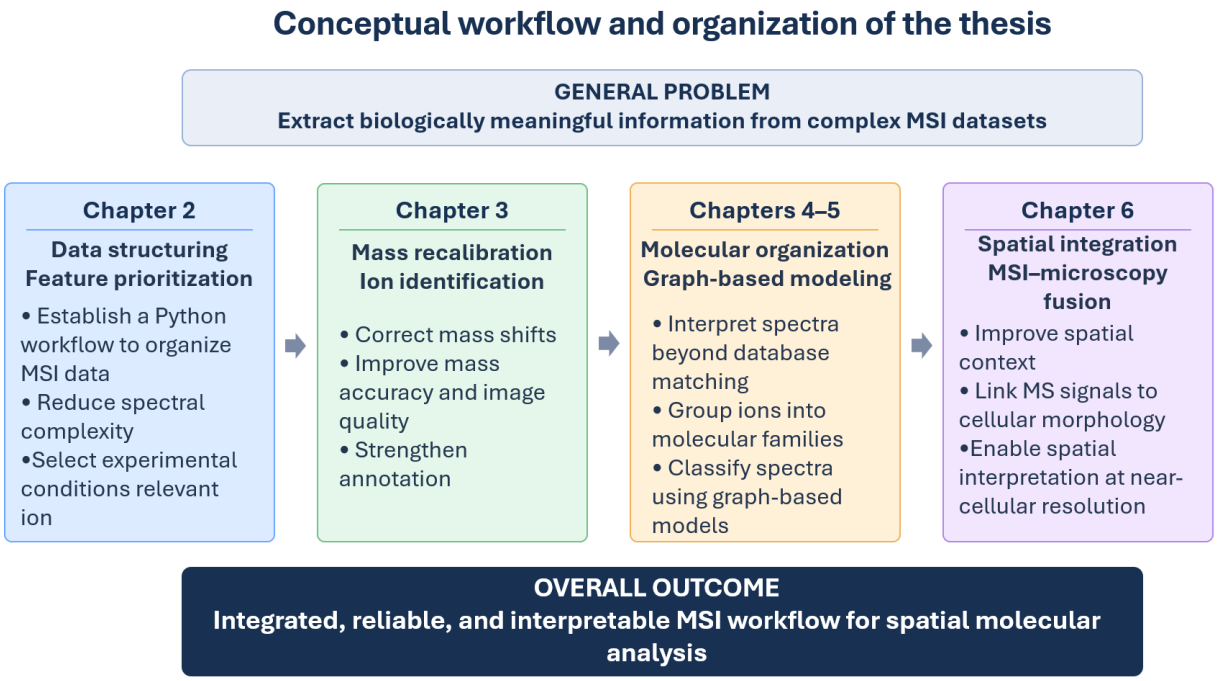


Figure 1.9. *Conceptual organization of the thesis illustrating the sequential methodological developments introduced to address major limitations in mass spectrometry imaging (MSI) data processing. The workflow progresses from data structuring and feature prioritization to spectral stabilization, molecular organization, and spatial integration, ultimately enabling biologically meaningful interpretation of MSI data.*

The thesis is organized as follows. Each chapter proposes a solution to a well-defined limitation of MSI.

Chapter 2

We introduce the development of a general MSI processing workflow for feature prioritization. The chapter describes how raw MSI data can be transformed into a structured representation suitable for systematic comparison across samples and conditions, through data conversion, mean-spectrum construction, peak detection, and ion-image generation. It then addresses one of the central challenges of MSI analysis: reducing a large, complex set of detected *m/z* values into a smaller subset of ions most relevant to the biological question under study. To this end, the chapter presents Image Structure Filtering, which exploits condition-specific signal detection and the spatial organization of ion images. This general workflow is subsequently applied to a case study in natural product discovery, illustrating its practical value for identifying ions associated with specific bacterial growth conditions. Beyond this application, the chapter also establishes the conceptual basis of the *deltamsi* package, which generalizes these processing steps into a reusable computational framework that is extended in the later chapters of the thesis.

Chapter 3

We discuss the mass shift phenomenon in mass spectrometry imaging and its impact on data interpretation. Molecular peak assignments in MSI are generally based on information directly obtained from the mass spectrum (exact mass, localization), often without additional MS/MS information. To attribute molecular formulas to MS peaks with a minimum of false positives, high mass accuracy must be maintained during the acquisition of the entire image and over the full *m/z* range. However, depending on the instrument and sample heterogeneity, small fluctuations in detected *m/z* values and mass shifts typically occur. This chapter proposes a post-acquisition, data-adaptive recalibration methodology to reduce mass shifts in MSI data. The approach relies on

the robust and automated selection of suitable calibrant signals in each pixel independently. We demonstrate the impact of recalibration on mass shifts by comparing ion images before and after automated internal recalibration and by evaluating improvements in molecular identification.

Chapter 4

We build on unsupervised methods and data compression and investigate how mass, mass defect, and mass patterns can be exploited to reorganize and compress MSI data. We first revisit the Kendrick mass defect (KMD) analysis as a way to reorder complex spectra and rapidly visualize chemically related compounds without relying solely on database *m/z* matching. Using KMD plots and the MSKF tool, we show how matrix and background signals can be filtered out and how molecular families such as lipids or lipopeptides can be highlighted in microbial and tissue MSI datasets. We then move beyond one-dimensional KMD plots and introduce a new representation of a spectrum as a graph of mass differences, in which peaks are connected if their *m/z* difference matches a list of chemically meaningful mass shifts and, in the MSI context, if their ion images are sufficiently similar. We then use this graph to cluster thousands of peaks into a limited number of molecular families, acting as a powerful data-compression layer that denoises and structures the spectral space. This graph-based view naturally prepares the ground for subsequent machine-learning models, in particular graph neural networks, to perform robust and interpretable spectrum and image classification.

Chapter 5

We build directly on the graph-based representations of high-resolution mass spectra introduced in Chapter 4 and extend them into a supervised deep learning framework for MSI classification and interpretation. While Chapter 4 uses graphs of mass differences and unsupervised clustering to compress spectra into molecular families, Chapter 5 investigates how a similar representation can be exploited by a Relational

Graph Convolutional Network (R-GCN) to perform supervised data compression tailored to classification tasks.

Each high-resolution spectrum is modeled as a graph whose nodes are centroid peaks and whose edges encode chemically meaningful mass differences (adducts, isotopes, neutral losses, transformations) derived from the same mass-pattern logic as in Chapter 4. The R-GCN aggregates information over this graph to produce a compact embedding that serves as input to a classifier, allowing the model to use both intensity and mass-relationship information rather than intensities alone. We benchmark this approach against standard machine-learning and 1D CNN baselines, focusing on robustness to realistic MSI artifacts such as mass shifts (Chapter 3), peak loss, and signal mixing, without relying on dataset-specific pre-processing. Finally, we show how class activation mapping can be projected back onto graph nodes to identify informative *m/z* values and molecular families.

Chapter 5 thus closes the thesis, our investigation of MSI pre-processing and of spectral representations for classification and interpretation. The spatial dimension of MSI is less extensively explored in this thesis, but it is nonetheless addressed in an annex chapter, Chapter 6, where we briefly investigate how deep learning can be used to link microscopic features to ionic spatial distributions.

Chapter 6

The spatial resolution of MALDI-MSI is often limited by sensitivity, file size, and spatial delocalization. As a result, spatial features that are clearly visible in routine microscope images are not always observed with the same clarity in MSI. To address this, microscope images of the same sample are associated with their corresponding MSI data to explore how morphological cell features relate to mass signal expression. We assess the potential of a convolutional U-Net architecture to automate the fusion of mass and morphological information. Image fusion methods are often demonstrated on complex tissue images composed of many cell types, which complicates the

interpretation of fusion results. Here, we instead evaluate our fusion strategy on a simpler dataset composed of a mixture of two cell types (MCF-7/WI-38). The model is trained and evaluated on a co-culture where the two cell types are mixed and cannot be distinguished at the MSI spatial resolution, and then tested on monocultures of each cell type. We show that the model can correctly predict different phospholipids detected by MALDI-MSI and segment cells at the microscope image resolution.

1.5 Recent advances and the specific gaps addressed by this thesis

This thesis was carried out between 2018 and 2022, and several of its components were consolidated more recently, namely the graph-based classifier published in 2025 and the open-source *deltamsi* package released in 2026, which compiles different pipelines developed during the thesis. Over the same interval, the field has advanced rapidly, driven by deep learning, foundation models for mass spectra, the push toward single-cell resolution, and a proliferation of MSI software. This section, therefore, situates the work against more recent literature, focusing only on the problems it addresses and the specific gap it fills in each case.

1.5.1 Software, preprocessing, and annotation

Beyond the individual processing and annotation steps already described (Section 1.3.2), the field has moved towards reproducible, end-to-end and modality-aware software, from Cardinal[64], MZmine[65] and the FDR-controlled METASPACE annotation engine[46,66] to msiFlow, an open, vendor-neutral pipeline spanning raw multimodal MSI-microscopy data to visualization[67]. Two limitations remain directly relevant here: most pipelines still offer little support for reducing thousands of *m/z* to the few ions relevant to a given comparison, and MS1-only annotation stays acutely sensitive to mass accuracy, so that uncorrected pixel-to-pixel mass shifts translate into missed or

erroneous identifications. The thesis targets both. Chapter 2 contributes a spatially-informed feature-prioritization workflow, consolidated as *deltamsi*, and Chapter 3 a data-adaptive per-pixel recalibration that measurably increases the number of confident annotations[68].

1.5.2 Spectrum representation, foundation models and graph-based classification

High-resolution mass information is key to interpreting a mass signal: the mass defect of a peak and the recurrent mass differences between peaks help estimate the molecular composition present in a spectrum, and can even point to the molecular family of a given ion. Capturing and representing this information is therefore important both for manual interpretation and for the representation given to a machine learning model, whereas collapsing a spectrum into a fixed intensity vector, as is common, discards much of it.

The particularity of this work is that it focuses on MS1 information from imaging and uses a graph representation to capture it. Chapter 4 builds a graph of the spectrum, in which peaks are the nodes and edges encode both a chemically meaningful mass difference and, specifically for imaging, spatial co-localization of the corresponding ion images, thereby aiding manual interpretation of the mass signal. In Chapter 5, we developed a Relational Graph Convolutional Network that learns from this graph to derive, from high-resolution mass and mass-difference information alone, both a representation of the whole spectrum and a representation of each peak it contains.

Two more recent lines of work each capture part of what such a model needs, but not both. The first brings graphs to MSI segmentation: SagMSI[69] and GraphMSI[70] classify pixels with graph neural networks, but their graphs are built over the image, with nodes as pixels and edges connecting spatial neighbors, so they encode spatial proximity

rather than the mass relationships within a spectrum. (Graph neural networks are used elsewhere in mass spectrometry, but on the molecular structure graph in order to predict a spectrum, as in MassFormer[71], again leaving the measured spectrum's mass features unexploited.) The second line does exploit high-resolution mass features and learns rich spectral representations through deep learning, DreaMS[72], for instance, is a transformer pre-trained in a self-supervised way on millions of MS/MS spectra by masking peaks and learning to predict their *m/z*; it represents each peak's *m/z* with high-resolution Fourier features that preserve fine mass detail, and injects the pairwise mass differences between peaks directly into its self-attention, so that neutral losses and mass relationships are learned from the data rather than hard-coded.

1.5.3 Single-cell analysis, super-resolution and microscopy integration

The single-cell strategies introduced in Section 1.3.1 have advanced along two lines. Instrumentally, MALDI-2 and transmission-mode geometries now reach pixel sizes of one micrometer and below for genuine single-cell and sub-cellular imaging, often coupled to in-source microscopy[73,74]. Computationally, where that resolution is not reached, cellular information is recovered by integrating MSI with higher-resolution modalities on a single section, such as imaging mass cytometry[75], spatial transcriptomics[76] or single-cell immunofluorescence[76,77], extending the MSI-microscopy principle of SpaceM[50]. Chapter 6 follows this computational route, using a U-Net to link the morphology of a microscopy image to the molecular signal of MSI on a controlled two-cell-type system and to predict cell-type-specific lipids at microscope resolution. Far from being superseded, this is precisely the strategy the field is now generalizing: GHIST, for example, predicts spatial gene expression at single-cell resolution directly from routine histology images by deep learning[78], the same morphology-to-molecule inference that Chapter 6 performs from microscopy to lipids. The approach, therefore,

remains a relevant, low-cost route to enriching a morphological image with a spatially resolved molecular layer.

REFERENCES:

(1) Yang, L. G.; March, Z. M.; Stephenson, R. A.; Narayan, P. S. Apolipoprotein E in Lipid Metabolism and Neurodegenerative Disease. *Trends Endocrinol. Metab.* **2023**, *34* (8), 430–445. https://doi.org/10.1016/j.tem.2023.05.002.

(2) Planque, M.; Igelmann, S.; Ferreira Campos, A. M.; Fendt, S.-M. Spatial Metabolomics Principles and Application to Cancer Research. *Curr. Opin. Chem. Biol.* **2023**, *76*, 102362. https://doi.org/10.1016/j.cbpa.2023.102362.

(3) Buchberger, A. R.; DeLaney, K.; Johnson, J.; Li, L. Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights. *Anal. Chem.* **2018**, *90* (1), 240–265. https://doi.org/10.1021/acs.analchem.7b04733.

(4) Gross, J. H. *Mass Spectrometry*; Springer International Publishing: Cham, 2017. https://doi.org/10.1007/978-3-319-54398-7.

(5) Gilmore, I. S.; Heiles, S.; Pieterse, C. L. Metabolic Imaging at the Single-Cell Scale: Recent Advances in Mass Spectrometry Imaging. *Annu. Rev. Anal. Chem.* **2019**, *12* (1), 201–224. https://doi.org/10.1146/annurev-anchem-061318-115516.

(6) Takáts, Z.; Wiseman, J. M.; Gologan, B.; Cooks, R. G. Mass Spectrometry Sampling Under Ambient Conditions with Desorption Electrospray Ionization. *Science* **2004**, *306* (5695), 471–473. https://doi.org/10.1126/science.1104404.

(7) Caprioli, R. M.; Farmer, T. B.; Gile, J. Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. *Anal. Chem.* **1997**, *69* (23), 4751–4760. https://doi.org/10.1021/ac970888i.

(8) Svatoš, A. Mass Spectrometric Imaging of Small Molecules. *Trends Biotechnol.* **2010**, *28* (8), 425–434. https://doi.org/10.1016/j.tibtech.2010.05.005.

(9) Zhu, X.; Xu, T.; Peng, C.; Wu, S. Advances in MALDI Mass Spectrometry Imaging Single Cell and Tissues. *Front. Chem.* **2022**, *9*, 782432. https://doi.org/10.3389/fchem.2021.782432.

(10) Zhang, H.; Lu, K. H.; Ebbini, M.; Huang, P.; Lu, H.; Li, L. Mass Spectrometry Imaging for Spatially Resolved Multi-Omics Molecular Mapping. *Npj Imaging* **2024**, *2* (1), 20. https://doi.org/10.1038/s44303-024-00025-3.

(11) Xing, L.; Zhao, C.-L.; Mou, H.-Z.; Pan, J.; Kang, B.; Chen, H.-Y.; Xu, J.-J. Next Generation of Mass Spectrometry Imaging: From Micrometer to Subcellular Resolution. *Chem. Biomed. Imaging* **2023**, *1* (8), 670–682. https://doi.org/10.1021/cbmi.3c00061.

(12) Wiseman, J. M.; Ifa, D. R.; Song, Q.; Cooks, R. G. Tissue Imaging at Atmospheric Pressure Using Desorption Electrospray Ionization (DESI) Mass Spectrometry. *Angew. Chem. Int. Ed.* **2006**, *45* (43), 7188–7192. https://doi.org/10.1002/anie.200602449.

(13) Roach, P. J.; Laskin, J.; Laskin, A. Nanospray Desorption Electrospray Ionization: An Ambient Method for Liquid-Extraction Surface Sampling in Mass Spectrometry. *The Analyst* **2010**, *135* (9), 2233. https://doi.org/10.1039/c0an00312c.

(14) Tillner, J.; Wu, V.; Jones, E. A.; Pringle, S. D.; Karancsi, T.; Dannhorn, A.; Veselkov, K.; McKenzie, J. S.; Takats, Z. Faster, More Reproducible DESI-MS for Biological Tissue Imaging. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (10), 2090–2098. https://doi.org/10.1007/s13361-017-1714-z.

(15) Karas, M.; Bachmann, D.; Bahr, U.; Hillenkamp, F. Matrix-Assisted Ultraviolet Laser Desorption of Non-Volatile Compounds. *Int. J. Mass Spectrom. Ion Process.* **1987**, *78*, 53–68. https://doi.org/10.1016/0168-1176(87)87041-6.

(16) Tanaka, K.; Waki, H.; Ido, Y.; Akita, S.; Yoshida, Y.; Yoshida, T.; Matsuo, T. Protein and Polymer Analyses up to *m/z* 100 000 by Laser Ionization Time-of-flight Mass Spectrometry.

Rapid Commun. Mass Spectrom. **1988**, *2* (8), 151–153. https://doi.org/10.1002/rcm.1290020802.
(17) Spengler, B.; Hubert, M.; Kaufmann, R. MALDI Ion Imaging and Biological Ion Imaging with a New Scanning UV-Laser Microprobe. *42nd ASMS Conference on Mass Spectrometry and Allied Topics*, 1994.
(18) Dreisewerd, K. Recent Methodological Advances in MALDI Mass Spectrometry. *Anal. Bioanal. Chem.* **2014**, *406* (9–10), 2261–2278. https://doi.org/10.1007/s00216-014-7646-6.
(19) *MALDI MS: A Practical Guide to Instrumentation, Methods and Applications*, 1st ed.; Hillenkamp, F., Peter-Katalinić, J., Eds.; Wiley, 2007. https://doi.org/10.1002/9783527610464.
(20) Körber, A.; Anthony, I. G. M.; Heeren, R. M. A. Mass Spectrometry Imaging. *Anal. Chem.* **2025**, *97* (29), 15517–15549. https://doi.org/10.1021/acs.analchem.4c05249.
(21) Goodwin, R. J. A. Sample Preparation for Mass Spectrometry Imaging: Small Mistakes Can Lead to Big Consequences. *J. Proteomics* **2012**, *75* (16), 4893–4911. https://doi.org/10.1016/j.jprot.2012.04.012.
(22) Seeley, E. H.; Oppenheimer, S. R.; Mi, D.; Chaurand, P.; Caprioli, R. M. Enhancement of Protein Sensitivity for MALDI Imaging Mass Spectrometry after Chemical Treatment of Tissue Sections. *J. Am. Soc. Mass Spectrom.* **2008**, *19* (8), 1069–1077. https://doi.org/10.1016/j.jasms.2008.03.016.
(23) Groseclose, M. R.; Andersson, M.; Hardesty, W. M.; Caprioli, R. M. Identification of Proteins Directly from Tissue: *In Situ* Tryptic Digestions Coupled with Imaging Mass Spectrometry. *J. Mass Spectrom.* **2007**, *42* (2), 254–262. https://doi.org/10.1002/jms.1177.
(24) Esteve, C.; Tolner, E. A.; Shyti, R.; Van Den Maagdenberg, A. M. J. M.; McDonnell, L. A. Mass Spectrometry Imaging of Amino Neurotransmitters: A Comparison of Derivatization Methods and Application in Mouse Brain Tissue. *Metabolomics* **2016**, *12* (2), 30. https://doi.org/10.1007/s11306-015-0926-0.
(25) Dreisewerd, K. The Desorption Process in MALDI. *Chem. Rev.* **2003**, *103* (2), 395–426. https://doi.org/10.1021/cr010375i.
(26) Karas, M.; Gl□ckmann, M.; Sch□fer, J. Ionization in Matrix-Assisted Laser Desorption/Ionization: Singly Charged Molecular Ions Are the Lucky Survivors. *J. Mass Spectrom.* **2000**, *35* (1), 1–12. https://doi.org/10.1002/(SICI)1096-9888(200001)35:1%3C1::AID-JMS904%3E3.0.CO;2-0.
(27) Jaskolla, T. W.; Karas, M. Compelling Evidence for Lucky Survivor and Gas Phase Protonation: The Unified MALDI Analyte Protonation Mechanism. *J. Am. Soc. Mass Spectrom.* **2011**, *22* (6), 976–988. https://doi.org/10.1007/s13361-011-0093-0.
(28) Ho, Y.; Shu, L.; Yang, Y. Imaging Mass Spectrometry for Metabolites: Technical Progress, Multimodal Imaging, and Biological Interactions. *WIREs Syst. Biol. Med.* **2017**, *9* (5), e1387. https://doi.org/10.1002/wsbm.1387.
(29) Gemperline, E.; Rawson, S.; Li, L. Optimization and Comparison of Multiple MALDI Matrix Application Methods for Small Molecule Mass Spectrometric Imaging. *Anal. Chem.* **2014**, *86* (20), 10030–10035. https://doi.org/10.1021/ac5028534.
(30) Shimma, S.; Sugiura, Y. Effective Sample Preparations in Imaging Mass Spectrometry. *Mass Spectrom.* **2014**, *3* (Special_Issue), S0029–S0029. https://doi.org/10.5702/massspectrometry.S0029.
(31) Tsai, Y.-H.; Menger, R. F.; Drexler, D. M.; Yost, R. A.; Garrett, T. J. Ionization Sources and Mass Analyzers in MS Imaging. *Bioanalysis* **2015**, *7* (20), 2629–2637. https://doi.org/10.4155/bio.15.187.

(32) Norris, J. L.; Caprioli, R. M. Analysis of Tissue Specimens by Matrix-Assisted Laser Desorption/Ionization Imaging Mass Spectrometry in Biological and Clinical Research. *Chem. Rev.* **2013**, *113* (4), 2309–2342. https://doi.org/10.1021/cr3004295.
(33) Short, R. T.; Todd, P. J. Improved Energy Compensation for Time-of-Flight Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (8), 779–787. https://doi.org/10.1016/1044-0305(94)80011-1.
(34) Alexandrov, T. MALDI Imaging Mass Spectrometry: Statistical Data Analysis and Current Computational Challenges. *BMC Bioinformatics* **2012**, *13* (S16), S11. https://doi.org/10.1186/1471-2105-13-S16-S11.
(35) Murray, K. K. Resolution and Resolving Power in Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2022**, *33* (12), 2342–2347. https://doi.org/10.1021/jasms.2c00216.
(36) Brenton, A. G.; Godfrey, A. R. Accurate Mass Measurement: Terminology and Treatment of Data. *J. Am. Soc. Mass Spectrom.* **2010**, *21* (11), 1821–1835. https://doi.org/10.1016/j.jasms.2010.06.006.
(37) Kind, T.; Fiehn, O. Seven Golden Rules for Heuristic Filtering of Molecular Formulas Obtained by Accurate Mass Spectrometry. *BMC Bioinformatics* **2007**, *8* (1), 105. https://doi.org/10.1186/1471-2105-8-105.
(38) Dunham, S. J. B.; Ellis, J. F.; Li, B.; Sweedler, J. V. Mass Spectrometry Imaging of Complex Microbial Communities. *Acc. Chem. Res.* **2017**, *50* (1), 96–104. https://doi.org/10.1021/acs.accounts.6b00503.
(39) Ščupáková, K.; Balluff, B.; Tressler, C.; Adelaja, T.; Heeren, R. M. A.; Glunde, K.; Ertaylan, G. Cellular Resolution in Clinical MALDI Mass Spectrometry Imaging: The Latest Advancements and Current Challenges. *Clin. Chem. Lab. Med. CCLM* **2020**, *58* (6), 914–929. https://doi.org/10.1515/cclm-2019-0858.
(40) Škrášková, K.; Khmelinskii, A.; Abdelmoula, W. M.; De Munter, S.; Baes, M.; McDonnell, L.; Dijkstra, J.; Heeren, R. M. A. Precise Anatomic Localization of Accumulated Lipids in ***Mfp2*** Deficient Murine Brains Through Automated Registration of SIMS Images to the Allen Brain Atlas. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (6), 948–957. https://doi.org/10.1007/s13361-015-1146-6.
(41) Feucherolles, M.; Frache, G. MALDI Mass Spectrometry Imaging: A Potential Game-Changer in a Modern Microbiology. *Cells* **2022**, *11* (23), 3900. https://doi.org/10.3390/cells11233900.
(42) Thomen, A.; Najafinobar, N.; Penen, F.; Kay, E.; Upadhyay, P. P.; Li, X.; Phan, N. T. N.; Malmberg, P.; Klarqvist, M.; Andersson, S.; Kurczy, M. E.; Ewing, A. G. Subcellular Mass Spectrometry Imaging and Absolute Quantitative Analysis across Organelles. *ACS Nano* **2020**, *14* (4), 4316–4325. https://doi.org/10.1021/acsnano.9b09804.
(43) Spraggins, J. M.; Rizzo, D. G.; Moore, J. L.; Noto, M. J.; Skaar, E. P.; Caprioli, R. M. Next-generation Technologies for Spatial Proteomics: Integrating Ultra-high Speed MALDI-TOF and High Mass Resolution MALDI FTICR Imaging Mass Spectrometry for Protein Analysis. *PROTEOMICS* **2016**, *16* (11–12), 1678–1689. https://doi.org/10.1002/pmic.201600003.
(44) Knochenmuss, R.; Dubois, F.; Dale, M. J.; Zenobi, R. The Matrix Suppression Effect and Ionization Mechanisms in Matrix-Assisted Laser Desorption/Ionization. *Rapid Commun. Mass Spectrom.* **1996**, *10* (8), 871–877. https://doi.org/10.1002/(SICI)1097-0231(19960610)10:8%3C871::AID-RCM559%3E3.0.CO;2-R.
(45) Tressler, C.; Tilley, S.; Yang, E.; Donohue, C.; Barton, E.; Creissen, A.; Glunde, K. Factorial Design to Optimize Matrix Spraying Parameters for MALDI Mass Spectrometry Imaging. *J. Am. Soc. Mass Spectrom.* **2021**, *32* (12), 2728–2737. https://doi.org/10.1021/jasms.1c00081.

(46) Nguyen, D. D.; Saharuka, V.; Kovalev, V.; Stuart, L.; Del Prete, M.; Lubowiecka, K.; De Mot, R.; Venturi, V.; Alexandrov, T. Facilitating Imaging Mass Spectrometry of Microbial Specialized Metabolites with METASPACE. *Metabolites* **2021**, *11* (8), 477. https://doi.org/10.3390/metabo11080477.
(47) Spraker, J. E.; Luu, G. T.; Sanchez, L. M. Imaging Mass Spectrometry for Natural Products Discovery: A Review of Ionization Methods. *Nat. Prod. Rep.* **2020**, *37* (2), 150–162. https://doi.org/10.1039/C9NP00038K.
(48) Stasulli, N. M.; Shank, E. A. Profiling the Metabolic Signals Involved in Chemical Communication between Microbes Using Imaging Mass Spectrometry. *FEMS Microbiol. Rev.* **2016**, *40* (6), 807–813. https://doi.org/10.1093/femsre/fuw032.
(49) Marioni, J. C.; Arendt, D. How Single-Cell Genomics Is Changing Evolutionary and Developmental Biology. *Annu. Rev. Cell Dev. Biol.* **2017**, *33* (1), 537–553. https://doi.org/10.1146/annurev-cellbio-100616-060818.
(50) Rappez, L.; Stadler, M.; Triana, S.; Gathungu, R. M.; Ovchinnikova, K.; Phapale, P.; Heikenwalder, M.; Alexandrov, T. SpaceM Reveals Metabolic States of Single Cells. *Nat. Methods* **2021**, *18* (7), 799–805. https://doi.org/10.1038/s41592-021-01198-0.
(51) Alberts, D.; Pottier, C.; Smargiasso, N.; Baiwir, D.; Mazzucchelli, G.; Delvenne, P.; Kriegsmann, M.; Kazdal, D.; Warth, A.; De Pauw, E.; Longuespée, R. MALDI Imaging-Guided Microproteomic Analyses of Heterogeneous Breast Tumors—A Pilot Study. *PROTEOMICS – Clin. Appl.* **2018**, *12* (1), 1700062. https://doi.org/10.1002/prca.201700062.
(52) Dilillo, M.; Pellegrini, D.; Ait-Belkacem, R.; De Graaf, E. L.; Caleo, M.; McDonnell, L. A. Mass Spectrometry Imaging, Laser Capture Microdissection, and LC-MS/MS of the Same Tissue Section. *J. Proteome Res.* **2017**, *16* (8), 2993–3001. https://doi.org/10.1021/acs.jproteome.7b00284.
(53) Alexandrov, T. Spatial Metabolomics and Imaging Mass Spectrometry in the Age of Artificial Intelligence. *Annu. Rev. Biomed. Data Sci.* **2020**, *3* (1), 61–87. https://doi.org/10.1146/annurev-biodatasci-011420-031537.
(54) Vaysse, P.-M.; Heeren, R. M. A.; Porta, T.; Balluff, B. Mass Spectrometry Imaging for Clinical Research – Latest Developments, Applications, and Current Limitations. *The Analyst* **2017**, *142* (15), 2690–2712. https://doi.org/10.1039/C7AN00565B.
(55) Ogrinc, N.; Kruszewski, A.; Chaillou, P.; Saudemont, P.; Lagadec, C.; Salzet, M.; Duriez, C.; Fournier, I. Robot-Assisted SpiderMass for *In Vivo* Real-Time Topography Mass Spectrometry Imaging. *Anal. Chem.* **2021**, *93* (43), 14383–14391. https://doi.org/10.1021/acs.analchem.1c01692.
(56) Hu, H.; Helminiak, D.; Yang, M.; Unsihuay, D.; Hilger, R. T.; Ye, D. H.; Laskin, J. High-Throughput Mass Spectrometry Imaging with Dynamic Sparse Sampling. *ACS Meas. Sci. Au* **2022**, *2* (5), 466–474. https://doi.org/10.1021/acsmeasuresciau.2c00031.
(57) Hu, H.; Laskin, J. Emerging Computational Methods in Mass Spectrometry Imaging. *Adv. Sci.* **2022**, *9* (34), 2203339. https://doi.org/10.1002/advs.202203339.
(58) Verbeeck, N.; Caprioli, R. M.; Van de Plas, R. Unsupervised Machine Learning for Exploratory Data Analysis in Imaging Mass Spectrometry. *Mass Spectrom. Rev.* **2020**, *39* (3), 245–291. https://doi.org/10.1002/mas.21602.
(59) Inglese, P.; Correia, G.; Pruski, P.; Glen, R. C.; Takats, Z. Colocalization Features for Classification of Tumors Using Desorption Electrospray Ionization Mass Spectrometry Imaging. *Anal. Chem.* **2019**, *91* (10), 6530–6540. https://doi.org/10.1021/acs.analchem.8b05598.

(60) McDonnell, L. A.; Van Remoortere, A.; Van Zeijl, R. J. M.; Deelder, A. M. Mass Spectrometry Image Correlation: Quantifying Colocalization. *J. Proteome Res.* **2008**, *7* (8), 3619–3627. https://doi.org/10.1021/pr800214d.
(61) Bemis, K. D.; Harry, A.; Eberlin, L. S.; Ferreira, C. R.; Van De Ven, S. M.; Mallick, P.; Stolowitz, M.; Vitek, O. Probabilistic Segmentation of Mass Spectrometry (MS) Images Helps Select Important Ions and Characterize Confidence in the Resulting Segments. *Mol. Cell. Proteomics* **2016**, *15* (5), 1761–1772. https://doi.org/10.1074/mcp.O115.053918.
(62) Veselkov, K. A.; Mirnezami, R.; Strittmatter, N.; Goldin, R. D.; Kinross, J.; Speller, A. V. M.; Abramov, T.; Jones, E. A.; Darzi, A.; Holmes, E.; Nicholson, J. K.; Takats, Z. Chemo-Informatic Strategy for Imaging Mass Spectrometry-Based Hyperspectral Profiling of Lipid Signatures in Colorectal Cancer. *Proc. Natl. Acad. Sci.* **2014**, *111* (3), 1216–1221. https://doi.org/10.1073/pnas.1310524111.
(63) Ovchinnikova, K.; Kovalev, V.; Stuart, L.; Alexandrov, T. OffsampleAI: Artificial Intelligence Approach to Recognize off-Sample Mass Spectrometry Images. *BMC Bioinformatics* **2020**, *21* (1), 129. https://doi.org/10.1186/s12859-020-3425-x.
(64) Bemis, K. A.; Föll, M. C.; Guo, D.; Lakkimsetty, S. S.; Vitek, O. Cardinal v.3: A Versatile Open-Source Software for Mass Spectrometry Imaging Analysis. *Nat. Methods* **2023**, *20* (12), 1883–1886. https://doi.org/10.1038/s41592-023-02070-z.
(65) Schmid, R.; Heuckeroth, S.; Korf, A.; Smirnov, A.; Myers, O.; Dyrlund, T. S.; Bushuiev, R.; Murray, K. J.; Hoffmann, N.; Lu, M.; Sarvepalli, A.; Zhang, Z.; Fleischauer, M.; Dührkop, K.; Wesner, M.; Hoogstra, S. J.; Rudt, E.; Mokshyna, O.; Brungs, C.; Ponomarov, K.; Mutabdžija, L.; Damiani, T.; Pudney, C. J.; Earll, M.; Helmer, P. O.; Fallon, T. R.; Schulze, T.; Rivas-Ubach, A.; Bilbao, A.; Richter, H.; Nothias, L.-F.; Wang, M.; Orešič, M.; Weng, J.-K.; Böcker, S.; Jeibmann, A.; Hayen, H.; Karst, U.; Dorrestein, P. C.; Petras, D.; Du, X.; Pluskal, T. Integrative Analysis of Multimodal Mass Spectrometry Data in MZmine 3. *Nat. Biotechnol.* **2023**, *41* (4), 447–449. https://doi.org/10.1038/s41587-023-01690-2.
(66) Palmer, A.; Phapale, P.; Chernyavsky, I.; Lavigne, R.; Fay, D.; Tarasov, A.; Kovalev, V.; Fuchser, J.; Nikolenko, S.; Pineau, C.; Becker, M.; Alexandrov, T. FDR-Controlled Metabolite Annotation for High-Resolution Imaging Mass Spectrometry. *Nat. Methods* **2017**, *14* (1), 57–60. https://doi.org/10.1038/nmeth.4072.
(67) Spangenberg, P.; Bessler, S.; Widera, L.; Bottek, J.; Richter, M.; Thiebes, S.; Siemes, D.; Krauß, S. D.; Migas, L. G.; Kasarla, S. S.; Phapale, P.; Kleesiek, J.; Führer, D.; Moeller, L. C.; Heuer, H.; Van De Plas, R.; Gunzer, M.; Soehnlein, O.; Soltwisch, J.; Shevchuk, O.; Dreisewerd, K.; Engel, D. R. msiFlow: Automated Workflows for Reproducible and Scalable Multimodal Mass Spectrometry Imaging and Microscopy Data Analysis. *Nat. Commun.* **2025**, *16* (1), 1065. https://doi.org/10.1038/s41467-024-55306-7.
(68) La Rocca, R.; Kune, C.; Tiquet, M.; Stuart, L.; Eppe, G.; Alexandrov, T.; De Pauw, E.; Quinton, L. Adaptive Pixel Mass Recalibration for Mass Spectrometry Imaging Based on Locally Endogenous Biological Signals. *Anal. Chem.* **2021**, *93* (8), 4066–4074. https://doi.org/10.1021/acs.analchem.0c05071.
(69) Shah, M.; Wang, L.; Guo, L.; Xie, C.; Lam, T. K.-Y.; Deng, L.; Xu, X.; Xu, J.; Dong, J.; Cai, Z. SagMSI: A Graph Convolutional Network Framework for Precise Spatial Segmentation in Mass Spectrometry Imaging. *Anal. Chim. Acta* **2025**, *1358*, 344098. https://doi.org/10.1016/j.aca.2025.344098.
(70) Guo, L.; Xie, P.; Shen, X.; Lam, T. K. Y.; Deng, L.; Xie, C.; Xu, X.; Wong, C. K. C.; Xu, J.; Fang, J.; Wang, X.; Xiong, Z.; Luo, S.; Wang, J.; Dong, J.; Cai, Z. Unraveling Spatial Heterogeneity in Mass Spectrometry Imaging Data with GraphMSI. *Adv. Sci.* **2025**, *12* (8), 2410840. https://doi.org/10.1002/advs.202410840.

(71) Young, A.; Röst, H.; Wang, B. Tandem Mass Spectrum Prediction for Small Molecules Using Graph Transformers. *Nat. Mach. Intell.* **2024**, *6* (4), 404–416. https://doi.org/10.1038/s42256-024-00816-8.
(72) Bushuiev, R.; Bushuiev, A.; Samusevich, R.; Brungs, C.; Sivic, J.; Pluskal, T. Self-Supervised Learning of Molecular Representations from Millions of Tandem Mass Spectra Using DreaMS. *Nat. Biotechnol.* **2026**, *44* (4), 630–640. https://doi.org/10.1038/s41587-025-02663-3.
(73) Potthoff, A.; Schwenzfeier, J.; Niehaus, M.; Bessler, S.; Hoffmann, E.; Soehnlein, O.; Höhndorf, J.; Dreisewerd, K.; Soltwisch, J. Spatial Biology Using Single-Cell Mass Spectrometry Imaging and Integrated Microscopy. *Nat. Commun.* **2025**, *16* (1), 9129. https://doi.org/10.1038/s41467-025-64603-8.
(74) Zhang, H.; Ebbini, M.; Duong, T.; Selby, K. G.; Lu, K. H.; Hubecky, E.; Lu, H.; Li, L. Single-Cell Mass Spectrometry Imaging: Emerging Advancements and Future Insights. *Anal. Chim. Acta* **2026**, *1382*, 344825. https://doi.org/10.1016/j.aca.2025.344825.
(75) Nunes, J. B.; Ijsselsteijn, M. E.; Abdelaal, T.; Ursem, R.; Van Der Ploeg, M.; Giera, M.; Everts, B.; Mahfouz, A.; Heijs, B.; De Miranda, N. F. C. C. Integration of Mass Cytometry and Mass Spectrometry Imaging for Spatially Resolved Single-Cell Metabolic Profiling. *Nat. Methods* **2024**, *21* (10), 1796–1800. https://doi.org/10.1038/s41592-024-02392-6.
(76) Hendriks, T. F. E.; Eijkel, G. B.; Visvikis, T.; Balluff, B.; Heeren, R. M. A.; Cuypers, E. One Section, Two Worlds: Single-Cell Integration of MALDI-MSI and Spatial Transcriptomics on the Same Single Tissue Section. *Sci. Rep.* **2025**, *15* (1), 42660. https://doi.org/10.1038/s41598-025-26735-1.
(77) Bessler, S.; Richter, M.; Schwenzfeier, J.; Reinartz Groba, S. N.; Trebicka, J.; Dreisewerd, K.; Soltwisch, J.; Soehnlein, O. Integrated Single-Cell Mass Spectrometry Imaging and Immunofluorescence Microscopy for Multimodal Characterization of Human Immune Cells. *Anal. Chem.* **2026**, *98* (14), 10338–10344. https://doi.org/10.1021/acs.analchem.6c00381.
(78) Fu, X.; Cao, Y.; Bian, B.; Wang, C.; Graham, D.; Pathmanathan, N.; Patrick, E.; Kim, J.; Yang, J. Y. H. Spatial Gene Expression at Single-Cell Resolution from Histology Using Deep Learning with GHIST. *Nat. Methods* **2025**, *22* (9), 1900–1910. https://doi.org/10.1038/s41592-025-02795-z.

CHAPTER TWO

***

Development of a General MSI Workflow for Feature Prioritization

Chapter 2: Content

2.1 Introduction: Detecting Condition-Specific Ions 51

2.2 MSI processing pipeline 53

2.2.1 Mass Spectrometry Image Handling for Data Processing 56

2.2.2 Average Spectrum and Peak Detection.....59

2.2.3 Generation of Ion Images for Each Peak 62

2.2.4 Scoring ion image structure 67

2.3 Exploiting Differential Signal Filtering (DSF) and Image Structure Filtering (ISF) Methods for Untargeted Mass Spectrometry Imaging of Bacterial Metabolites 72

2.4 Generalization of the workflow in the deltamsi package..... 75

2.4.1 Software for MSI data analysis 75

2.4.2 General description of deltamsi 79

2.5 Conclusion and transition to other chapters..... 81

2.1 Introduction: Detecting Condition-Specific Ions

Mass Spectrometry Imaging (MSI) is a key tool in biological research, with the primary goal of detecting ions associated with specific regions of interest or particular phenotypes. Once the molecular identities of these ions are determined, it becomes possible to unravel the underlying biochemical mechanisms that govern the observed regions or phenotypes. Over the past few decades, MSI has emerged as a powerful tool for studying the interactions between microorganisms, such as microbial-microbial or microbe-host interactions, and their direct environmental interfaces. The advantage of MSI is its ability to generate 2D molecular maps (ion images) of *in situ* bacterial colonies, unveiling chemical gradients and the precise locations of metabolite production and diffusion within bacterial colonies or at the interface of interacting species[1,2]. This is particularly useful for understanding localized chemical signaling, competitive interactions, and cooperative behavior among microorganisms. Moreover, the ability to visualize the spatial distribution of ions enables us to discover novel molecules, especially those confined to specific regions of a microbial colony or to the interaction zone between species, which might be missed with other analytical methods. Matrix-Assisted Laser Desorption/Ionization (MALDI)-MSI, particularly, is well-suited for these applications, offering high spatial resolution and sensitivity[3].

The hyperspectral nature of MSI makes it challenging to analyze, as typical MSI experiments often involve tens of thousands of pixels and several thousand *m/z* peaks. Moreover, MALDI-MSI data are complex to interpret due to background ions, the possibility of a molecule being ionized with different adducts, and the presence of fragment ions that may result from in-source fragmentation (to a lesser extent) or from in vivo degradation processes [4,5]. Hence, reducing the number of *m/z* values to a smaller set, referred to as "ions of interest," appears crucial. Decreasing the list of *m/z* values also helps prevent being overwhelmed by excessive information, which is essential before manually interpreting the biological significance of these ions.

Manual interpretation is often assisted by specialized software, which enables the users to explore ion images and extract biological insight[6]. However, these tools still require substantial manual effort to refine the processed data and identify ions under a specific experimental condition. In studies of bacterial secondary metabolism, we often try to stimulate the production of unknown metabolites by changing growth conditions or the environment. This traditional approach to natural product discovery typically yields gene expression patterns that are either fully on or fully off. Univariate tests or discriminant analyses can detect ions whose intensity differs between two conditions, but these approaches do not incorporate the spatial distribution of ions[7,8].

For this reason, this chapter presents a general MSI preprocessing and feature-prioritization workflow to identify biologically relevant ions in comparative experiments (Figure 2.1). The workflow is separated into four steps. The workflow begins with the conversion of raw MSI data into an analysis-ready representation, followed by the construction of a common feature space, the generation of ion images, and the comparison of these ion images between control and experimental groups. Particular emphasis is placed on the idea that ion prioritization can be improved by combining conventional signal-based criteria with metrics that capture the spatial organization of the images.

Then, this scoring concept is operationalized as the Image Structure Filtering (ISF) algorithm and evaluated against a simpler Differential Signal Filtering (DSF) baseline based on on/off detection. The two algorithms are compared on a manually curated reference set and then jointly applied to bacterial case studies aimed at the discovery of condition-specific natural products.

This workflow also provided the basis for the subsequent development of *deltamsi*, an open-source Python package dedicated to high-resolution MSI data processing and interpretation. As discussed at the end of the chapter, *deltamsi* was designed to formalize and generalize the main steps introduced here (preprocessing and feature prioritization), including multi-sample MSI data handling, construction of analysis-ready feature matrices, recalibration, preprocessing, ion-image extraction, spatial scoring, and differential comparison between groups. In that sense, the present chapter introduces not only a methodological strategy for condition-specific ion discovery but also the conceptual foundations of a reusable computational framework for MSI preprocessing that will be extended in the following chapters.

The chapter is organized as follows. Section 2.2 describes the MSI preprocessing and feature-prioritization workflow for converting raw data into a structured representation suitable for downstream analysis and introduces the principle of ion-image scoring as a complementary criterion for feature prioritization. Section 2.3 then presents the DSF/ISF study as an application of this workflow to bacterial natural product discovery. Finally, Section 2.4 explains how the same analytical logic was consolidated and extended in the *deltamsi* package, thereby providing a bridge between the methodological developments of this chapter and the broader software framework used throughout the thesis.

2.2 MSI processing pipeline

Generally, MSI preprocessing workflows involve converting raw data into a more accessible representation, detecting prominent *m/z* peaks, aligning/recalibrating *m/z* peaks between spectra to create features, normalizing the detected signal, filtering features according to their *m/z* and/or spatial information, defining spatial regions of interest, and finally linking ions to the spatial areas of interest[7,9]. The application of these processing steps may vary depending on the analysis's end goal. However, they all rely on transforming raw data into a more accessible representation for specific data processing. This chapter establishes a mass spectrometry imaging pipeline (Figure 2.1). This process involves transforming raw data into a data cube and scoring the mass-to-charge ratios (*m/z*) according to their spatial distribution. This allows us to identify *m/z* values of interest by comparing them to a control group.

Within this section, Subsections 2.2.1–2.2.3 describe the generic preprocessing components, data conversion, mean-spectrum peak detection, and ion-image generation that produce the data cube on which all downstream analyses operate. Subsection 2.2.4 then introduces the conceptual pivot of the chapter: an image-structure score that moves beyond intensity-only descriptors and provides a spatially aware criterion for feature prioritization.

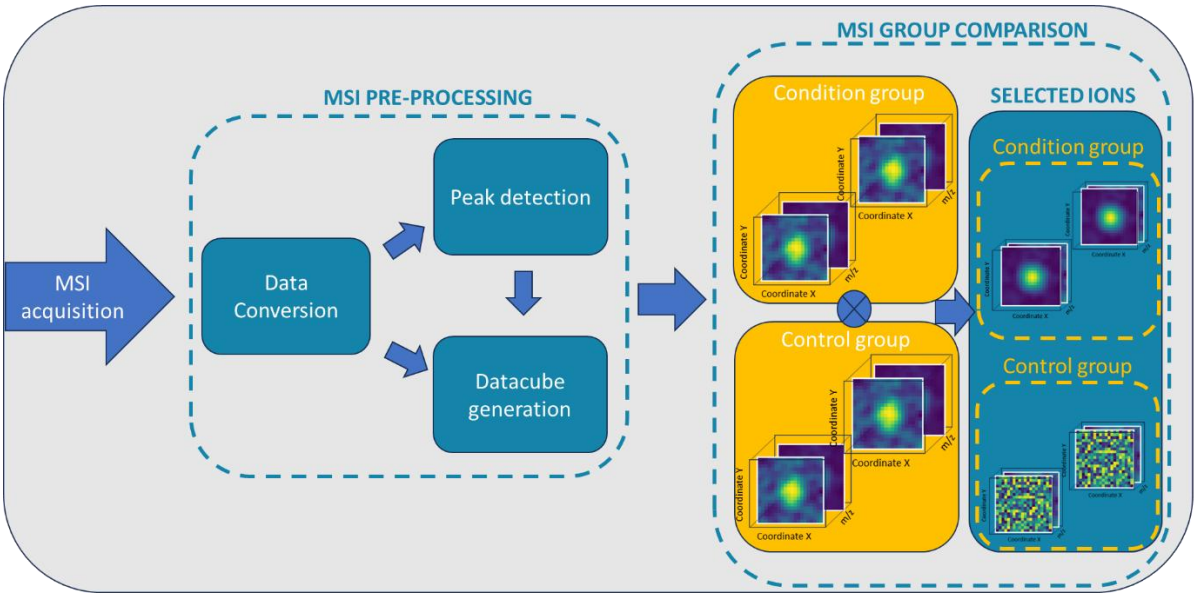


Figure 2.1. *Overview of the MSI preprocessing workflow described in this chapter. MSI data are acquired from multiple samples and converted from vendor-specific formats into imzML (Section 2.2.1). A mean spectrum is computed across the image and used for peak detection to define a common m/z axis (Section 2.2.2). Ion images are then generated for each selected m/z, forming a data cube for each sample (Section 2.2.3). Finally, data cubes are separated into control and condition groups, and ion images are compared between groups to extract ions of interest, i.e. ions expressed in the condition but absent (or not structured) in the controls (Section 2.2.4).*

2.2.1 Mass Spectrometry Image Handling for Data Processing

After acquiring mass spectrometry imaging (MSI) data, the resulting files are typically stored in vendor-specific proprietary formats. For downstream analysis, it is therefore convenient to convert these files into an open, standardized format. In this work, we use the imzML format[10], which has become a widely adopted standard for MSI and is supported by most instrument manufacturers through dedicated export or conversion tools. imzML stores, for each pixel of the image, the corresponding mass spectrum together with its spatial coordinates and acquisition metadata, and thus provides a flexible basis for quality control, visualization, and subsequent processing. The imzML format facilitates, already, diverse analytical approaches, including visualizing spatial distributions, performing compound identification, or more advanced analyses such as Kendrick mass defect analysis (see Chapter 4). During MSI acquisition, the mass spectra can be stored in profile or centroid mode (Figure 2.2). In profile mode, all data points that draw the *m/z* peak are preserved, allowing detailed inspection of peak shapes and assessment of the quality of mass-spectrum acquisition, such as the peak resolution. This mode is particularly useful for quality control, but can result in large file sizes. In contrast, the centroid mode is more effective for analyzing large volumes of MSI data. Ideally, one data point is retained for each detected *m/z* peak (centroid) in this mode. As a result, the imzML files in centroid mode contain significantly fewer data points per peak, making them much smaller and more manageable. This reduction in data points is significant in high-resolution mass spectrometry, where spectra from typical images can reach up to 100 GB in profile mode. Centroid mode reduces the file size, facilitating data handling and storage. However, the peak-picking method is often not optimized, resulting in multiple centroids being retained for a single peak, which complicates further analysis.

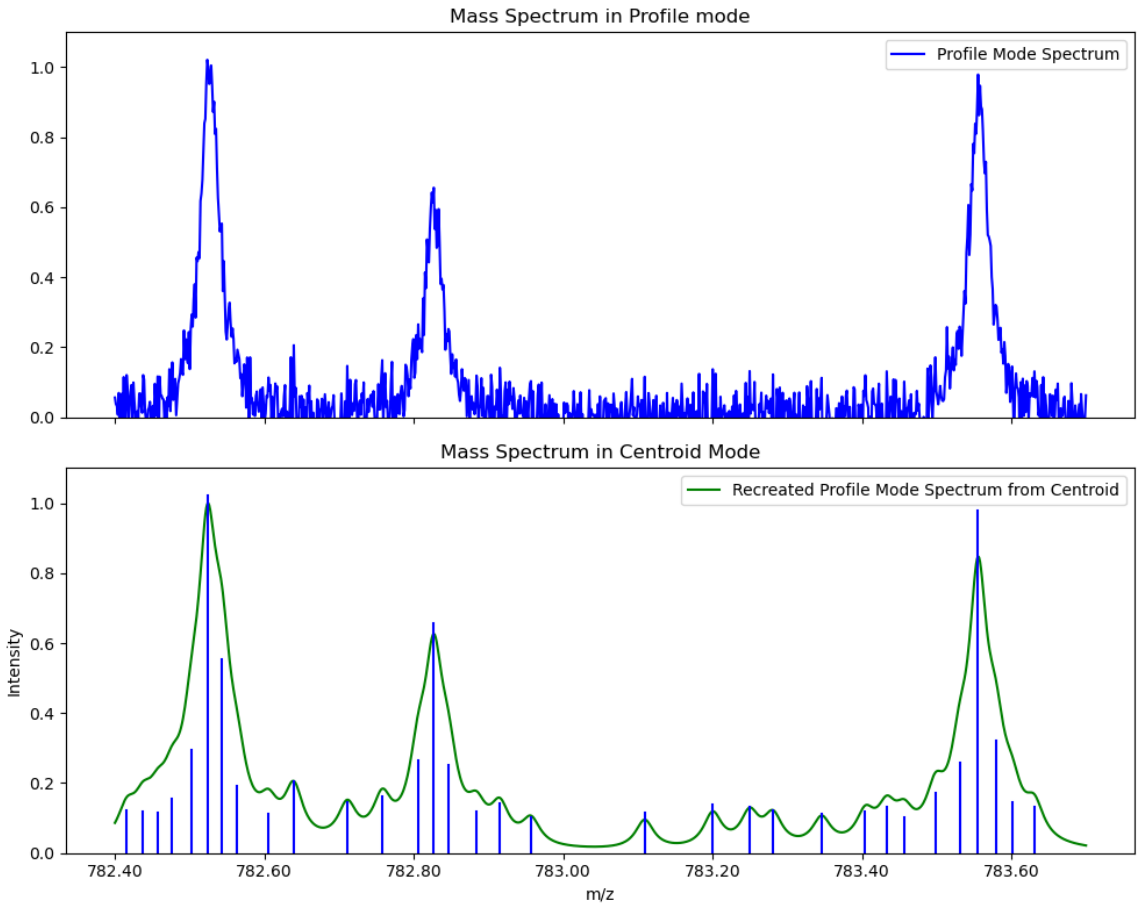


Figure 2.2. *shows a mass spectrum in profile mode at the top, where multiple data points are used to represent the signal. Below, the same spectrum is displayed in centroid mode, which uses fewer data points to describe the signal. The green trace represents a reconstructed profile mode derived from the centroid peaks. This reconstructed profile can be utilized to compute an average spectrum from centroid MSI. It is important to note that centroid mode employs a peak detection algorithm that does not always yield a single point for each peak.*

For many advanced analyses, it is advantageous to reorganize the MSI data into an intermediate structure known as a data cube (Figure 2.3). In this representation, each ion image corresponds to a specific *m/z* value from a predefined list, so that ion images can be rapidly accessed and compared by index.

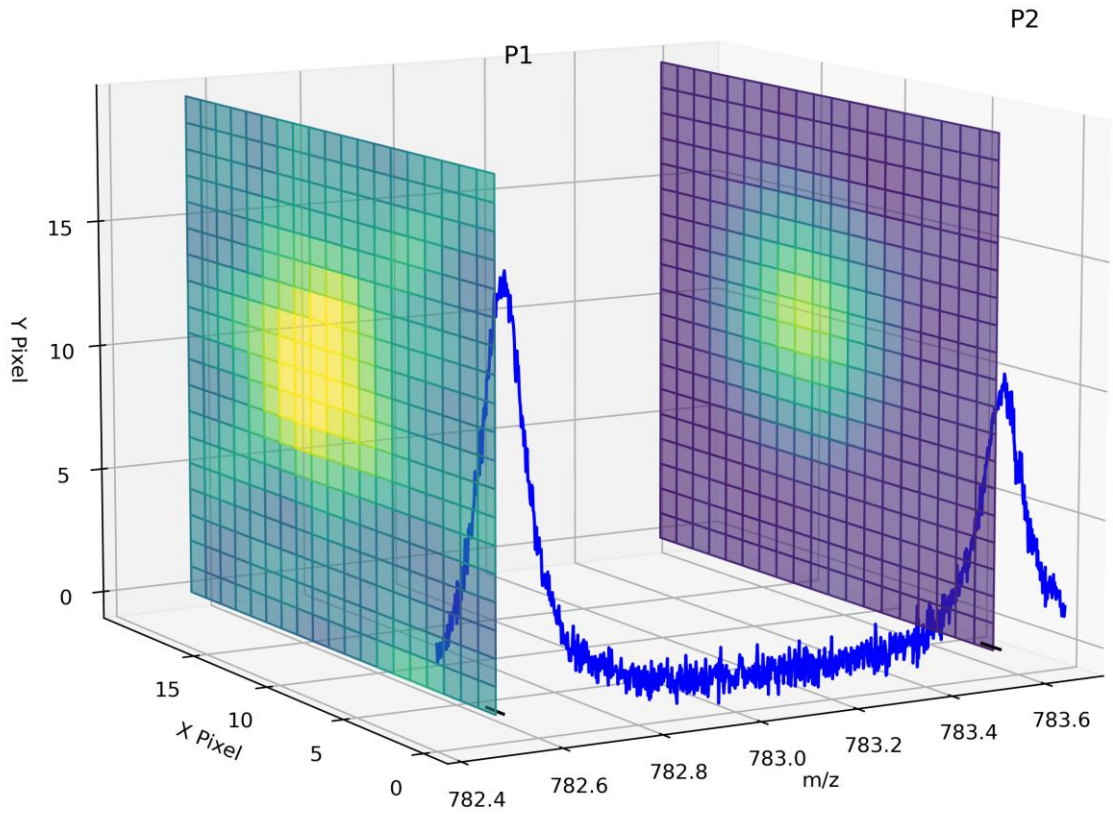


Figure 2.3. *Representation of an MSI data cube. In the foreground is a representation of the average spectrum of the MSI. On this spectrum, two peaks are well-defined: P1 and P2. Each MS peak's ion image is generated by finding the maximum peak intensity in each spectrum from the 20x20 spectra that compose the MSI. Only well-defined peaks are used to generate ion images; noisy signals in between the peaks do not generate ion images. The ion images of those two peaks are represented by a heatmap, with maximum intensity represented by yellow pixels and minimum intensity represented by dark blue pixels.*

In practice, once the MSI data is converted into a data cube, an ion image for any *m/z* value in the cube can be retrieved instantly. This is a significant improvement over the imzML format, where direct access to specific ion images is not feasible. The data cube

format streamlines data analysis, enabling faster, more efficient exploration of ion distributions within the sample. This improved data-handling capability is especially useful for tasks that involve frequent ion-image comparisons. For instance, it can be used for analyzing the differences between experimental conditions or identifying specific chemical signatures in complex biological samples, which are directly relevant to this work.

2.2.2 Average Spectrum and Peak Detection

To generate a data cube from an imzML file, we first need to define a list of *m/z* values for which ion images will be created. Because we do not know in advance which ions will be biologically relevant, it is important to retain as many MSI-detected *m/z* values as reasonably possible. A common strategy is therefore to select the most intense peaks from an average (mean) spectrum computed over the MSI dataset, or over a set of MSIs.

The average spectrum is constructed in several steps. First, the mass spectrum associated with each MSI pixel is extracted; each spectrum consists of a list of *m/z* values and their corresponding intensities. For each *m/z* value along the spectral axis, intensities from all pixels are summed, and this sum is then divided by the number of pixels. The resulting mean spectrum represents the global intensity distribution of ions across the image and provides a compact one-dimensional summary from which prominent *m/z* peaks can be selected.

In centroid MSI, each peak is ideally represented by a single centroid *m/z* value per spectrum. In practice, however, the same ion peak measured in two different spectra may be reported at slightly different *m/z* values due to limited mass accuracy and rounding, so that peaks do not align exactly at high decimal precision (typically the fourth or fifth decimal, depending on the instrument). As a consequence, simply summing intensities at identical *m/z* values may underestimate the true average signal. To mitigate this, the peaks in each spectrum can be artificially “widened” before or during averaging. This can be achieved either by replicating the centroid intensity over

a small *m/z* window centered on the centroid or by reconstructing a Gaussian-like peak profile around the centroid *m/z* (Figure 2.2). When iterating over the pixels to build the mean spectrum, this peak widening can be applied on the fly.

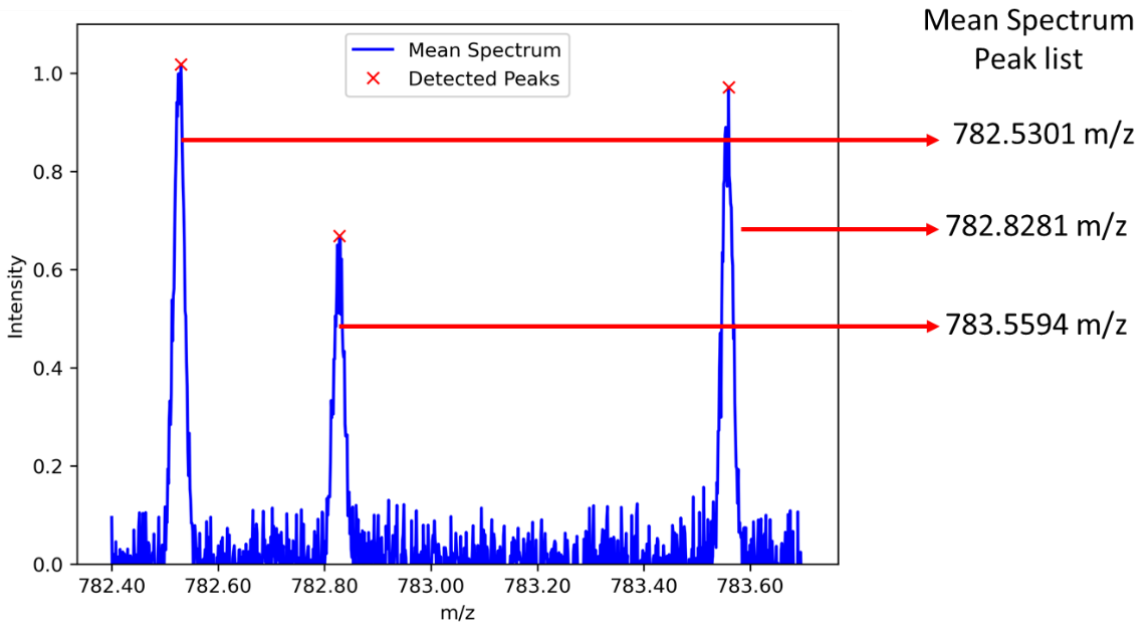


Figure 2.4. *illustrates an average spectrum and the corresponding peak-detection step. Three prominent peaks can be visually identified, and their apices are determined by locating local maxima (window size = 0.01 m/z) and comparing each candidate peak with its neighboring m/z values. The m/z positions of these detected peaks are then used as candidate entries for the data cube.*

Formally, once the mean spectrum has been computed, peak picking proceeds by scanning a contiguous *m/z* window across the full spectrum (Figure 2.4). Within each window, local maxima are identified as points whose intensity is higher than that of their immediate neighbors. If multiple local maxima are detected within the same window, only the most intense is retained. After this procedure, the most intense *m/z* values are selected. This number, usually between 2,000 and 5,000, is chosen empirically to provide broad coverage of MSI signals while limiting the risk of discarding ions of interest that may be weakly expressed or spatially localized. In practice, the optimal number of peaks depends on sample complexity and may need to be adjusted when averaging across multiple MSI datasets.

Additional signal-processing steps can, in principle, be applied to improve peak detection. Noise-filtering methods can reduce background noise and improve peak visibility; baseline-correction algorithms can remove non-chemical contributions and baseline drift; and smoothing filters, such as the Savitzky–Golay filter[11], can attenuate high-frequency fluctuations while preserving peak shape. However, applying these operations alters the peak intensities in the underlying spectra and, consequently, the spatial distribution of ion intensities in the reconstructed ion images. Because our downstream analysis explicitly relies on these spatial distributions, we chose not to apply such intensity-transforming preprocessing steps in this work.

2.2.3 Generation of Ion Images for Each Peak

Generating ion images for each detected peak is a crucial step in MSI data processing, as it enables the spatial visualization of specific ion distributions within the sample. In our workflow, the list of significant *m/z* peaks obtained from the mean-spectrum peak-picking step (Section 2.2.2) is used as the basis for ion-image generation. For each selected *m/z* value, we construct an ion image by scanning all pixels in the MSI dataset and extracting the intensity value corresponding to the selected *m/z* (Figure 2.5).

For each spectrum, a mass window Δ*m/z* is defined around the target *m/z* value to account for small mass deviations and to ensure robust intensity estimation. All centroid peaks falling within this window are considered, and if multiple centroids are present, only the most intense one is retained. Multiple centroids within such a narrow window are assumed to be redundant detections of the same ion rather than distinct species, since Δ*m/z* is set below the instrument's resolvable mass separation, typically of the order of ±0.01 *m/z* for TOF instruments and ±0.005 *m/z* (or a few ppm) for FT-ICR. Closely spaced centroids in this range, therefore, most often reflect imperfect vendor peak-picking (Section 2.2.1) rather than genuinely separate ions. To suppress contributions from noise, we also apply a minimum intensity threshold, which empirically corresponds to the noise level in our data. If no centroid above this threshold is found within the mass window, the intensity at that pixel is set to 0 in the corresponding ion image.

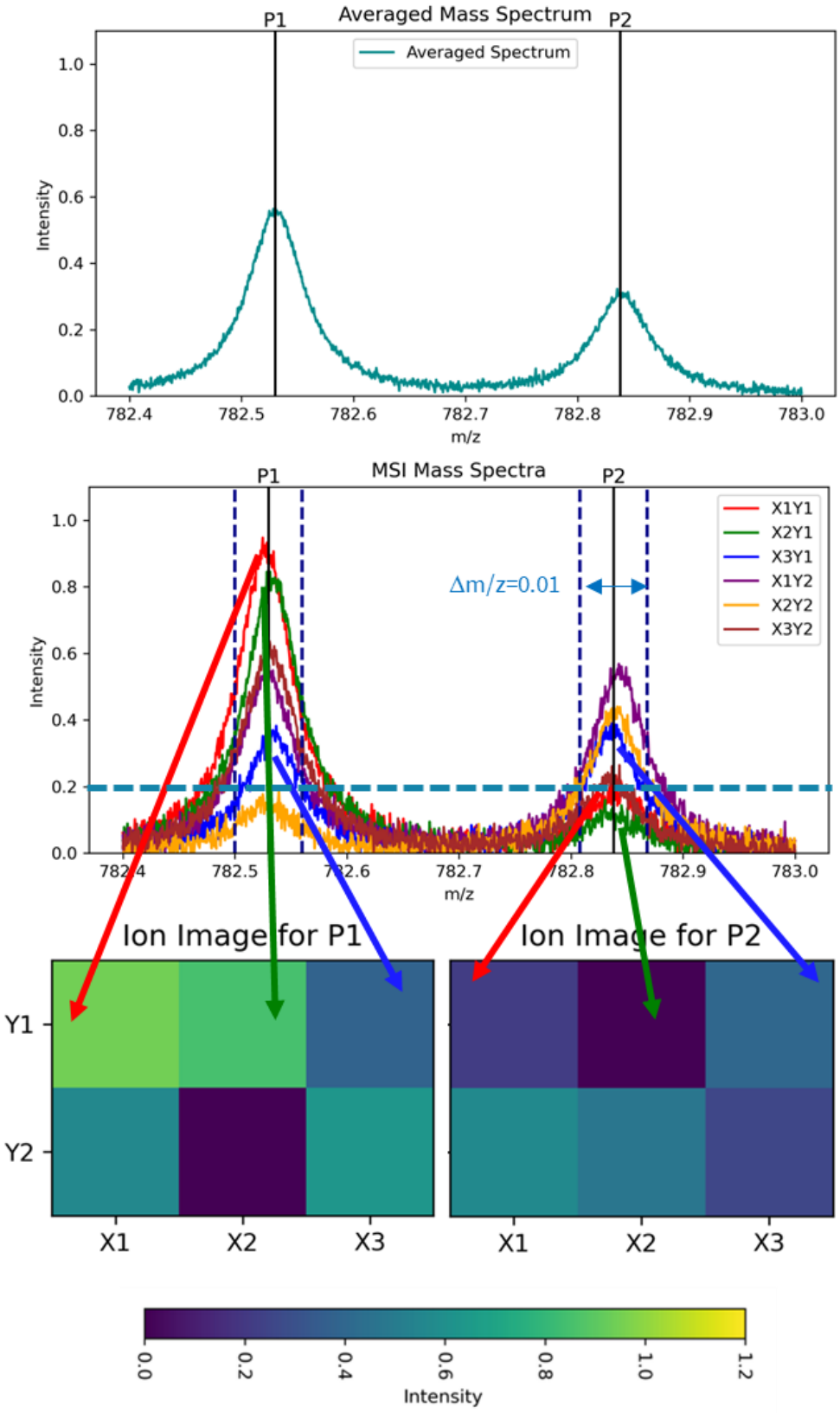

Averaged Mass Spectrum
P1
P2
Averaged Spectrum
Intensity
m/z
MSI Mass Spectra
X1Y1
X2Y1
X3Y1
X1Y2
X2Y2
X3Y2
Δm/z=0.01
Ion Image for P1
Ion Image for P2
Y1
Y2
X1
X2
X3
Intensity

Figure 2.5. *illustrates the generation of ion images for two peaks (P1 and P2) in a mass spectrometry image (MSI) analysis with a spatial resolution of 2x3 pixels. The m/z values of the apex of the prominent peaks (P1 and P2) are identified from an average spectrum (shown at the top). A mass window of ±0.03 m/z is selected around each peak m/z, marked by the dark blue line. The mass spectra for each of the six pixels in the MSI are shown as different-colored traces in the bottom mass spectrum. Ion images for the two peaks are generated by selecting the maximum intensity value within this tolerance window for each mass spectrum. A minimum intensity tolerance of 0.2 is set; therefore, the yellow trace for P1 and the green trace for P2 each result in a pixel intensity of 0 in their respective ion images (dark blue).*

Peak widening (Section 2.2.2) and the *m/z* extraction window (Section 2.2.3) address the same underlying problem: centroid *m/z* values do not align perfectly between pixel spectra. Peak widening broadens each centroid (by replicating its intensity over a small *m/z* range or reconstructing a Gaussian profile) so that nearly-coincident centroids from different pixels contribute coherently to the mean spectrum. The extraction window does the reverse: once a target *m/z* has been selected from the mean spectrum, it defines a tolerance ±Δ*m/z* within which each pixel spectrum is searched, and the most intense centroid is retained as the pixel intensity. Peak widening, therefore, combines centroids across spectra to build a consistent feature axis, whereas the extraction window locates and reduces centroids within each spectrum to populate an ion image. Both window widths are governed by the typical mass deviation of the instrument and should be chosen consistently.

The recorded intensity values are then reassembled into a two-dimensional array, preserving each pixel's X and Y coordinates. This array forms the ion image, which displays the spatial distribution of the ion's intensity across the sample. To facilitate comparison between different ion images, intensity values may be normalized. This normalization involves scaling the intensity values to a common range (e.g., between 0 and 1) or applying other techniques to account for variations in ionization efficiency or sample preparation. A common normalization method is TIC (total ion count) normalization, which divides the intensity of each spectrum by the sum of all ion

intensities. This normalization assumes that the overall spectral content of each spectrum is similar, but the intensity of an ion can vary between pixels[12]. TIC normalization has several complementary effects. Within a single spectrum, dividing every intensity by the same TIC preserves the relative ratios between peaks; only the absolute scale changes. Across pixels, the operation is applied independently to each spectrum, so for a given *m/z*, the ion-image value becomes the ion's fractional contribution to the local TIC. An ion present at the same absolute amount in two pixels, therefore, appears more intense in the pixel with the lower TIC, which can introduce apparent spatial localizations driven by differences in overall chemical content rather than by the ion's distribution. In the mean spectrum, each pixel contributes with equal weight rather than being implicitly weighted by its TIC, so ions from low-TIC pixels (e.g., background) become more prominent while ions specific to high-TIC regions are suppressed, which can shift the *m/z* values retained during peak detection. Across ions, mean ion-image intensities also become only relatively comparable: two ions can show similar mean normalized values while differing substantially in absolute abundance. TIC normalization, therefore, preserves relative comparisons within a constant chemical background but can bias them whenever sample regions differ markedly in overall content, which is why it was not applied to the bacterial MSI datasets in Section 2.3.

The generated ion images can be visualized using software such as SCiLS™ or custom Python scripts, typically displayed in false color to highlight regions of varying ion intensity. Regions with high intensity correspond to areas of strong ion detection and may indicate zones of active metabolite production, diffusion, or accumulation.

High mass accuracy and resolving power are essential for distinguishing closely spaced *m/z* values and minimizing signal overlap between different compounds. In practice, small mass drifts between pixels can cause peaks belonging to the same ion to fall partly outside the chosen mass window, leading to incomplete ion images and

apparent loss of structure (Figure 2.6). The same drift also broadens peaks in the average spectrum and reduces effective mass resolution.

Recalibration and alignment of *m/z* values are therefore critical to preserve both mass resolution and spatial fidelity. By correcting systematic *m/z* shifts across the image, recalibration ensures that peaks for a given ion fall consistently within the extraction window, improving ion-image completeness and the quality of the mean spectrum. The recalibration algorithm itself is detailed in Chapter 3.

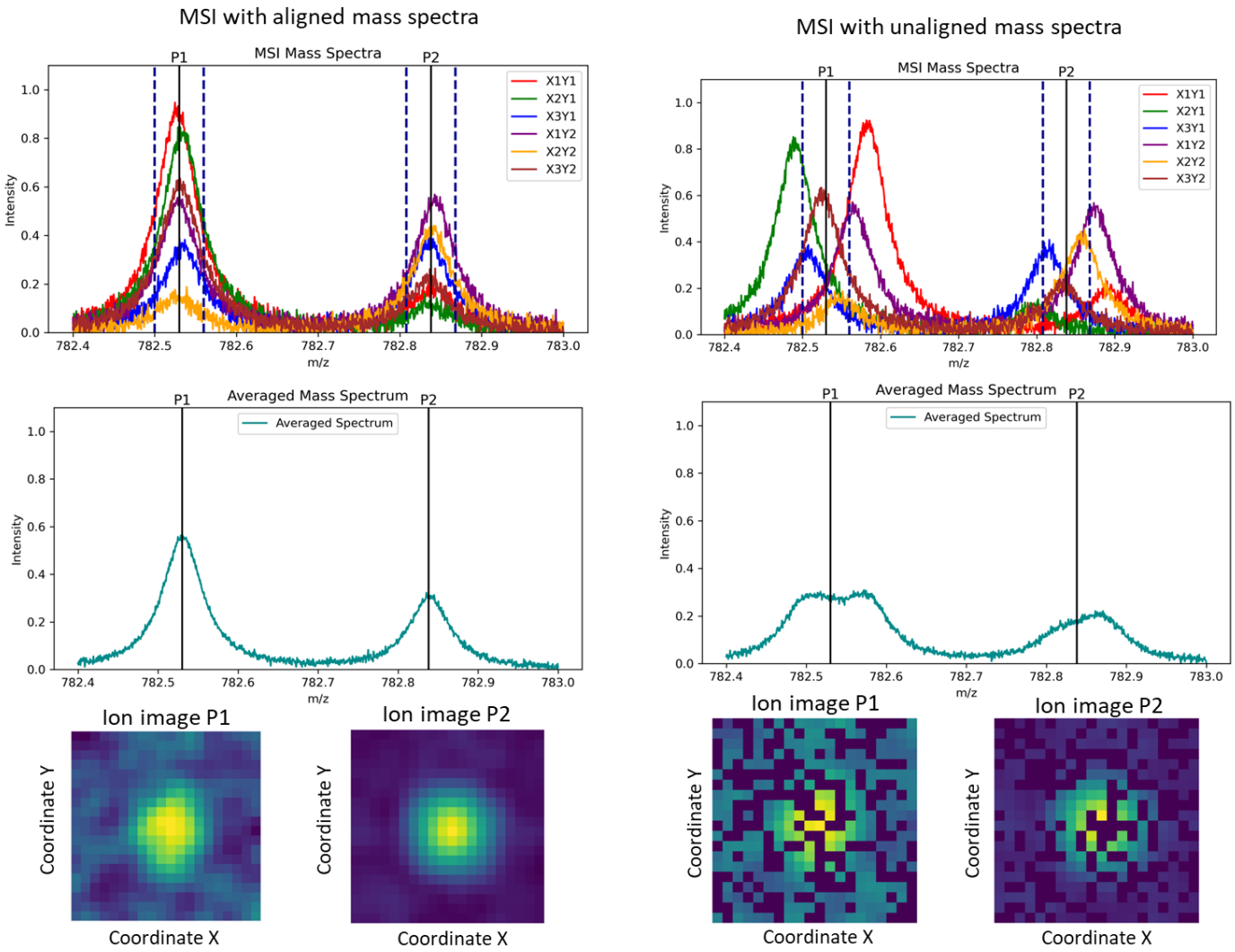


Figure 2.6. *Illustration of the generation of ion images for two peaks (P1 and P2) in a mass spectrometry image (MSI), with a spatial resolution of 20x20 pixels. The figure is organized into two columns and three rows. The first column represents the generation of ion images for an MSI without mass shifts, which appears as MSI with aligned spectra. The second column depicts the generation of ion images with a mass shift and thus MSI with unaligned spectra. The first row presents a plot of mass spectra obtained from 6 pixels of the 20x20 image, with each spectrum displayed in a different color. The m/z values of the ion images are indicated by black vertical lines for P1 and P2. A mass window of ±0.03 m/z is selected around each ion image m/z, marked by the dark blue line. It is noted that some peaks in the unaligned MSI are shifted outside this mass window and will not be considered for image generation, explaining why dark pixels (set to 0) are visible in the image. The second row of the figure shows the average spectrum of the MSI. The unaligned MSI exhibits a lower resolution for the two main peaks compared to the MSI without a mass shift, despite both having comparable spectral resolutions. The third row corresponds to the ion images of the two peaks from the 20x20 image. As explained above, in the case of the unaligned MSI, the ion images appear incomplete because some pixels have an intensity value set to 0 because certain peaks were not included within the ±0.03 m/z window around the peak m/z, and then excluded from the image reconstruction process.*

2.2.4 Scoring ion image structure

Once ion images have been generated for each detected peak and assembled into a data cube, we can visualize the spatial distribution of each ion and select ions of interest. In metabolomics based on liquid chromatography-mass spectrometry (LC-MS), each feature is usually defined by its retention time and *m/z*, and a common way to identify ions specific to an experimental condition is to compute a feature fold change between groups. In that setting, each feature is represented by a scalar quantity, such as a peak area or peak height, which can be directly compared between samples.

In MSI, by contrast, features are represented as images rather than as scalar peak intensities. This representation makes it more challenging to define the “abundance” of a feature, because the ion signal is distributed across the sample surface. A simple global summary, such as the mean intensity of an ion image over all pixels, provides only a partial description. Two ions can have the same average intensity but very different spatial distributions: one may be strongly expressed in a small interaction

region, whereas another may be weak and diffuse over a larger part of the sample. An analysis based solely on global intensity would treat these ions as equivalent, even though they likely reflect distinct biological processes. Consequently, extracting biologically meaningful information from MSI requires methods that consider not only how much signal is present, but also how that signal is organized spatially.

In the present work, the objective is to identify ions that are specifically associated with a condition of interest, such as a microbial confrontation, while being absent, or at least not spatially structured in the corresponding control conditions. A straightforward strategy is to remove *m/z* values observed in the controls and retain only those detected in the condition of interest, without considering their spatial distribution. This idea underlies the Differential Signal Filtering (DSF) algorithm, which is described and evaluated in Section 2.3. DSF therefore provides a useful baseline based on an "on/off" detection logic.

However, this strategy does not account for the fact that noisy or artefactual ions may still display relatively high intensities. In MSI, noise is not defined only by low abundance, but also by the absence of coherent spatial organization. Noisy ions often appear as dispersed or fragmented pixels across the image, whereas biologically relevant ions tend to form structured regions, fronts, or localized zones. For this reason, the spatial organization of an ion image can be used as an additional descriptor to prioritize features of interest.

To capture this aspect, we used a variant of the spatial chaos approach introduced by Palmer and co-workers for MSI annotation[13], in which the informativeness of an ion image is related to the degree to which signal forms coherent spatial regions rather than fragmented, random patterns. In practice, the score used here measures the extent to which an ion image forms contiguous and spatially organized structures as opposed to

scattered, isolated pixels. Images with highly fragmented signals receive low scores, whereas images with coherent spatial regions receive high scores.

The detailed mathematical definition of the score is provided in the Supplementary Information (S1). Briefly, each ion image is normalized and transformed into a set of binary images across a series of intensity thresholds. For each thresholded image, the number of connected signal regions is evaluated relative to the number of positive pixels, and these quantities are aggregated into a single score reflecting the degree of spatial structure. Although the terminology “spatial chaos score” is retained for consistency with the original literature, higher values correspond in practice to lower spatial chaos, that is, to stronger spatial organization.

The spatial chaos score provides a compact descriptor of the spatial structure of an ion image. A low score indicates that the signal is fragmented into many small disconnected regions, as expected for noisy or weakly informative features. A high score indicates that the signal is concentrated in fewer and larger connected regions, as expected for ions associated with biologically meaningful spatial domains.

More generally, the score depends primarily on spatial organization, the number, size, and connectivity of signal regions, rather than on absolute ion intensity, which is removed by the per-image normalization (Section S1). Two ions with similar spatial patterns but very different absolute abundances therefore receive similar scores, while two ions with similar mean intensities but different organizations (e.g., one concentrated in a coherent region, the other diffuse and fragmented) receive markedly different ones. Combined with the noise threshold applied during ion-image generation (Section 2.2.3), this means that weak but spatially coherent ions are not penalized relative to intense ones. A residual sensitivity to intensity remains in two ways: the shape of the intensity distribution affects the geometry of the binary images at each threshold (a sharp focus and a broad gradient yield different cluster patterns), and ions

whose intensity sits close to the noise level can be penalized because low-threshold binary images then mix structured signal with residual noise pixels, fragmenting the connected components. In both cases, intensity influences the score only through its spatial expression, not through its absolute scale.

Because the spatial chaos score is computed directly on ion-image values, TIC normalization (Section 2.2.3) can influence the score by reshaping the ion images themselves. When TIC varies markedly across the image, dividing each pixel by its local TIC suppresses signal in chemically rich regions and boosts it in poorer ones. An ion genuinely localized in a high-TIC region will therefore appear less concentrated after normalization and may receive a lower score, whereas an ion present uniformly in absolute terms across regions of unequal TIC will appear artificially localized to the low-TIC regions and may receive a higher score despite having no real spatial structure. This sensitivity is one of the main reasons TIC normalization was not applied to the bacterial MSI datasets analyzed in Section 2.3.

Once a spatial chaos score has been assigned to each ion image, images of the same ion can be compared between groups. Therefore, the comparison is performed on a descriptor of spatial structure rather than on a scalar measure of abundance.

For each *m/z* value, the score is computed for all corresponding ion images in the control and condition groups. To summarize each group, the maximum score observed among the corresponding MSI datasets is retained. This choice makes it possible to detect ions that form a clear and biologically relevant structure in at least one replicate, even if the pattern is weaker in the others.

The structure-based fold change is then defined as:

$$\mathrm{FC}_S(m) = \frac{S^{\max}_{\text{condition}}(m)}{S^{\max}_{\text{control}}(m)}$$

where $S^{\max}_{\text{condition}}(m)$ and $S^{\max}_{\text{control}}(m)$ denote the maximum spatial chaos scores observed for the ion in the condition and control groups, respectively.

This ratio provides a direct measure of how the spatial organization of an ion image changes between groups. Values greater than 1 indicate that the ion is more spatially structured in the condition group than in the controls, suggesting the emergence of a coherent pattern under the condition of interest. Values close to 1 indicate a similar level of structure in both groups, whereas values below 1 indicate that the ion is more structured in the controls than in the condition group.

By ranking *m/z* values according to this structure-based fold change, it becomes possible to prioritize ions that are not only detected in the condition of interest, but also form coherent spatial patterns in that condition. In the following section, this spatial criterion is considered together with DSF in order to select ions produced specifically under the bacterial experimental conditions of interest.

This section has presented the rationale and implementation of the Image Structure Filtering (ISF) algorithm, which prioritizes ions according to the spatial organization of their ion images. In Section 2.3, ISF is applied and evaluated, together with the DSF baseline, in a bacterial case study aimed at identifying condition-specific ions.

2.3 Exploiting Differential Signal Filtering (DSF) and Image Structure Filtering (ISF) Methods for Untargeted Mass Spectrometry Imaging of Bacterial Metabolites

Using the MSI workflow developed in this chapter, we analyze MSI data from bacterial colonies to identify condition-induced natural products. In the developed workflow, we created an Image Structure Filtering (ISF) algorithm that scores each ion image based on the randomness of its spatial distribution and compares this score between conditions. We compare it to a Differential Signal Filtering (DSF) algorithm that uses a more classical on/off detection and does not explicitly consider spatial structure. DSF therefore provides a natural baseline to quantify the added value of ISF.

We evaluated ISF and DSF by manually selecting ions of interest in a confrontation experiment between *Escherichia coli* and *Streptomyces coelicolor*. The manually curated list provided a reference set to assess the ability of both algorithms to retrieve condition-specific ions.

Once evaluated, the two methods were successfully applied in combination on two case studies: (i) cellobiose-mediated induction of the pathogenicity of *Streptomyces scabies*, the causative agent of common scab on root and tuber crops, and (ii) iron-repressed production of siderophores of *Streptomyces scabies*. They effectively highlight strain-associated natural products.

In this work, raw data were acquired and converted to imzML files in centroid mode (Section 2.2.1), so that each pixel spectrum is represented by its own list of (*m/z*, intensity) centroids rather than by the densely sampled point cloud of profile mode. Profile-mode data are not used at any stage of this analysis. The centroid imzML files were recalibrated (see Chapter 3 for details) before being converted into one of two centroid-based representations, each tailored to a different algorithm. For ISF, the recalibrated centroids are projected onto a data cube in which all spectra share a

common *m/z* axis defined by peak picking on the mean spectrum (Section 2.2.2): the per-spectrum *m/z* values are replaced by their nearest match on this shared axis, which enables fast index-based access to ion images. For DSF, the same recalibrated centroids are stored in an ICK file, in which each spectrum retains its original *m/z* values without being forced onto a shared axis. The two representations, therefore, differ in how they handle the per-spectrum variability inherent to centroid data: the data cube trades faithful preservation of original *m/z* values for indexed ion-image access, whereas ICK preserves the original *m/z* values of each centroid at the cost of slower ion-image extraction and comparison. Because our MSI datasets contain distinct structures with markedly different total ion counts, we did not apply TIC normalization. In this context, TIC normalization can introduce biases and artificially alter the spatial distribution of ions, precisely what ISF aims to capture.

The complete work is detailed in the following publication: “Exploiting Differential Signal Filtering (DSF) and Image Structure Filtering (ISF) Methods for Untargeted Mass Spectrometry Imaging of Bacterial Metabolites”, written by Pierre Burguet, Raphaël La Rocca, Christopher Kune, Déborah Tellatin, Nudzejma Stulanovic, Augustin Rigolet, Johann Far, Marc Ongena, Sébastien Rigali, and Loïc Quinton and published in J. Am. Soc. Mass Spectrom. 2024, 35, 8, 1743–1755.

My contributions to this work were: (i) development of the MSI processing script; (ii) design and implementation of the ISF ion-image scoring algorithm; (iii) recalibration of the MSI datasets; and (iv) application of ISF to the different experimental conditions and analysis of the resulting data.

The complete publication is available in the 'Associated Papers' section of the thesis, Chapter 2, A1.

ACS Partner Journal

Journal of the American Society for Mass Spectrometry

pubs.acs.org/jasms

Article

Exploiting Differential Signal Filtering (DSF) and Image Structure Filtering (ISF) Methods for Untargeted Mass Spectrometry Imaging of Bacterial Metabolites

Pierre Burguet, Raphaël La Rocca, Christopher Kune, Déborah Tellatin, Nudzejma Stulanovic, Augustin Rigolet, Johann Far, Marc Ongena, Sébastien Rigali, and Loïc Quinton*

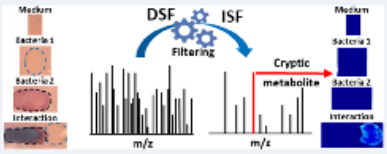
Cite This: J. Am. Soc. Mass Spectrom. 2024, 35, 1743−1755

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

Downloaded via UNIV DE LIEGE on September 2, 2024 at 17:06:33 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ABSTRACT: Matrix-assisted laser desorption/ionization (MALDI) mass spectrometry imaging (MSI) is a label-free technique, producing images where pixels contain mass spectra. The technique allows the visualization of the spatial distribution of (bio)molecules from metabolites to proteins, on surfaces such as tissues sections or bacteria culture media. One particularly exciting example of MALDI-MSI use rests on its potential to localize ionized compounds produced during microbial interactions and chemical communication, offering a molecular snapshot of metabolomes at a given time. The huge size and the complexity of generated MSI data make the processing of the data challenging, which requires the use of computational methods. Despite recent advances, currently available commercial software relies mainly on statistical tools to identify patterns, similarities, and differences within data sets. However, grouping m/z values unique to a given data set according to microbiological contexts, such as coculture experiments, still requires tedious manual analysis. Here we propose a nontargeted method exploiting the differential signals between negative controls and tested experimental conditions, i.e., differential signal filtering (DSF), and a scoring of the ion images using image structure filtering (ISF) coupled with a fold change score between the controls and the conditions of interest. These methods were first applied to coculture experiments involving *Escherichia coli* and *Streptomyces coelicolor*, revealing specific MS signals during bacterial interaction. Two case studies were also investigated: (i) cellobiose-mediated induction for the pathogenicity of *Streptomyces scabiei*, the causative agent of common scab on root and tuber crops, and (ii) iron-repressed production of siderophores of *S. scabiei*. This report proposes guidelines for MALDI-MSI data treatment applied in the case of microbiology contexts, with enhanced ion peak annotation in specific culture conditions. The strengths and weaknesses of the methods are discussed.



INTRODUCTION

Matrix-assisted laser desorption/ionization mass spectrometry imaging (MALDI-MSI) is a label-free technique used to visualize the spatial distribution of molecules,[1−3] such as metabolites,[4] peptides,[5] lipids,[6] or proteins[7] on various surfaces. Information obtained from comprehensive studies based on mass spectrometry (MS) and the determination of the spatial distributions of molecules in samples offer significant knowledge about the molecular characterization of biological samples, ranging from tissue sections (organ biopsies, plant slices, ...) to biopolymer thin films.[8] Over the last decades, MSI methods have been increasingly explored in microbiology to produce images of metabolite synthesis and distribution by various interacting microorganisms, usually grown on a thin-layer of agar based medium.[2,9−11] The molecular "snapshot" of the metabolome of strains grown under controlled experimental conditions allows inferring molecular crosstalk and induced metabolite expression within and between microorganisms. Competition for resources and inhibition or elimination of antagonist species would greatly benefit *in situ* and real-time monitoring, making MALDI-MSI well suited for biomarkers, chemical communication elucidation, or new molecule discovery.[12]

Despite the recent advances of MSI applied in the field of microbiology, concerning the reproducibility of sample preparation, homogeneous matrix deposition or data acquisition,[12] the MSI data generated have a huge memory footprint and are challenging to process and analyze. Actually, each image file contains thousands of spectra (and consequently thousands of pixels), each of which displays thousands of mass m/z signals (peaks), especially when MALDI-MSI and high-resolution mass spectrometry (HRMS) are engaged. These $m/$

Received: March 29, 2024
Revised: June 27, 2024
Accepted: June 28, 2024
Published: July 15, 2024



ACS Publications © 2024 American Society for Mass Spectrometry. Published by American Chemical Society. All rights reserved.

https://doi.org/10.1021/jasms.4c00129
J. Am. Soc. Mass Spectrom. 2024, 35, 1743−1755

2.4 Generalization of the workflow in the deltamsi package

2.4.1 Software for MSI data analysis

As discussed in Sections 2.1–2.3, mass spectrometry imaging (MSI) data analysis faces two recurring difficulties. First, MSI datasets typically contain a very large number of detected *m/z* signals, many of which are redundant because they arise from adduct formation, isotopic patterns, or in-source fragmentation. Second, MSI experiments often generate large volumes of data, and local pixel-level variability can complicate ion-image reconstruction when calibration and alignment are not carefully handled. These characteristics make it difficult to move directly from raw data to a concise and biologically interpretable set of ions of interest.

To address these challenges, a broad software ecosystem has emerged around MSI. At the data-format level, the imzML standard has played a central role by providing an open and interoperable representation of spectra together with spatial coordinates and acquisition metadata, thereby facilitating downstream visualization and processing[14]. Building on this standard, many tools have been developed for interactive exploration of MSI datasets, often through graphical interfaces that allow rapid inspection of spectra, ion images, and regions of interest. In the context of the present chapter, this type of visualization is exemplified by software such as SCiLS™ as well as by custom Python scripts. Among open-source graphical environments, MSiReader is one of the tools and provides both targeted and untargeted MSI analysis capabilities while supporting common imaging mass spectrometry formats[15].

Alongside these interactive tools, several scriptable frameworks have been proposed to support reproducible MSI preprocessing and statistical analysis. In the R ecosystem, Cardinal offers a comprehensive environment for MSI data handling, including preprocessing, spatial segmentation, and classification or discriminant analysis, and is well-suited to more complex experimental designs involving multiple tissues or

conditions[16]. More general mass spectrometry processing libraries, such as MALDIquant, are also frequently used as building blocks within MSI pipelines[17]. Galaxy MSI has sought to make reproducible MSI workflows more accessible by integrating preprocessing, statistical analysis, and image co-registration within a single workflow-oriented framework[18].

Despite this diversity, an important limitation remains the balance between analytical breadth and computational scalability, especially for high-resolution MSI data. Several tools are still constrained by in-memory processing, which becomes problematic as dataset sizes increase. This issue has motivated the development of solutions explicitly designed for larger datasets. SpectralAnalysis, for example, was proposed as an end-to-end MSI processing framework with particular attention to the increasing size of imaging datasets[19]. In the R ecosystem, rMSIproc similarly implements a complete MSI processing workflow while emphasizing scalable computation[20]. More general infrastructures, such as matter, further extend this idea by enabling on-disk manipulation of array-like datasets that exceed available memory[21]. Together, these developments reflect a broader movement in MSI informatics toward more scalable and reproducible computational workflows, as highlighted in recent reviews of MSI preprocessing, multivariate analysis, and software tools[22].

Another major dimension of MSI analysis concerns molecular interpretation. Beyond preprocessing and visualization, one must also reduce redundancy among detected signals and support the annotation of chemically meaningful features. Important advances in this area include METASPACE, through the pySM framework, which introduced statistically controlled metabolite annotation for high-resolution MSI using false-discovery-rate-based procedures[23]. Other approaches focus more specifically on redundancy and spectral interpretation. Kendrick mass defect filtering (chapter 4), for instance, provides an efficient way to reveal homologous series and their spatial distributions in MSI datasets[24]. In parallel, tools such as rMSIcleanup and rMSIfragment

were developed to facilitate the annotation of matrix-related peaks and in-source fragments, respectively[5,25].

Taken together, current MSI software covers most of the major analytical needs of the field, including data standardization, visualization, preprocessing, segmentation, statistical comparison, annotation, and chemistry-aware filtering. However, these capabilities are often distributed across distinct tools, programming environments, and data models. In practice, this fragmentation can make it difficult to build a single reproducible workflow that moves seamlessly from scalable preprocessing to feature prioritization, structure-aware filtering, annotation, and downstream statistical analysis while preserving a coherent representation of samples and features throughout the analysis. It is precisely this methodological and practical gap that motivated the development of *deltamsi*, a programmable Python framework designed to integrate the different layers of MSI analysis into a unified, extensible data model.

2.4.2 General description of *deltamsi*

To support the processing workflow described in Sections 2.2–2.3 and to facilitate re-use across projects, we developed *deltamsi*, an open-source Python package dedicated to high-resolution MSI data processing and interpretation. The package was designed to (i) enable efficient access to multiple large MSI datasets within a single analysis framework, (ii) integrate *m/z* recalibration and mass-shift correction as robust and controllable preprocessing steps, and (iii) offer dedicated tools to reduce feature redundancy, enhance feature prioritization, and accelerate interpretation, particularly through Kendrick-based exploration and graph-based de-complexification driven by mass differences.

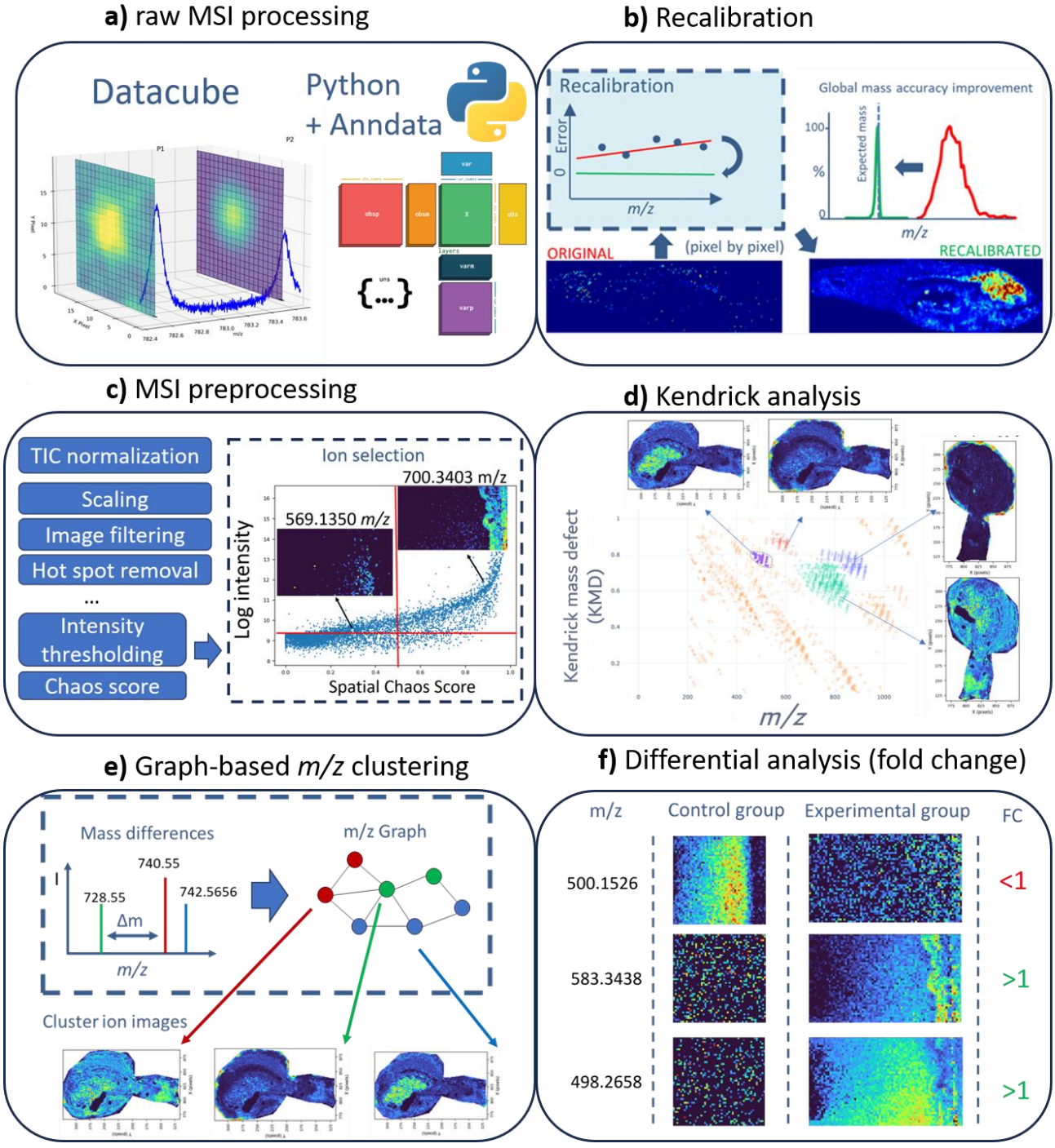


Figure 2.7. *Overview of the main functionalities implemented in deltamsi for high-resolution MSI processing and interpretation. (a) Raw MSI processing and data representation: conversion of raw MSI data into a data cube-like representation in Python, organized in an AnnData-backed structure for efficient access and metadata handling. (b) Recalibration: pixel-wise (or spectrum-wise) mass-shift correction to improve mass accuracy, illustrated by the reduction of m/z error and improved peak alignment, leading*

to cleaner and more consistent ion images after recalibration. (c) MSI preprocessing and feature selection: typical preprocessing steps (e.g., TIC normalization, scaling, image filtering, hot-spot removal) followed by intensity thresholding and spatial-structure scoring (chaos score) to support ion/feature selection. (d) Kendrick analysis: Kendrick mass defect visualization used for exploratory analysis and interactive selection/labeling of ions of interest, with corresponding ion images. (e) Graph-based m/z clustering: construction of an ion graph from expected mass differences (Δm) and/or spatial similarity (colocalization), followed by clustering to obtain molecular groups and cluster-level aggregated ion images. (f) Differential analysis between conditions: comparison of control versus experimental groups using fold-change metrics (intensity and/or spatial-structure/chaos-derived criteria), illustrated by ion images and fold-change direction.

Figure 2.7 provides an overview of the main analysis blocks implemented in *deltamsi*. Importantly, the present chapter introduces the core concepts underlying the package's workflow, which constitute its central processing logic. These concepts include the representation of MSI data from profile or centroid spectra, the computation of sample-level and global mean spectra, peak detection from these mean spectra, the generation of ion images, and the construction of analysis-ready feature matrices for downstream comparison. This chapter also introduces the spatial chaos score, which is useful for identifying ion images of interest on the basis of their spatial organization, together with score-based comparisons between control and condition groups for feature prioritization. In *deltamsi*, these operations form the core processing layer implemented through the object MSICube.

A key design choice in *deltamsi* is to represent MSI data using AnnData, an annotated matrix format originally developed for single-cell data[26]. In this representation, MSI pixels (or aggregated spatial units) correspond to observations (obs), *m/z* features correspond to variables (var), and spatial coordinates are stored in multidimensional fields such as obsm["spatial"]. This is the same abstraction used in Scanpy for single-cell transcriptomics[27] and in Squidpy for spatial omics[28]. Using AnnData is advantageous for MSI because both MSI and transcriptomics workflows rely on large feature matrices coupled to structured metadata and downstream multivariate analysis. This choice makes *deltamsi* interoperable with the broader Python single-cell and spatial ecosystem while preserving MSI-specific operations such as *m/z* handling, tolerance-

based extraction, ion-image generation, and score computation. Processing traces and derived results can be stored directly in the AnnData object, for example in uns, layers, varm, and varp, enabling the full analysis state to be serialized and reloaded reproducibly.

Within *deltamsi*, MSI-specific logic is centralized in the MSICube object, which converts the raw imzML files to a processed, metadata-rich AnnData object and exposes consistent interfaces for preprocessing, visualization, scoring, and downstream analysis. Other major concepts and functions of the package are developed in the following chapters of this thesis. Mass recalibration and mass-shift correction are described in Chapter 3, whereas Kendrick-space analysis and annotation, as well as graph-based mass clustering, are described in Chapter 4. Thus, Chapter 2 introduces the conceptual and computational foundation of *deltamsi*, while later chapters expand on specialized modules built on top of this core workflow.

The main functions of the package are presented in Chapter 2, Supplementary Information S2, which serves as the primary function-description document for *deltamsi* and details the principal methods implemented in the package, with particular emphasis on the MSICube workflow and its associated processing modules. In practice, *deltamsi* stores not only the main intensity matrix, but also derived matrices such as ion–ion similarity measures, including cosine colocalization, as well as graph representations in structured slots of the AnnData object, enabling memory-aware workflows through sparse and top-k storage strategies when needed. *deltamsi* is distributed as open-source software and is publicly available from the project repository on GitHub (https://github.com/LaRoccaRaphael/deltamsi/).

2.5 Conclusion and transition to other chapters

MSI experiments typically generate a very large number of molecular features, making efficient data processing essential before biological interpretation can begin. The objective of preprocessing is not only to detect and extract signals, but also to reduce the complexity of the dataset and prioritize a subset of ions for detailed manual investigation. This prioritization relies on their spatial distribution, their variation across experimental conditions, and their potential biological relevance. Feature prioritization, therefore, represents a key intermediate step between raw MSI acquisition and downstream molecular interpretation.

In this chapter, we developed a general MSI preprocessing and analysis workflow, implemented in the publicly available package *deltamsi*. This workflow provides a structured framework for transforming raw MSI data into analysis-ready features and for systematically comparing experimental conditions. Its application to a natural product discovery case study showed that the combined use of DSF and ISF can efficiently highlight ions associated with specific bacterial growth conditions, thereby facilitating the selection of molecular features of potential biological interest.

Once ions of interest have been prioritized, the next step is to interpret their *m/z* values and propose candidate molecular identities through matching against databases of known compounds. This step depends critically on mass accuracy and, consequently, on appropriate recalibration of the MSI data. Although recalibration was already applied in this chapter to a subset of selected ions, it is addressed more systematically in Chapter 3, where we describe the recalibration strategy in detail and examine its impact on database matching and molecular annotation.

The present analysis also showed that, even after ion selection, many detected *m/z* values do not correspond to independent molecular entities. Instead, several features may originate from the same underlying molecule through adduct formation, isotopic

patterns, *in vivo* degradation processes or in-source fragmentation. To avoid overcounting such related signals, it is therefore necessary to dereplicate them by grouping features that arise from a common molecular source. Dereplication here refers strictly to recognizing and clustering redundant *m/z* values into chemical families, not to subsequent intensity aggregation (averaging, summing, or otherwise combining the corresponding ion images), which is an optional downstream step once the groups are defined. Chapters 4 and 5 address dereplication by presenting approaches to automatically identify and cluster related *m/z* values, thereby simplifying downstream biological interpretation.

Finally, once molecular identities have been assigned, their chemical annotation can be integrated with their spatial distribution to provide a richer understanding of their biological roles. This combined perspective makes it possible to determine where a molecule is produced, under which conditions it appears, and how its localization relates to specific cell types or microenvironments. In Chapter 6, the workflow developed here will be applied to cell-culture MSI to identify lipid ions and interpret their spatial distribution in terms of cell-type-specific lipid composition.

REFERENCES:

(1) Watrous, J. D.; Dorrestein, P. C. Imaging Mass Spectrometry in Microbiology. *Nat. Rev. Microbiol.* **2011**, *9* (9), 683–694. https://doi.org/10.1038/nrmicro2634.

(2) Dunham, S. J. B.; Ellis, J. F.; Li, B.; Sweedler, J. V. Mass Spectrometry Imaging of Complex Microbial Communities. *Acc. Chem. Res.* **2017**, *50* (1), 96–104. https://doi.org/10.1021/acs.accounts.6b00503.

(3) Yang, J. Y.; Phelan, V. V.; Simkovsky, R.; Watrous, J. D.; Trial, R. M.; Fleming, T. C.; Wenter, R.; Moore, B. S.; Golden, S. S.; Pogliano, K.; Dorrestein, P. C. Primer on Agar-Based Microbial Imaging Mass Spectrometry. *J. Bacteriol.* **2012**, *194* (22), 6023–6028. https://doi.org/10.1128/JB.00823-12.

(4) Baquer, G.; Sementé, L.; Ràfols, P.; Martín-Saiz, L.; Bookmeyer, C.; Fernández, J. A.; Correig, X.; García-Altares, M. rMSIfragment: Improving MALDI-MSI Lipidomics through Automated in-Source Fragment Annotation. *J. Cheminformatics* **2023**, *15* (1), 80. https://doi.org/10.1186/s13321-023-00756-2.

(5) Baquer, G.; Sementé, L.; García-Altares, M.; Lee, Y. J.; Chaurand, P.; Correig, X.; Ràfols, P. rMSIcleanup: An Open-Source Tool for Matrix-Related Peak Annotation in Mass Spectrometry Imaging and Its Application to Silver-Assisted Laser Desorption/Ionization. *J. Cheminformatics* **2020**, *12* (1), 45. https://doi.org/10.1186/s13321-020-00449-0.

(6) Hu, H.; Laskin, J. Emerging Computational Methods in Mass Spectrometry Imaging. *Adv. Sci.* **2022**, *9* (34), 2203339. https://doi.org/10.1002/advs.202203339.

(7) Buchberger, A. R.; DeLaney, K.; Johnson, J.; Li, L. Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights. *Anal. Chem.* **2018**, *90* (1), 240–265. https://doi.org/10.1021/acs.analchem.7b04733.

(8) Carreira, R. J.; Shyti, R.; Balluff, B.; Abdelmoula, W. M.; Van Heiningen, S. H.; Van Zeijl, R. J.; Dijkstra, J.; Ferrari, M. D.; Tolner, E. A.; McDonnell, L. A.; Van Den Maagdenberg, A. M. J. M. Large-Scale Mass Spectrometry Imaging Investigation of Consequences of Cortical Spreading Depression in a Transgenic Mouse Model of Migraine. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (6), 853–861. https://doi.org/10.1007/s13361-015-1136-8.

(9) Alexandrov, T. MALDI Imaging Mass Spectrometry: Statistical Data Analysis and Current Computational Challenges. *BMC Bioinformatics* **2012**, *13* (S16), S11. https://doi.org/10.1186/1471-2105-13-S16-S11.

(10) Römpp, A.; Schramm, T.; Hester, A.; Klinkert, I.; Both, J.-P.; Heeren, R. M. A.; Stöckli, M.; Spengler, B. imzML: Imaging Mass Spectrometry Markup Language: A Common Data Format for Mass Spectrometry Imaging. In *Data Mining in Proteomics*; Hamacher, M., Eisenacher, M., Stephan, C., Eds.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2011; Vol. 696, pp 205–224. https://doi.org/10.1007/978-1-60761-987-1_12.

(11) Gibb, S.; Strimmer, K. MALDIquant: A Versatile R Package for the Analysis of Mass Spectrometry Data. *Bioinformatics* **2012**, *28* (17), 2270–2271. https://doi.org/10.1093/bioinformatics/bts447.

(12) Deininger, S.-O.; Cornett, D. S.; Paape, R.; Becker, M.; Pineau, C.; Rauser, S.; Walch, A.; Wolski, E. Normalization in MALDI-TOF Imaging Datasets of Proteins: Practical Considerations. *Anal. Bioanal. Chem.* **2011**, *401* (1), 167–181. https://doi.org/10.1007/s00216-011-4929-z.

(13) Palmer, A.; Phapale, P.; Chernyavsky, I.; Lavigne, R.; Fay, D.; Tarasov, A.; Kovalev, V.; Fuchser, J.; Nikolenko, S.; Pineau, C.; Becker, M.; Alexandrov, T. FDR-Controlled Metabolite Annotation for High-Resolution Imaging Mass Spectrometry. *Nat. Methods* **2017**, *14* (1), 57–60. https://doi.org/10.1038/nmeth.4072.

(14) Römpp, A.; Schramm, T.; Hester, A.; Klinkert, I.; Both, J.-P.; Heeren, R. M. A.; Stöckli, M.; Spengler, B. imzML: Imaging Mass Spectrometry Markup Language: A Common Data Format for Mass Spectrometry Imaging. In *Data Mining in Proteomics*; Hamacher, M., Eisenacher, M., Stephan, C., Eds.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2011; Vol. 696, pp 205–224. https://doi.org/10.1007/978-1-60761-987-1_12.

(15) Bokhart, M. T.; Nazari, M.; Garrard, K. P.; Muddiman, D. C. MSiReader v1.0: Evolving Open-Source Mass Spectrometry Imaging Software for Targeted and Untargeted Analyses. *J. Am. Soc. Mass Spectrom.* **2018**, *29* (1), 8–16. https://doi.org/10.1007/s13361-017-1809-6.

(16) Bemis, K. D.; Harry, A.; Eberlin, L. S.; Ferreira, C.; Van De Ven, S. M.; Mallick, P.; Stolowitz, M.; Vitek, O. *Cardinal* : An R Package for Statistical Analysis of Mass Spectrometry-Based Imaging Experiments. *Bioinformatics* **2015**, *31* (14), 2418–2420. https://doi.org/10.1093/bioinformatics/btv146.

(17) Gibb, S.; Strimmer, K. MALDIquant: A Versatile R Package for the Analysis of Mass Spectrometry Data. *Bioinformatics* **2012**, *28* (17), 2270–2271. https://doi.org/10.1093/bioinformatics/bts447.

(18) Föll, M. C.; Moritz, L.; Wollmann, T.; Stillger, M. N.; Vockert, N.; Werner, M.; Bronsert, P.; Rohr, K.; Grüning, B. A.; Schilling, O. Accessible and Reproducible Mass Spectrometry Imaging Data Analysis in Galaxy. *GigaScience* **2019**, *8* (12), giz143. https://doi.org/10.1093/gigascience/giz143.

(19) Race, A. M.; Palmer, A. D.; Dexter, A.; Steven, R. T.; Styles, I. B.; Bunch, J. SpectralAnalysis: Software for the Masses. *Anal. Chem.* **2016**, *88* (19), 9451–9458. https://doi.org/10.1021/acs.analchem.6b01643.

(20) Ràfols, P.; Heijs, B.; Del Castillo, E.; Yanes, O.; McDonnell, L. A.; Brezmes, J.; Pérez-Taboada, I.; Vallejo, M.; García-Altares, M.; Correig, X. rMSIproc: An R Package for Mass Spectrometry Imaging Data Processing. *Bioinformatics* **2020**, *36* (11), 3618–3619. https://doi.org/10.1093/bioinformatics/btaa142.

(21) Bemis, K. A.; Vitek, O. Matter: An R Package for Rapid Prototyping with Larger-than-Memory Datasets on Disk. *Bioinformatics* **2017**, *33* (19), 3142–3144. https://doi.org/10.1093/bioinformatics/btx392.

(22) Ràfols, P.; Vilalta, D.; Brezmes, J.; Cañellas, N.; Del Castillo, E.; Yanes, O.; Ramírez, N.; Correig, X. Signal Preprocessing, Multivariate Analysis and Software Tools for MA(LDI)-TOF Mass Spectrometry Imaging for Biological Applications. *Mass Spectrom. Rev.* **2018**, *37* (3), 281–306. https://doi.org/10.1002/mas.21527.

(23) Palmer, A.; Phapale, P.; Chernyavsky, I.; Lavigne, R.; Fay, D.; Tarasov, A.; Kovalev, V.; Fuchser, J.; Nikolenko, S.; Pineau, C.; Becker, M.; Alexandrov, T. FDR-Controlled Metabolite Annotation for High-Resolution Imaging Mass Spectrometry. *Nat. Methods* **2017**, *14* (1), 57–60. https://doi.org/10.1038/nmeth.4072.

(24) Kune, C.; McCann, A.; Raphaël, L. R.; Arias, A. A.; Tiquet, M.; Van Kruining, D.; Martinez, P. M.; Ongena, M.; Eppe, G.; Quinton, L.; Far, J.; De Pauw, E. Rapid Visualization of Chemically Related Compounds Using Kendrick Mass Defect As a Filter in Mass Spectrometry Imaging. *Anal. Chem.* **2019**, *91* (20), 13112–13118. https://doi.org/10.1021/acs.analchem.9b03333.

(25) Baquer, G.; Sementé, L.; Ràfols, P.; Martín-Saiz, L.; Bookmeyer, C.; Fernández, J. A.; Correig, X.; García-Altares, M. rMSIfragment: Improving MALDI-MSI Lipidomics through Automated in-Source Fragment Annotation. *J. Cheminformatics* **2023**, *15* (1), 80. https://doi.org/10.1186/s13321-023-00756-2.

(26) Virshup, I.; Rybakov, S.; Theis, F. J.; Angerer, P.; Wolf, F. A. Anndata: Access and Store Annotated Datamatrices. *J. Open Source Softw.* **2024**, *9* (101), 4371. https://doi.org/10.21105/joss.04371.

(27) Wolf, F. A.; Angerer, P.; Theis, F. J. SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis. *Genome Biol.* **2018**, *19* (1), 15. https://doi.org/10.1186/s13059-017-1382-0.
(28) Palla, G.; Spitzer, H.; Klein, M.; Fischer, D.; Schaar, A. C.; Kuemmerle, L. B.; Rybakov, S.; Ibarra, I. L.; Holmberg, O.; Virshup, I.; Lotfollahi, M.; Richter, S.; Theis, F. J. Squidpy: A Scalable Framework for Spatial Omics Analysis. *Nat. Methods* **2022**, *19* (2), 171–178. https://doi.org/10.1038/s41592-021-01358-2.

CHAPTER THREE

***

Mass Shift in Mass Spectrometry Imaging, Data Interpretation and Recalibration

Chapter 3: Contents

3.1 Problem of mass shift in mass spectrometry imaging 89

3.1.1 Impact of mass shift 89

3.1.2 Post-acquisition and pre-acquisition approaches to correct mass shift 90

3.1.3 Context and scientific production 91

3.2 Adaptive pixel mass recalibration for mass spectrometry imaging based on locally endogenous biological signals 97

3.2.1 Abstract 97

3.2.2 Introduction 98

3.2.3 Experimental Section 101

3.2.3.1 Materials 101

3.2.3.2 Methods 102

3.2.4 Results and Discussion 110

3.2.4.1 Recalibration Impact on Data Analysis 110

3.2.4.2 Impact of Recalibration on the Number of Molecular Annotations 112

3.2.4.3 FDR and Tolerance Selection 114

3.2.4.4 Evaluation on a Large Set of Public MSI Data Sets 115

3.2.5 Conclusions 120

3.2.6 Associated Content and Acknowledgments 122

3.1 Problem of mass shift in mass spectrometry imaging

3.1.1 Impact of mass shift

The preceding chapter introduced the processing of Mass Spectrometry Imaging (MSI) data, aiming to facilitate a comprehensive understanding of the majority of signals. Regardless of the chosen processing methods, a fundamental prerequisite for MSI data processing lies in the alignment and calibration of peaks within spectra. The perfect case being that ions detected across spectra/pixel exhibit the same *m/z* (mass-to-charge ratio) values and correspond to the theoretical *m/z* values of the ions. This ideal situation, however, is never met for several reasons.

Indeed, in both conventional and high-resolution mass spectrometry, a common issue is mass shifts. Mass shifts are characterized by deviations in the detected *m/z* values of ions from the anticipated/theoretical values. In imaging, these deviations exhibit inconsistent, unpredictable variations from one spectrum to another. When these deviations cause peaks of the same ion to fall outside the accepted mass tolerance range, it directly impacts data interpretation by affecting ion annotation and ion image quality.

In a single spectrum, mass shifts cause experimental peaks to deviate from their theoretical *m/z* values. If the identification tolerance window remains unchanged, peaks that shift beyond it are missed, and the number of correct identifications drops. Alternatively, if we decrease the *m/z* tolerance to compensate for them, it erodes the confidence associated with these identifications and can ultimately lead to misinterpretations. In the context of Mass Spectrometry Imaging (MSI), it is customary to detect all peaks by computing the average spectrum derived from the image. However, when mass shifts are present, identifications based on the average spectrum suffer from artificially degraded resolution: each peak is broadened by the pixel-to-pixel spread of its *m/z*, and closely spaced peaks can merge into a single feature. This thereby

diminishes the ability to make distinct identifications and, consequently, decreases the overall number of successful identifications.

Then, the presence of larger peaks within the average spectrum can significantly influence the generation of ion images. Initially, if the *m/z* tolerance is not sufficiently broad to encompass the entire deviation range, certain regions of the ion image may show no detected signals (see Chapter 2). This affects not only manual interpretation of the image but also data-processing-based interpretation, including segmentation algorithms. Such limitations can have a cascading effect. For instance, segmentation algorithms may primarily identify regions that share similar mass shifts rather than regions that differ in molecular composition (i.e., in the set of detected ions and their relative intensities), as was initially expected.

3.1.2 Post-acquisition and pre-acquisition approaches to correct mass shift

Mass shifts in imaging can be addressed by strategies applied either before or after data acquisition.

In the post-acquisition approach, the image is already acquired and processed, allowing identification of peaks corresponding to known ions (analytes or internal calibrants). Subsequently, this information enables estimation of mass shifts within each spectrum, thereby facilitating recalibration. In this chapter, we discuss the issues associated with this method and introduce a novel, automated solution for mass shift estimation and correction.

Pre-acquisition methods can be applied when the root cause of a mass shift has been identified, enabling mitigation of its impact by addressing the underlying factors. Various sources of mass shifts can be identified, such as the intrinsic performances of the analyzer, the sample itself (composition), its preparation (thickness, matrix deposition), and the calibration method. For instance, images obtained using an FT-ICR

mass analyzer suffer from mass shifts often attributed to space-charge effects occurring within the ICR cell. In contrast, images acquired with a MALDI-TOF instrument show that deviations in *m/z* are mostly linked to the sample topography (specifically, variations in thickness across the sample), which directly influences the kinetic energy imparted to the ions and their time-of-flight, and consequently results in deviations in *m/z* measurements.

3.1.3 Context and scientific production

In this chapter, our primary focus is the post-acquisition recalibration method developed in this work, along with its evaluation across a range of diverse datasets.

It's worth noting that this algorithm maintains a neutral stance regarding the specific mass analyzer in use and doesn't differentiate between recalibration functions based on the type of mass analyzer. Instead, the algorithm operates on the premise that mass shifts are primarily constant with *m/z* or exhibit a linear correlation with m/z values, acknowledging that this assumption may limit its effectiveness when recalibrating datasets afflicted by substantial initial mass shifts. However, it is important to highlight that the simplicity of the algorithm's parameters and its practical application often yield satisfactory recalibration results across diverse MSIs.

The identifications facilitated by these methods provide enhanced confidence in assigning molecular formulas to the detected ions. This is primarily achieved by reducing the *m/z*-based identification mass tolerance after recalibration, thereby improving the accuracy of attributions. The work proposed in this chapter has been published in Analytical Chemistry:

"Adaptive pixel mass recalibration for mass spectrometry imaging based on locally endogenous biological signals," written by Raphaël La Rocca, Christopher Kune, Mathieu Tiquet, Lachlan Stuart, Gauthier Eppe, Theodore Alexandrov, Edwin De Pauw, and Loïc Quinton, and published by Analytical Chemistry, ***2021****; 93, 8, 4066–4074.*

analytical chemistry

pubs.acs.org/ac Article

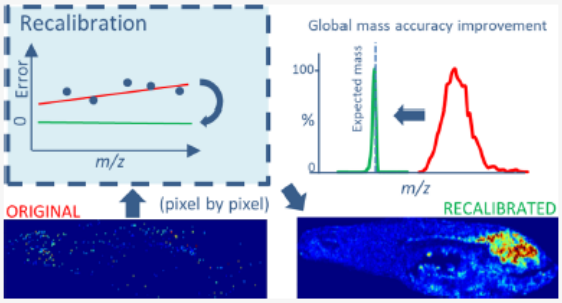
Adaptive Pixel Mass Recalibration for Mass Spectrometry Imaging Based on Locally Endogenous Biological Signals

Raphaël La Rocca,* Christopher Kune, Mathieu Tiquet, Lachlan Stuart, Gauthier Eppe, Theodore Alexandrov, Edwin De Pauw, and Loïc Quinton

Cite This: https://dx.doi.org/10.1021/acs.analchem.0c05071 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Mass spectrometry imaging (MSI) is a powerful and convenient method for revealing the spatial chemical composition of different biological samples. Molecular annotation of the detected signals is only possible if a high mass accuracy is maintained over the entire image and the m/z range. However, the change in the number of ions from pixel-to-pixel of the biological samples could lead to small fluctuations in the detected m/z-values, called mass shift. The use of internal calibration is known to offer the best solution to avoid, or at least to reduce, mass shifts. Their "a priori" selection for a global MSI acquisition is prone to false positive detection and therefore to poor recalibration. To fill this gap, this work describes an algorithm that recalibrates each spectrum individually by estimating its mass shift with the help of a list of pixel-specific internal calibrating ions, automatically generated in a data-adaptive manner (https://github.com/LaRoccaRaphael/MSI_recalibration). Through a practical example, we applied the methodology to a zebrafish whole-body section acquired at a high mass resolution to demonstrate the impact of mass shift on data analysis and the capability of our algorithm to recalibrate MSI data. In addition, we illustrate the broad applicability of the method by recalibrating 31 different public MSI data sets from METASPACE from various samples and types of MSI and show that our recalibration significantly increases the numbers of METASPACE annotations (gaining from 20 up to 400 additional annotations), particularly the high-confidence annotations with a low false discovery rate.



1. INTRODUCTION

In biology and medicine, the in situ determination of the molecular environment is of prime importance to understand biological processes and pathology evolution.[1] This growing interest pushes the development of analytical methods that correlate spatial distribution with the detection of different biological molecules such as peptides,[2] metabolites, and lipids.[3] Mass spectrometry imaging (MSI) has been demonstrated as particularly powerful as it can rapidly reveal, in an untargeted manner, a wide range of compounds present in small amounts in biological samples as varied as whole-body sections,[4,5] tissue sections, bacteria colonies,[6] plants,[7] or again single cells.[8,9] MSI is a particular application of mass spectrometry (MS) in which spectra are recorded, usually by matrix-assisted laser desorption ionization (MALDI) or to a lesser extent desorption electrospray ionisation[10] and secondary ion mass spectrometry,[11] at different positions, called pixels, over a sample forming a two-dimensional (2D) image. A pixel is then a spectrum, identified by its coordinates (x, y), containing m/z values (channels) and their intensities.[12,13]

While high mass resolution is required to distinguish isobaric and quasi-isobaric compounds in complex mixtures and to exploit isotope signatures, high accuracy of m/z measurement is essential for reliable compound identification. For this reason, high-resolution mass spectrometry (HRMS), combining both high resolution and mass accuracy, is encouraged for the annotation of molecular signatures of biological samples.[14] However, the MSI data acquired by HRMS have shown that they suffer from an inconsistent variation of measured m/z values, from pixel to pixel. There are various reasons for mass-shift phenomenon in MSI, most of which depend on the design of the mass analyzer. For example, it has been shown that the number of ions in a Fourier-transform ion cyclotron resonance cell (FT-ICR) is correlated with mass shifts.[15] The analysis of MSI images is based on a so-called average spectrum, representing all ions detected in the image, that is, by summing each spectrum in each pixel. The pixel-to-pixel

Received: December 3, 2020
Accepted: January 28, 2021

© XXXX American Chemical Society

https://dx.doi.org/10.1021/acs.analchem.0c05071
Anal. Chem. XXXX, XXX, XXX−XXX

Moreover, the developed MSI pre-processing programs and the post-acquisition recalibration method described in this chapter enabled two additional publications. The first one concerned mass shift occurring during image acquisition by a MALDI-FT-ICR. In this work, we proposed an analytical workflow that monitors the total ion current to mitigate pixel-to-pixel *m/z* shifts. Adjustment of the laser parameters is proposed to maintain high spectral resolution and mass accuracy measurement within the instrument specifications during MSI analyses. The optimized method has been successfully employed in replicates to perform high-quality MALDI MS images at resolving power above 1,000,000 in the lipid mass range across the whole image for superconducting magnets of 7T and 9.4T using 1ω and 2ω detection. We developed a program that allows the estimation of mass shift of images, and the evaluation of the proposed method. This work has also been published in Analytical Chemistry:

"FT-ICR Mass Spectrometry Imaging at Extreme Mass Resolving Power Using a Dynamically Harmonized ICR Cell with 1ω or 2ω Detection" written by Mathieu Tiquet, **Raphaël La Rocca**, Stefan Kirnbauer, Samuele Zoratto, Daan Van Kruining, Loïc Quinton, Gauthier Eppe, Pilar Martinez-Martinez, Martina Marchetti-Deschmann, Edwin De Pauw, and Johann Far, and published by Analytical Chemistry, in **2022**, 94, 26, 9316-9326.

The complete publication is available in the 'Associated Papers' section of the thesis, Chapter 3, A3.

analytical chemistry

pubs.acs.org/ac Article

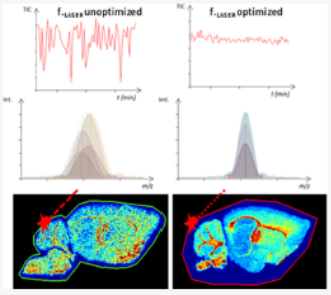
FT-ICR Mass Spectrometry Imaging at Extreme Mass Resolving Power Using a Dynamically Harmonized ICR Cell with 1ω or 2ω Detection

Mathieu Tiquet, Raphaël La Rocca, Stefan Kirnbauer, Samuele Zoratto, Daan Van Kruining, Loïc Quinton, Gauthier Eppe, Pilar Martinez-Martinez, Martina Marchetti-Deschmann, Edwin De Pauw, and Johann Far*

Cite This: *Anal. Chem.* 2022, 94, 9316−9326 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: MALDI mass spectrometry imaging (MALDI MSI) is a powerful analytical method for achieving 2D localization of compounds from thin sections of typically but not exclusively biological samples. The dynamically harmonized ICR cell (ParaCell) was recently introduced to achieve extreme spectral resolution capable of providing the isotopic fine structure of ions detected in complex samples. The latest improvement in the ICR technology also includes 2ω detection, which significantly reduces the transient time while preserving the nominal mass resolving power of the ICR cell. High-resolution MS images acquired on FT-ICR instruments equipped with 7T and 9.4T superconducting magnets and the dynamically harmonized ICR cell operating at suboptimal parameters suffered severely from the pixel-to-pixel shifting of m/z peaks due to space-charge effects. The resulting profile average mass spectra have depreciated mass measurement accuracy and mass resolving power under the instrument specifications that affect the confidence level of the identified ions. Here, we propose an analytical workflow based on the monitoring of the total ion current to restrain the pixel-to-pixel m/z shift. Adjustment of the laser parameters is proposed to maintain high spectral resolution and mass accuracy measurement within the instrument specifications during MSI analyses. The optimized method has been successfully employed in replicates to perform high-quality MALDI MS images at resolving power (FWHM) above 1,000,000 in the lipid mass range across the whole image for superconducting magnets of 7T and 9.4T using 1 and 2ω detection. Our data also compare favorably with MALDI MSI experiments performed on higher-magnetic-field superconducting magnets, including the 21T MALDI FT-ICR prototype instrument of the NHMFL group at Tallahassee, Florida.



1. INTRODUCTION

Matrix-assisted laser desorption/ionization mass spectrometry imaging (MALDI MSI) has emerged as a label-free analytical method for monitoring the relative abundance (despite severe limitations due to suppression effects) and spatial distribution of a wide variety of analytes, especially for biological samples.[1−3] To properly distinguish isobaric compounds[4] inherent to the complexity of biological samples, a high resolving power at full width at half-maximum ($R.P._{FWHM}$ > 300,000 at 400m/z) and a reliable mass measurement accuracy (MMA) are required in the absence of an upstream separation method (such as ion mobility). These performances are commonly achieved by a Fourier transform mass analyzer such as Fourier transform-ion cyclotron resonance (FT-ICR).[5] The Bruker dual ion source ESI/MALDI FT-ICR (solariX and scimaX) is a hybrid instrument equipped with multipoles, a quadrupole, and a collision cell for precursor ion selection and fragmentation. Ions produced by electrospray (ESI) and MALDI are also accumulated in the multipole region to prepare the ion packet to be introduced into the ICR cell. Consequently, almost any combination of MALDI laser settings is compatible with any transient time for mass spectra acquisition. Recent developments introduced by Nikolaev et al.[6−8] led to the dynamically harmonized ICR cell commercialized by Bruker in the solariX XR and scimaX XR brand FT-ICR mass spectrometers under the name ParaCell. This new cell offers the highest mass resolving power (R.P.) currently achievable for such instruments[9] (around 1,000,000 in the lipid mass range in broadband mode), and mass measurement accuracy typically is in the sub-ppm range. These improvements drastically increase the confidence level in the determination of the empiric formula of precursor ions, especially when including the fine isotopic structure.[10]

Received: February 14, 2022
Accepted: May 9, 2022
Published: May 23, 2022



ACS Publications © 2022 The Authors. Published by American Chemical Society

https://doi.org/10.1021/acs.analchem.2c00754
Anal. Chem. 2022, 94, 9316−9326

Chapter 3: Mass Shift in Mass Spectrometry Imaging, Data Interpretation and Recalibration

In the second publication, we report a critical analysis and recommendations to mitigate sources of mass shift in MALDI-TOF and MALDI-FT-ICR instruments. The spatial distribution of mass shifts reveals their origins across different experimental settings. We reported that, for TOF instruments, the mass shift can be reduced by optimizing the sample preparation (*i.e.,* improving the sample flatness and the homogeneity of the matrix deposition). The main factor influencing mass shift in MALDI-FT-ICR, in contrast, is the variation in the total ion current (TIC), which is directly linked to the number of ions measured in the cell. Maintaining the TIC below a given value and thereby avoiding Coulomb expansion during acquisition facilitates a drastic reduction in the mass shift in these MS analyzers. Finally, post-processing recalibration, detailed in Chapter 3, allows for further reduction of the mass shifts, regardless of whether pre-acquisition recommendations were applied. However, the performance of this algorithm depends on the initial quality of the MSI data, as better recalibrations are obtained only if lower-level mass shifts were initially present. Consequently, it is recommended to control the mass shift at every step of the MSI workflow to achieve optimal MSI post-processing recalibration.

"*Mass shift in mass spectrometry imaging: comprehensive analysis and practical corrective workflow*" written by Andréa McCann, Sophie Rappe, **Raphaël La Rocca**, Mathieu Tiquet, Loïc Quinton, Gauthier Eppe, Johann Far, Edwin De Pauw, Christopher Kune, and published by Analytical and Bioanalytical Chemistry, **2021**; 413, 2831-2844.

The complete publication is available in the 'Associated Papers' section of the thesis, Chapter 3, A2.

Analytical and Bioanalytical Chemistry
https://doi.org/10.1007/s00216-021-03174-1

RESEARCH PAPER

Mass shift in mass spectrometry imaging: comprehensive analysis and practical corrective workflow

Andréa McCann[1] · Sophie Rappe[1] · Raphaël La Rocca[1] · Mathieu Tiquet[1] · Loïc Quinton[1] · Gauthier Eppe[1] · Johann Far[1] · Edwin De Pauw[1] · Christopher Kune[1]

Received: 14 October 2020 / Revised: 22 December 2020 / Accepted: 13 January 2021
© Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
MALDI mass spectrometry imaging (MSI) allows the mapping and the tentative identification of compounds based on their *m/z* value. In typical MSI, a spectrum is taken at incremental 2D coordinates (pixels) across a sample surface. Single pixel mass spectra show the resolving power of the mass analyzer. Mass shift, i.e., variations of the *m/z* of the same ion(s), may occur from one pixel to another. The superposition of shifted masses from individual pixels peaks apparently degrades the resolution and the mass accuracy in the average spectrum. This leads to low confidence annotations and biased localization in the image. Besides the intrinsic performances of the analyzer, the sample properties (local composition, thickness, matrix deposition) and the calibration method are sources of mass shift. Here, we report a critical analysis and recommendations to mitigate these sources of mass shift. Mass shift 2D distributions were mapped to illustrate its effect and explore systematically its origin. Adapting the sample preparation, carefully selecting the data acquisition settings, and wisely applying post-processing methods (i.e., *m/z* realignment or individual *m/z* recalibration pixel by pixel) are key factors to lower the mass shift and to improve image quality and annotations. A recommended workflow, resulting from a comprehensive analysis, was successfully applied to several complex samples acquired on both MALDI ToF and MALDI FT-ICR instruments.

Keywords Mass spectrometry imaging · FT-ICR · Mass shift · Data recalibration · Time of flight · Mass accuracy

Introduction

Mass spectrometry imaging (MSI) is a powerful mass spectrometry (MS) method to visualize 2D molecular species distribution in a sample. A mass spectrometry image consists of a set of mass spectra obtained from a thin section of the sample moved step-by-step in a regular grid under an ionization source. Each spot, also called pixel, contains a whole mass spectrum. A common method to produce ions at different spots is the use of laser-based ion source, such as laser desorption ionization (LDI), surface-assisted laser desorption ionization (SALDI), and especially matrix-assisted laser desorption ionization (MALDI). Desorption electrospray ionization (DESI) and secondary ion mass spectrometry are also popular ion sources fitting well with mass spectrometry imaging [1]. Less frequently, images are obtained in microscopic mode using a position sensitive detector [2]. In scanning MSI methods, all the individual mass spectra (i.e., from all pixels) are compiled to produce 2D maps that can be reconstructed for each *m/z* [3] or for families of compounds [4].

The use of maximum mass accuracy and sufficient mass resolving power allows a molecular formula to be assigned to each peak, as long as the peak is free of isobaric interferences and detected with a sufficient sensitivity [5], while MS/MS strategies are usually required to resolve atomic connectivity and to identify the ions with more confidence [6–8]. Any additional descriptors, such as the collision cross section (i.e., from ion mobility MS) can help in confirming the identification without requiring MS/MS steps [9–11].

Peak annotation in mass spectrometry imaging will therefore rely primarily on the *m/z* values and isotopic ratios. MSI gains its full power when accurate mass determination at high

Published in the topical collection *Mass Spectrometry Imaging 2.0* with guest editors Shane R. Ellis and Tiffany Porta Siegel.

Andréa McCann and Sophie Rappe are co-first authors of this paper.

Andréa McCann and Sophie Rappe contributed equally to this work.

✉ Christopher Kune
c.kune@uliege.be

[1] Mass Spectrometry Laboratory, MolSys Research Unit, Department of Chemistry, University of Liège, Allée du Six Août, 11 – Quartier Agora, 4000 Liège, Belgium

Published online: 30 January 2021

3.2 Adaptive pixel mass recalibration for mass spectrometry imaging based on locally endogenous biological signals

3.2.1 Abstract

Mass spectrometry imaging (MSI) is a powerful and convenient method for revealing the spatial chemical composition of different biological samples. Molecular annotation of the detected signals is only possible if high mass accuracy is maintained throughout the image and across the entire *m/z* range. However, changes in ion counts from pixel to pixel in biological samples can lead to small fluctuations in measured *m/z* values, known as mass shifts. Internal calibration offers the best solution to avoid, or at least to reduce, mass shifts. Selecting these calibrant ions a priori for global MSI acquisition is prone to false positive detection and, therefore, to poor recalibration. To fill this gap, this work describes an algorithm that recalibrates each spectrum in an MSI dataset by estimating its mass shift using a list of internal calibration ions, automatically generated in a data-adaptive manner (https://github.com/LaRoccaRaphael/MSI_recalibration). Using a practical example, we applied the methodology to a zebrafish whole-body section acquired at high mass resolution to demonstrate the impact of mass shifts on data analysis and the capability of our algorithm to recalibrate MSI data and improve results. In addition, we illustrate the broad applicability of the method by recalibrating 31 different public MSI data sets from METASPACE from various samples and types of MSI and show that our recalibration significantly increases the number of METASPACE annotations (gaining from 20 up to 400 additional annotations), particularly the high-confidence annotations with a low false discovery rate.

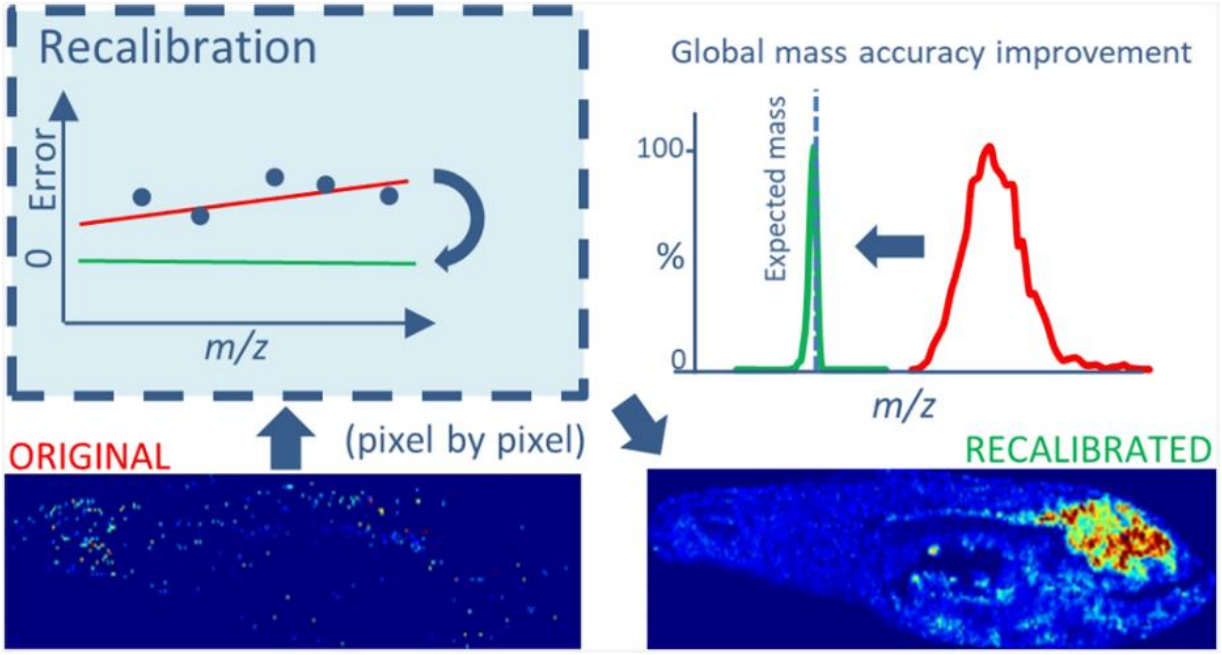


Figure 3.1. *Overview of the recalibration method when applied to zebrafish MSI. Recalibration allows for obtaining better ion images due to a decrease in variations of the peak's m/z.*

3.2.2 Introduction

In biology and medicine, the *in situ* determination of the molecular environment is of prime importance for understanding biological processes and the evolution of pathology[1]. This growing interest pushes the development of analytical methods that correlate spatial distribution with the detection of different biological molecules such as peptides[2], metabolites, and lipids[3]. Mass spectrometry imaging (MSI) has been demonstrated as particularly powerful as it can rapidly reveal, in an untargeted manner, a wide range of compounds present in biological samples as varied as whole-body sections[4,5], tissue sections, bacteria colonies[6], plants[7], or even single cells[8,9]. MSI is a particular application of mass spectrometry (MS) in which spectra are recorded, usually by matrix-assisted laser desorption ionization (MALDI) or, to a lesser extent, desorption electrospray ionisation[10] and secondary ion mass spectrometry[11], at different positions, called pixels, over a sample forming a two-dimensional (2D) image. A pixel is then a spectrum, identified by its coordinates *(x, y),* containing *m/z* values

(channels) and their associated intensities[12,13]. While high mass resolution is required to distinguish isobaric and quasi-isobaric compounds in complex mixtures and to exploit isotope signatures, high accuracy of *m/z* measurement is essential for reliable compound identification. For this reason, high-resolution mass spectrometry (HRMS), combining both high resolution and mass accuracy, is encouraged for the annotation of molecular signatures of biological samples[14]. However, the MSI data acquired by HRMS have shown that they suffer from an inconsistent variation of measured *m/z* values, from pixel to pixel. There are various reasons for mass-shift phenomenon in MSI, most of which depend on the design of the mass analyzer. For example, it has been shown that the number of ions in a Fourier-transform ion cyclotron resonance cell (FT-ICR) is correlated with mass shifts[15]. The analysis of MSI images is based on a so-called average spectrum, representing all ions detected in the image, that is, by summing each spectrum in each pixel. The pixel-to-pixel fluctuation of measured *m/z* reaches up to several ppm and strongly influences the accuracy and resolution of the MSI average spectra[16–18]. It thus has a crucial impact on MSI interpretation. Indeed, mass shifts affect the quality of the results by (i) decreasing the number of identifications, (ii) increasing the number of false identifications, (iii) reducing the confidence of each identification, and (iv) having an impact on the ability to reconstruct the appropriate spatial distributions of specific detected species. It has already been shown that automated annotation of a metabolite for MSI data is critically dependent on *m/z* accuracy and requires the accuracy of at least 3 ppm in *m/z*[19]. In summary, mass shifts strongly weaken the advantages of HRMS instruments for MSI data analysis.

The mass accuracy depends mainly on the quality of the instrument calibration, while the mass resolving power is linked to the mass analyzer device[20]. Instrument calibration is performed by locking the experimental *m/z* signals to their theoretical *m/z* values using a suitable mathematical function (e.g., linear, quadratic, or cubic functions). In general, the MS calibration function is estimated from the signal of a calibrating substance acquired separately from the MSI sample (i.e., external calibration)[21]. This

function is then applied uniformly to each pixel. Since mass-shift is a pixel-dependent effect, the application of a unique MS calibration function cannot reduce this phenomenon in an efficient way[16,22]. Recalibration of MSI data using individual spectrum recalibration functions estimated directly from specific signals in the spectrum (i.e., internal calibration) appears as a method of choice for removing the mass-shift effect[16,23]. The signal of additive molecules[16,18] or biological matrix signals[23,24] can be exploited to perform the MSI recalibration. The use of endogenous molecules has the advantage of avoiding the ion suppression effect generated by the added reference molecules. In this approach, the set of potential calibrating ions has to be known in advance and has to be detected in each pixel to effectively and efficiently reduce the mass-shift contribution[17]. This limits its application to highly homogenous MSI samples whose expected signal is already known (i.e., the signal of the matrix ions or of the ions already identified). Moreover, the application of this method to MSI involving complex spectra increases the chance of selecting inappropriate calibration ions.

Other approaches, such as MS spectra alignment method, have been investigated to counteract the mass-shift phenomenon. However, despite the increase in spectra consistency, alignment does not necessarily correct for mass shifts, contrary to recalibration[25–28].

In this context, this work aims to propose a post-acquisition data-adaptive recalibration methodology to reduce mass shifts in MS data based on a robust and automated selection of adequate calibrating signals in each pixel independently. In this case, the method is applied on MSI by using confidently identified lipids and metabolites from MSI data sets in METASPACE (an open and free access platform) to constitute these calibrating ion lists.

3.2.3 Experimental Section

3.2.3.1 Materials

In this work, two types of data sets have been analyzed: a zebrafish cryo-section and a set of 31 public MSI data sets from METASPACE. First, 12 µ*m* thick slices of 1-month-old zebrafish embedded in gelatin were produced by the Cryostar NX70 (Thermo Scientific), then placed on an indium tin oxide glass slide (Bruker, Bremen, Germany). All manipulations on zebrafish are approved by the Animal Welfare Committee of the University of Liège and are performed according to Belgian federal regulations for animal protection. After 15 min of desiccation, tissue slices were covered by CHCA matrix (97% purity, Sigma-Aldrich, Taufkirchen, Germany) using an automatic sprayer SunCollect System (SunChrom). MSI acquisition was performed on a SolariX XR 9.4T (Bruker) using the automation software FlexImaging 5.0 (Bruker, Bremen, Germany). The acquisition method consists of 400 laser shots per pixel fired at 1000 Hz with the laser power fixed at 70%. The minimum laser focus was employed with a raster width of 60 µm, leading to images ranging from 10 to 15k pixels. The image was converted into centroided imzML[29] format (common data format for MSI) using FlexImaging 5.0. On the other hand, 31 public MSI data sets were selected in METASPACE as representatives for different ion sources (MALDI/DESI), analyzers (Orbitrap/ FT-ICR), polarities (positive/negative), and MALDI matrices (DHB for positive mode, DAN for negative mode, CHCA/Norharmane for both modes). To represent a maximum of the METASPACE samples, we have selected human/mouse samples when possible. In addition, 6 data sets from Waters were included to cover several TOF-based analyzers. All MSIs were downloaded as centroided imzML[29] (see Supporting Information, S1 for more details). The metadata about the sample preparation was, however, only partially provided (see Supporting Information, S1), and the parameters used to convert the raw image into centroided imzML are not known.

3.2.3.2 Method

General Approach

In this work, the term “hit” is used to depict a match between experimental signal *m/z* ($m/z_{exp.}$) and internal calibrating ion *m/z* ($m/z_{calib.}$) within a given mass tolerance from the m/z_{calib}. We will call “mass error” the difference in mass between the m/z_{exp} and its linked $m/z_{calib.}$ (eq 1). A hit is considered true if the experimental signal corresponds to the calibration ion. Otherwise, an erroneous match between the experimental signal and the internal calibrating ion is a false hit. In MSI, the mass spectra are information-rich and often contain on the order of 10^5 peaks per spectrum . Therefore, it is particularly difficult to discriminate true from false hits as multiple signals can be found in the mass range of a given m/z_{calib}. This difficulty is amplified by the presence of mass shifts since higher mass tolerance is required. However, increasing the mass tolerance inevitably increases the number of false hits, decreasing the recalibration performance. Therefore, the algorithm presented in this work optimizes the selection of true hits for recalibrating each pixel individually. We assume that true hits have mass errors directly correlated to the mass shifts. Then, the calculated mass error is used as a criterion for the selection of hits, considering that hits with the most frequent mass errors are predominantly true. These hits are thus selected to fit a linear model of the mass errors according to *m/z*. Finally, the recalibration is performed by removing the estimated errors in every detected *m/z* value.

This algorithm is divided into 5 steps (Figure 3.2). (1) The generation of the list of $m/z_{calib.}$ for the whole MSI data according to similar public MSI data sets from METASPACE (i.e., representing the same kind of biological samples), since we assume them to share metabolites with the sample of interest. (2) Centroid MS spectra are extracted from each pixel. From this step, each pixel is processed individually. (3) The calibrant hits are generated by computing the mass errors between the list of potential internal calibrating ions and the spectrum signals for each pixel. (4) For each pixel individually, the preferential calibrant hits are selected according to their mass errors by discarding

hits with infrequent mass errors. (5) A linear model for predicting the mass shifts based on the preferential hits is then constructed to correct every mass of the spectrum.

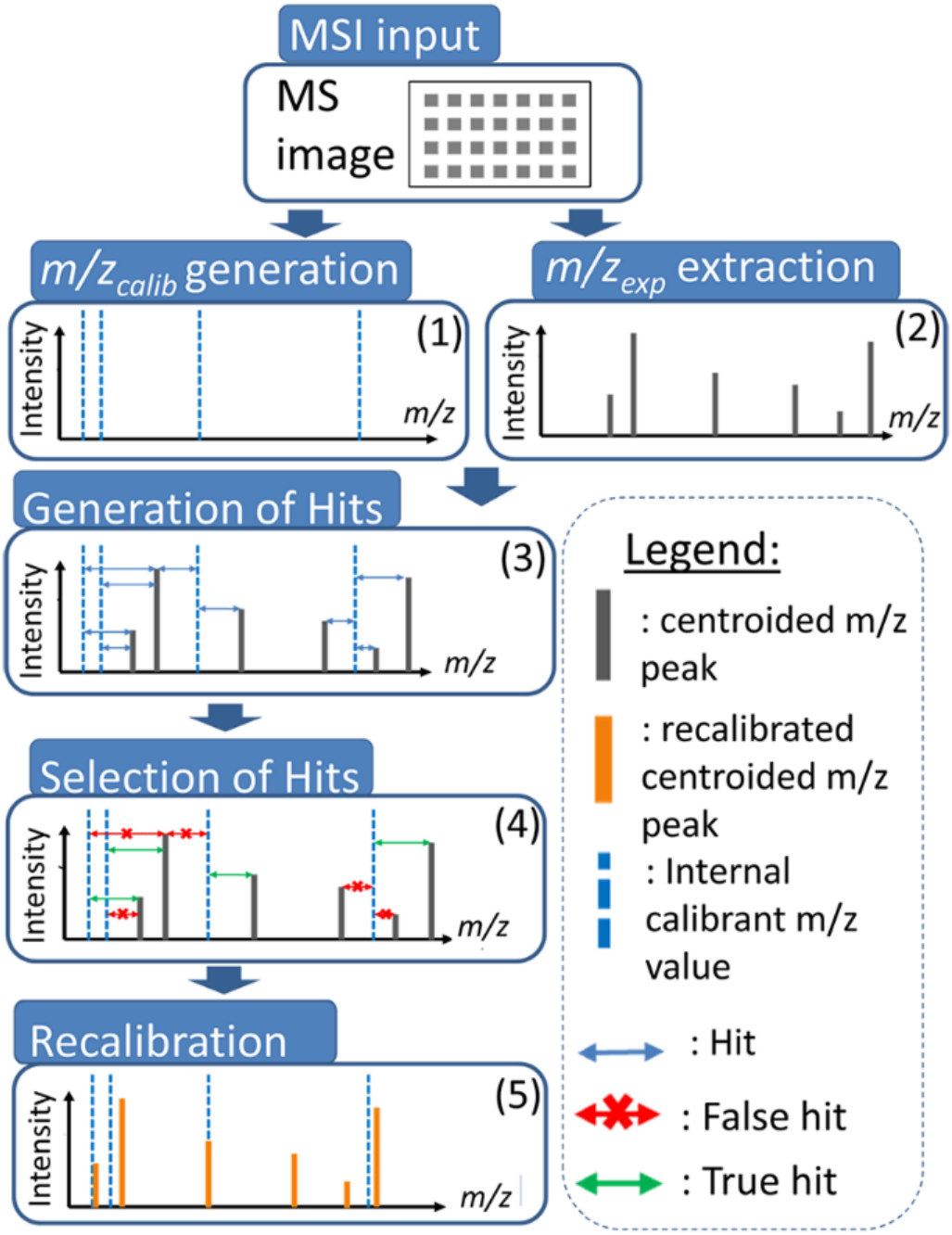


Figure 3.2. *Schematic overview of the five-step recalibration pipeline: (1) m/z_{calib} list generation, (2) m/z_{exp} extraction, (3) hit generation, (4) hit selection, and (5) pixel-wise recalibration. Steps are described in detail in the main text. Symbol meanings are given in the inset legend.*

Generation of the List of Internal Calibrating Ions

Generating the $m/z_{calib.}$ list covering the signals from the sample of interest is a critical step for accurate estimation of mass shifts as it directly influences the collection of true hits. Therefore, for each MSI data subject to recalibration, the most redundant annotations across similar public METASPACE datasets are first retrieved, and the theoretical exact *m/z* values corresponding to these annotations (sum formula + adduct) constitute the m/z_{calib} list. Only these theoretical *m/z* values are passed to the algorithm; the molecular identities themselves are not used in the recalibration procedure, nor are the measured *m/z* values from the source datasets. This ensures that the calibrating *m/z* values are exact and free of any pixel-specific mass shifts inherited from other acquisitions. However, it is not mandatory to use METASPACE to generate the $m/z_{calib.}$ as any list of *m/z* values could work; yet, using METASPACE will result in fewer false hits.

Generating the Hits

The m/z_{exp} list is made of the detected peaks for each pixel (Figure 3.2). The m/z_{exp} list is then compared to the list of $m/z_{calib.}$ within a mass tolerance called T1 in (Figure 3.2). This mass tolerance window is considered large enough to encompass most of the extreme mass shifts in the data and small enough to discard the contribution of non-relevant peaks and isotopes. The mass errors in Da, used to calculate the regression and the error distribution, are calculated for each hit using eq 1 where M_{error}, M_{exact}, and $M_{experiment}$ are, respectively, the mass error, the exact mass (from m/z_{calib}. list), and the experimental mass (from the m/z_{exp}. list).

$$M_{error} = M_{experiment} - M_{exact} \quad (1)$$

M_{error} values are expressed in Da instead of ppm to ensure a linear evolution of M_{error} along the *m/z* axis (see Figure 3.3a).

Selection of Hits

The selection of hits is made independently for each pixel. Since the mass tolerance window for finding true hits T1 (e.g., set at ±0.01 Da for all MSI data) can be large, a preselection of hits is necessary before fitting any model (Figure 3.2). The calculated mass errors of the hits (with eq 1) are used for discarding false hits. As discussed previously, the assumption is that true hits display similar mass errors and the most populated errors should therefore contain a maximum of true hits. A kernel density estimation of the distribution of the errors is estimated by a Gaussian Kernel[30] as illustrated in Figure 3.3a. The hits of interest are finally selected within a certain range from the maximum of this distribution. This range will be called T2 (i.e., set at ±0.002 Da for all MSI data treated here).

For any given calibrant *m/z*, the T1 window encompasses not only the genuine match but also background peaks from the dense biological spectrum that happen to fall nearby. Because these background peaks originate from a chemically diverse set of unrelated species with no systematic mass relationship to the calibrant, the corresponding mass errors lack any preferred value and are spread broadly across T1. Collectively, false hits therefore contribute a smooth, broadly distributed pedestal to the mass-error distribution, uniform to first order, rather than a sharp peak at any specific mass-error value.

True hits, by contrast, all reflect the same underlying physical phenomenon, the pixel-specific mass shift of the analyzer. Although this shift evolves continuously with *m/z* (linearly to first order), T1 is deliberately set to be wider than the total mass-error variation produced by the linear trend across the spectrum's *m/z* range. As a consequence, true hits do not span T1; they occupy only a narrow band whose width is governed by the slope of the trend, and this band appears as a localized peak in the mass-error distribution, superimposed on the smooth false-hit pedestal.

Kernel density estimation captures the bulk of the shift. These two contributions, a smooth false-hit pedestal and a sharp true-hit peak superimposed on it, are exactly the

conditions under which a kernel density estimate can isolate the true-hit population. The KDE peak locates the center of the true-hit band, while the broader false-hit background contributes only a slowly varying offset that does not compete with the peak. The algorithm does not require the false-hit distribution to be strictly uniform; it only requires that the false-hit contribution be smooth on the scale of the true-hit peak width, which is the operational consequence of the chemistry of biological spectra combined with the engineered narrowness of T1. Selecting hits within a narrow window T2 around the KDE maximum discards most of the smoothly distributed false hits while retaining nearly all true hits. This step already provides a coarse estimate of the per-pixel mass shift, essentially the intercept of the linear error-versus-*m/z* relationship, and it does so using a non-parametric method robust to the very high proportion of outliers present in T1.

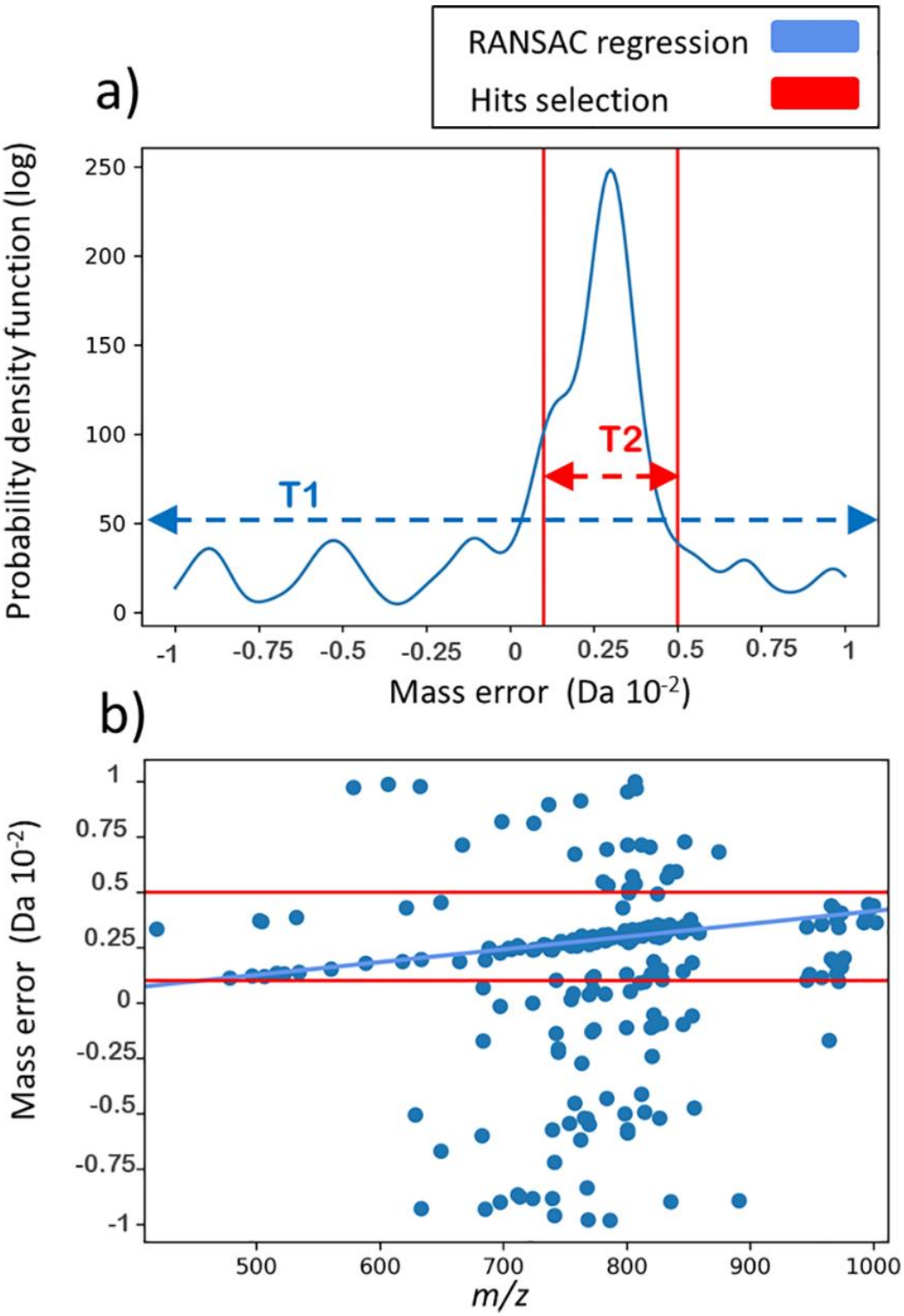


Figure 3.3. *Recalibration algorithm output example from the hits of a single pixel. (a) Selection of hits according to the most frequent errors and (b) mass errors estimation and recalibration.*

Mass Error Estimation and Recalibration

Hits selection reduces the probability of selecting false hits. Within the T2-selected hits, the ratio of true to false hits is now favorable, but the slope of the trend means that the mass error of a true hit still varies with its *m/z*. Unfortunately, the application of simple linear regression functions, such as OLS (ordinary least squares), leads to poor results since false hits are still present due to the complexity of biological signals. Therefore, RANSAC (random sample consensus) algorithm[31] is used for regression as it is more robust to outliers compared to OLS (Figure 3.3b); the MS spectrum (MSI pixel) recalibration is performed by removing the mass errors estimated by the linear fit from each detected *m/z* value. A minimum of 10 hits per pixel is required for recalibration to proceed; pixels with fewer hits are left uncorrected. The threshold ensures that enough true hits remain within T2, after accounting for the residual outlier fraction, to constrain the two parameters of the linear model and provide RANSAC with a meaningful consensus to work with. Below this floor, too few true hits survive for the regression to be stable. This step is made for each pixel of the MSI data (Figure 3.2).

RANSAC's effectiveness degrades sharply when the outlier fraction is very high: random sub-sampling then frequently selects configurations dominated by false hits, producing unreliable models. The KDE/T2 pre-filter is therefore essential; it transforms an intractable regression problem (many more false than true hits) into a tractable one (mostly true hits with a manageable outlier tail).

These two stages correspond to estimating the intercept and slope of the same linear model in sequence. The intercept is extracted first because it dominates the total shift and can be located by a peak-finder that tolerates a very high outlier fraction; the slope is extracted afterward, once enough outliers have been removed for a robust regression to converge. This decomposition also clarifies the role of T1: it must be set wider than the largest plausible mass shift and wider than the slope-induced variation across the *m/z* range, so that the true-hit peak in the KDE is fully contained within T1 rather than

being truncated at its edges, while remaining narrow enough to exclude isotope and adduct contamination that would distort the false-hit pedestal.

Although METASPACE annotations are not infallible, they are FDR-controlled rather than error-free; three-layered mechanisms limit the propagation of annotation errors into the recalibration. First, the cross-dataset redundancy filter requires that a candidate identity recur in several similar datasets, which is unlikely for one-off spurious annotations. Second, only theoretical *m/z* values are transferred, so pixel-level inaccuracies in the source measurements cannot be preserved under the identity. Third, any residual incorrect calibrant that remains after the first two filters generates hits in the experimental spectrum that are not coherent with the actual per-pixel mass shift; these hits behave like any other false hits, contributing to the smooth false-hit pedestal discarded by the T2 window and to the outlier tail rejected by RANSAC. The robustness of the two-stage estimation procedure is therefore exploited not only against random spectral background but also against the low residual rate of incorrect calibrants.

Figure 3.3 highlights the application of our algorithm to a single pixel of the zebrafish MSI data. The selection of hits based on the density estimation of the mass errors is shown in Figure 3.3a; the selected hits are the points located within the red lines. The goal of this step is to filter out as many as possible of those hits that have too high error deviations from true hits to maximize the chance of not including false hits during model estimation. The RANSAC linear model is then estimated on the selected hits. Since many outliers are still present in the selected hits, the use of a robust linear estimator is necessary for detecting only true hits (Figure 3.3b).

The reference implementation of the developed method is available at https://github.com/LaRoccaRaphael/MSI_ recalibration.

The full method is written in the programming language Python with the following libraries: NumPy, Pandas, METASPACE, SciPy, pyimzML, and scikit-learn. The processing

time is mainly attributable to the recalibration part of the method (steps 3-5) which depends on the number of pixels in the image. For example, the recalibration of the imzML of the zebrafish, containing 11,050 pixels, took 845 s on a computer with 16 GB of RAM, 2.6 GHz four-core Intel i7 processor with 8 MB of cache running Ubuntu.

A more detailed version of the method is described in Supporting Information, S2.

3.2.4 Results and Discussion

3.2.4.1 Recalibration Impact on Data Analysis

To evaluate the efficiency of our internal calibration algorithm, the MSI data of a zebrafish slice imaged with a MALDI FT-ICR have been compared before and after recalibration. The observed mass shift is the result of the highly heterogeneous molecular composition of the organs and the tissues of this sample leading to different amounts of generated ions in the FT-ICR cell[15]. Even for high accuracy instruments such as FT- ICR mass analyzer, the amplitude of mass shift is unknown for a given sample. For this reason, the values of T1 and T2 were set after investigating the mass shifts of various pixels by using the plot as depicted in Figure 3.3, and then T1 and T2 values were adopted if an increase in annotations was observed by the quality plots (Figure 3.5). Finally, T1 was fixed at ± 0.01 Da and T2 at ±0.002 Da for this data set. However, too few public MSI data sets on zebrafish in positive mode were found on METASPACE. Therefore, we decided to generate the list of internal calibrating ions from LIPID MAPS database[32] with a selection of glycerophospholipid and sphingolipid ions (Na^+, K^+, and H^+ adducts). Figure 3.4 shows a comparison of the average MS spectra of the image before and after recalibration for two different well-characterized intense and ubiquitous ions[33] (Figure 3.4 A,B) as well as the consequence on the reconstructed images of the ions for different tolerances (Figure 3.4 C-J). The two peaks chosen for this evaluation are the two following phosphatidylcholines $[C_{42}H_{82}NO_8P+Na]^+$ and $[C_{40}H_{80}NO_8P+Na]^+$ (*m/z* 782.5670 and 756.5514, respectively). Those two ions and all

adduct variants were not included in the list of internal calibrating ions. In Supporting Information S4, the *deltamsi* package was used to replicate the analysis.

Figure 3.4 A,B clearly shows a large distribution of the experimental *m/z* values (in red) scattered over 7.5 mDa and centered quite far from the exact theoretical *m/z* (2.1% and 3.3% of the pixels are located within 1 ppm from the exact mass of $[C_{40}H_{80}NO_8P+Na]^+$ and $[C_{42}H_{82}NO_8P+Na]^+$, respectively). After recalibration (in green), the distribution of the *m/z* is thinner (spread over ±0.5 mDa) and more accurate (97.35 and 98.74% of the pixels are located within 1 ppm from the exact mass of $[C_{40}H_{80}NO_8P+Na]^+$ and $[C_{42}H_{82}NO_8P+Na]^+$, respectively). The 2D intensity distribution (MSI image) of these two compounds extracted at 1 and 5 ppm around their theoretical *m/z* values are represented before calibration (Figure 3.4 C,G for 1 ppm and 3D,H for 5 ppm) and after recalibration (Figure 3.4 E,I for 1 ppm and 3F,J for 5 ppm).

Before recalibration, a mass tolerance window of 5 ppm was necessary to reconstruct the ion distribution as no image was obtained with 1 ppm. However, a large selection window increases the risk of getting additional ions in the selected window, resulting in the creation of composite images. After recalibration, most of the signals of the investigated ions are included in the mass selection window of ±1 ppm (green distributions). It results in these images being much more contrasted and more detailed, enhancing molecular description and interpretation. This highlights the impact of the mass shift on the reconstruction of *m/z* images. The comparison of MSI data before and after recalibration supports that our recalibration procedure avoids the loss in mass accuracy in the average MS signal due to mass shifts.

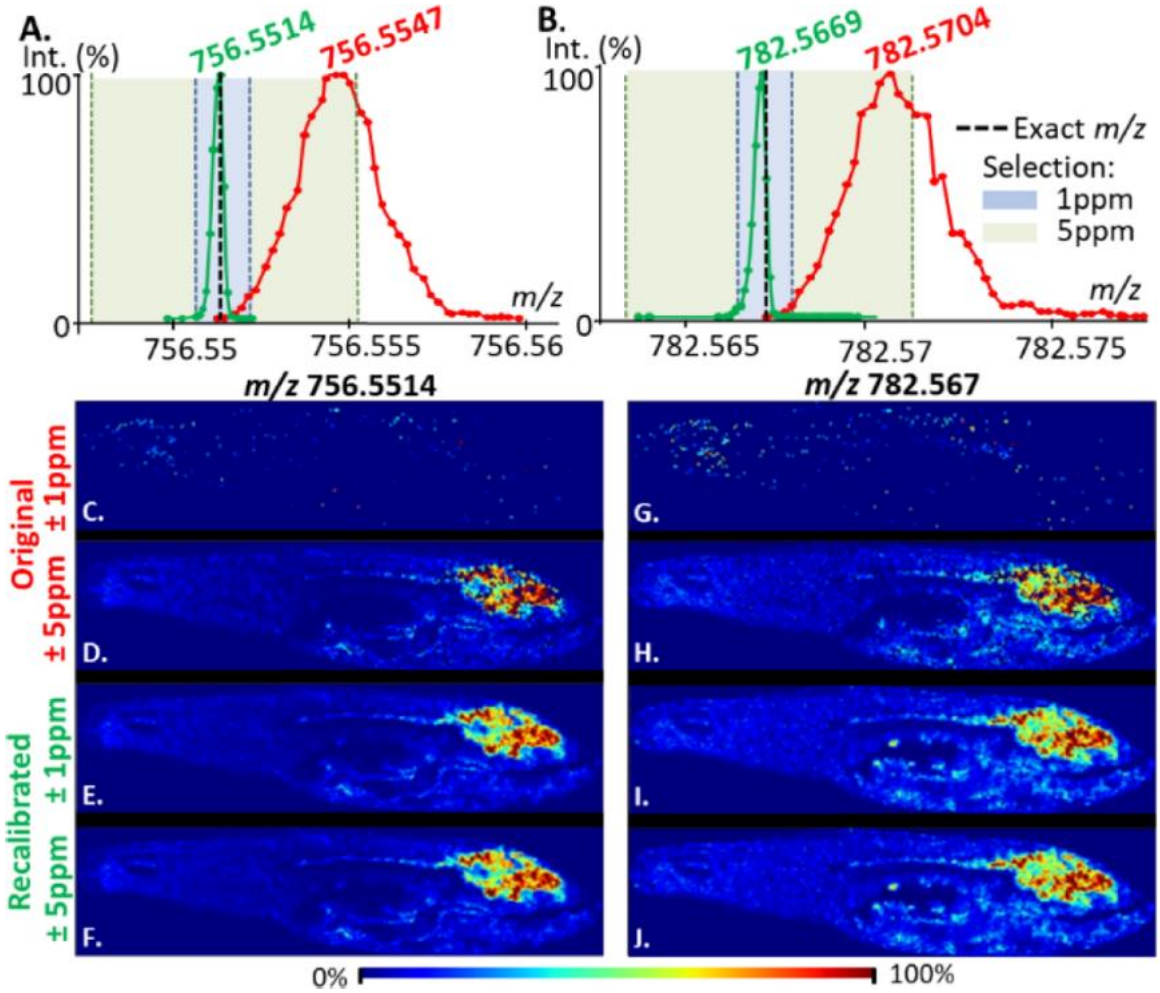


Figure 3.4. *From a zebrafish MSI data. Comparison of the average spectra before and after recalibration for two different well- characterized ions* $[C_{40}H_{80}NO_8P+Na]^+$ *and* $[C_{42}H_{82}NO_8P+Na]^+$*, with a theoretical m/z of 756.5514 and 782.5670 m/z, respectively (A,B), as well as the consequence on the reconstructed images at ± 1 ppm (C-G and E-I) and at ± 5 ppm (D-H and F-J). The images are reconstructed from the theoretical mass of the two ions.*

3.2.4.2 Impact of Recalibration on the Number of Molecular Annotations

High mass accuracy is a crucial parameter for obtaining accurate annotations of molecular ions in MSI in the absence of MS/MS (fragmentation information). As shown above, recalibration increases the accuracy of the MS measurement (mass error from 4.3 ppm to 0.12 ppm). Errors below 1 ppm are totally in line with the SolariX FT-ICR mass

analyzer accuracy expectations for a single spectrum (assuming adequate MS calibration). To evaluate the performance of the recalibration, a metric based on the errors of the hits could not be considered, as we do not know which of them are false or true. A new evaluation method is then consequently proposed. To quantify the impact of the recalibration on the annotation quality, both unprocessed (original) and recalibrated MSI data have been submitted to METASPACE for automatic annotation. Any METASPACE annotation is provided with its Metabolite-Signal Match (MSM) score that quantifies the likelihood of the measured signal to match the signal predicted for the molecule from a target database. The MSM score is computed by integrating (i) the measure of spatial chaos of the ion image at the selected *m/z*, (ii) the similarity between the experimental and theoretical isotopic patterns, and (iii) the spatial correlation between the reconstructed images of the isotopes. The estimation of false positives is made by employing a target-decoy approach where the decoy database contains implausible ions. From METASPACE, the target database can be selected among different popular options such as ChEBI (Chemical Entities of Biological Interest), HMDB (Human Metabolome Database), and LIPID MAPS. The false discovery rate[19] (FDR) is estimated as the proportion of signals that matches the decoy database for a given score against the signal that matches the target database for the same score. Mass shifts in MSI data should decrease the number of annotations for a given FDR by decreasing the true positive matches in the target database by increasing the possibility of matching decoy signals and by decreasing the structure of the spatial localization of an ion (Figure 3.4). The number of METASPACE annotations for a given FDR appears then as an adequate criterion to evaluate the performance of a recalibration strategy. Original and recalibrated MSI are annotated by METASPACE, using LIPID MAPS database, considering the following adducts: $[M+H]^+$, $[M+Na]^+$, and $[M+K]^+$. Mass tolerances used for the identifications are 0.5, 1.0, 1.5, 2.0, 2.5, and then from 3.0 to 10 ppm with a step of 1 ppm. The performance of the recalibration is assessed by comparing the number of annotations of the original and recalibrated MSI (Figure 3.5). Classically, annotations with an FDR of 10% or lower are kept for analysis, as lower FDR

corresponds to better annotation quality[19]. The most important increase would be at 1 ppm with an FDR of 10% where more than 200 additional compounds were identified. This increase of annotation strongly supports the effectiveness of our recalibration strategy for reducing the mass-shift effect and for increasing the accuracy of MSI data.

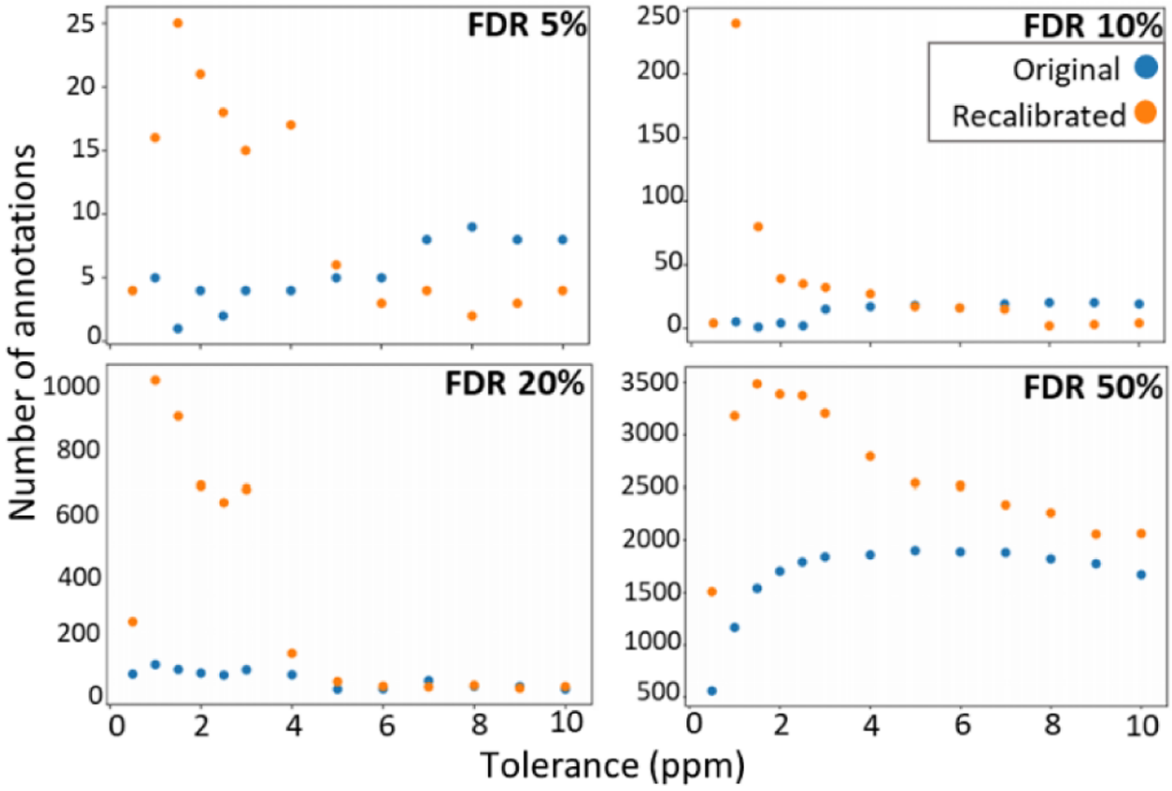


Figure 3.5. *Number of METASPACE annotations of the original (blue dots) and recalibrated (orange dots) zebrafish MSI for different mass tolerances (ppm). The number of annotations is shown for FDR ≤5, ≤10, ≤20, and ≤50%.*

3.2.4.3 FDR and Tolerance Selection

The evolution of annotation numbers with the tolerance for any FDR is due to two concomitant factors. First, the chances of matching decoy signals increase as the tolerance value increases, reducing the number of annotations. The second factor concerns the signal in the MSI data. When the tolerance value increases, the chance of matching isobaric or quasi-isobaric species instead of the expected signal increases as

well. If the isotopic pattern of these species is close to the suspected annotations, then the number of annotations will be falsely increased.

As a result, the variation in the number of annotations as a function of the mass tolerance depends on the predominance of these two factors, which is unpredictable for an unknown signal (Figure 3.5). However, the tolerance value in ppm should be low enough to optimize the number of annotations as it decreases the number of false positives. Therefore, the most interesting tolerance is the minimum tolerance value giving the highest number of annotations. FDRs at 5 or 10% are commonly used for the interpretation of sample composition[34]. FDRs below 10% are indicative of the quality of the sample signals, keeping only the most relevant identifications. However, higher FDR, such as 20 and 50%, may be useful when considering annotations with low signal intensities. Indeed, as annotation scoring also depends on spatial structures and isotope distribution, low-intensity signals have an impact on MSM scores and appear at higher FDR values. Since the recalibration shows a significant increase in annotations at higher FDRs (Figure 3.5), the majority of the signal is probably weak in the initial data.

3.2.4.4 Evaluation on a Large Set of Public MSI Data Sets

The proposed algorithm was evaluated with 31 different MSI data sets to assess its robustness under various experimental conditions. The recalibration of these data sets was performed using the settings detailed in the Material and Methods sections and Supporting Information, S2. The generation of calibrating ions for a given MSI is based on the most redundant ions from METASPACE annotated data sets with the same metadata (see Supporting Information, S2). The recalibration parameters were determined using the same approach as explained in the section above “Recalibration Impact on Data Analysis” leading to the same values. Finally, the same parameters were used for T1 and T2, that is, ± 0.01 and ±0.002 Da, respectively. Original and recalibrated MSI were then annotated by METASPACE. For MSI data in negative mode, only $[M-H]^-$ and

$[M+Cl]^-$ ions were considered, whereas for MSI data in positive mode, $[M+H]^+$, $[M+Na]^+$, and $[M+K]^+$ ions were taken into account. The mass tolerances set for the identification were 0.5 ppm and every unit from 1.0 to 8.0 ppm. The performance of the recalibration was assessed by comparing the number of annotations of the original and recalibrated MSI as discussed above. All possible databases available on the METASPACE platform were involved as target databases to generate the m/z_{calib} list. Among them, the database that provided the best number of annotations on the unprocessed MSI data at a mass tolerance of 3 ppm and FDR of 10% was chosen (default settings in METASPACE). Therefore, it should be kept in mind that annotations may come from different databases depending on the data set. Only the number of annotations at 5% FDR and a tolerance of 1 ppm was considered to normalize the evaluation of the algorithm's performance on this data set. For some data, the recalibration method seemed to perform better for other values (see Supporting Information, S3), but these were not taken into account because they were associated with a higher chance of false identifications (higher FDR or higher tolerance values). However, it could be useful to consider annotations at higher FDR values or higher tolerance for data mining as a perspective for this work. The benefit in terms of the number of annotations (i.e., the number of annotations in the recalibrated data minus the ones in the original data and thus the proportion of annotations gained) for each data set is reported in Figure 3.6.

Our algorithm was able to increase the annotation numbers for 75% of the data sets, without any prior knowledge of the samples. For 12 of the recalibrated data sets, more than 20 additional annotations were obtained. This is particularly useful since we considered high-quality annotations with FDR 5% that often correspond to the most relevant molecular signals. Recalibration could not be performed on 3 data sets (11, 12, and 13 in Figure 3.6 and Supporting Information, S3), as the number of hits is under the threshold criterion of hits for processing recalibration (i.e., a minimum of 10 hits; see the Methods section). After excluding these 3 MSI data sets, the recalibration resulted in

slightly fewer annotations for only 5 MSI data sets, which can be explained either by the poor quality of the data sets or by artefacts due to the recalibration.

Different hypotheses can be invoked to explain the cause of poor recalibration. In a general context, the quality of recalibration in each pixel is related to the number of true hits taken into account for the estimation of the recalibration function. The MS signal quantity (i.e., the amount of MS peaks over the noise) is the first key condition ensuring a robust recalibration of MSI data. Indeed, the recalibration of Orbitrap's MSI data, which globally contains more signals (more MS peaks) than other MSI data, gives the most impressive results (gaining from 20 up to 400 additional annotations) since more true hits are obtained. However, we cannot conclude that Orbitrap analyzers produce more MS signals than other mass analyzers in general, but this observation underlines the fact that recalibration performs better on data with a large number of MS signals. As already mentioned in the Method section, the poor performance of the selection of true hits can also be attributed to the mass error—mass trend (e.g., amplitude of the slope of the linear regression and the non-linearity evolution of the mass error in regard to the *m/z* values) (Figure 3.3b). However, taking these effects into account is challenging because the trend of the error may change from one pixel to another due to the heterogeneity of the signals in the MSI. Moreover, enough true hits must be achieved to properly determine the error trend over the mass range. One solution to overcome these effects is to increase the mass tolerance for the hits selection T2 and T1 for the most extreme mass-shift cases.

However, it should be kept in mind that this could increase the number of false hits. Different mass tolerances were tested but did not give better results (data not shown). The poor recalibration is consequently mainly attributed to a lack of true matches between m/z_{exp} and m/z_{calib} due to a weak signal. The values set for T1 and T2 seem to work well for all MSI data analyzed in this work which can be explained by the fact that all data were acquired with mass analyzers providing high mass accuracy measurement. This is not surprising since METASPACE annotates MSI data at ±3 ppm by

default, promoting the submission of MSI acquired with high-accuracy mass analyzers. Therefore, T1 and T2 could be used as constants for recalibrating MSI from METASPACE.

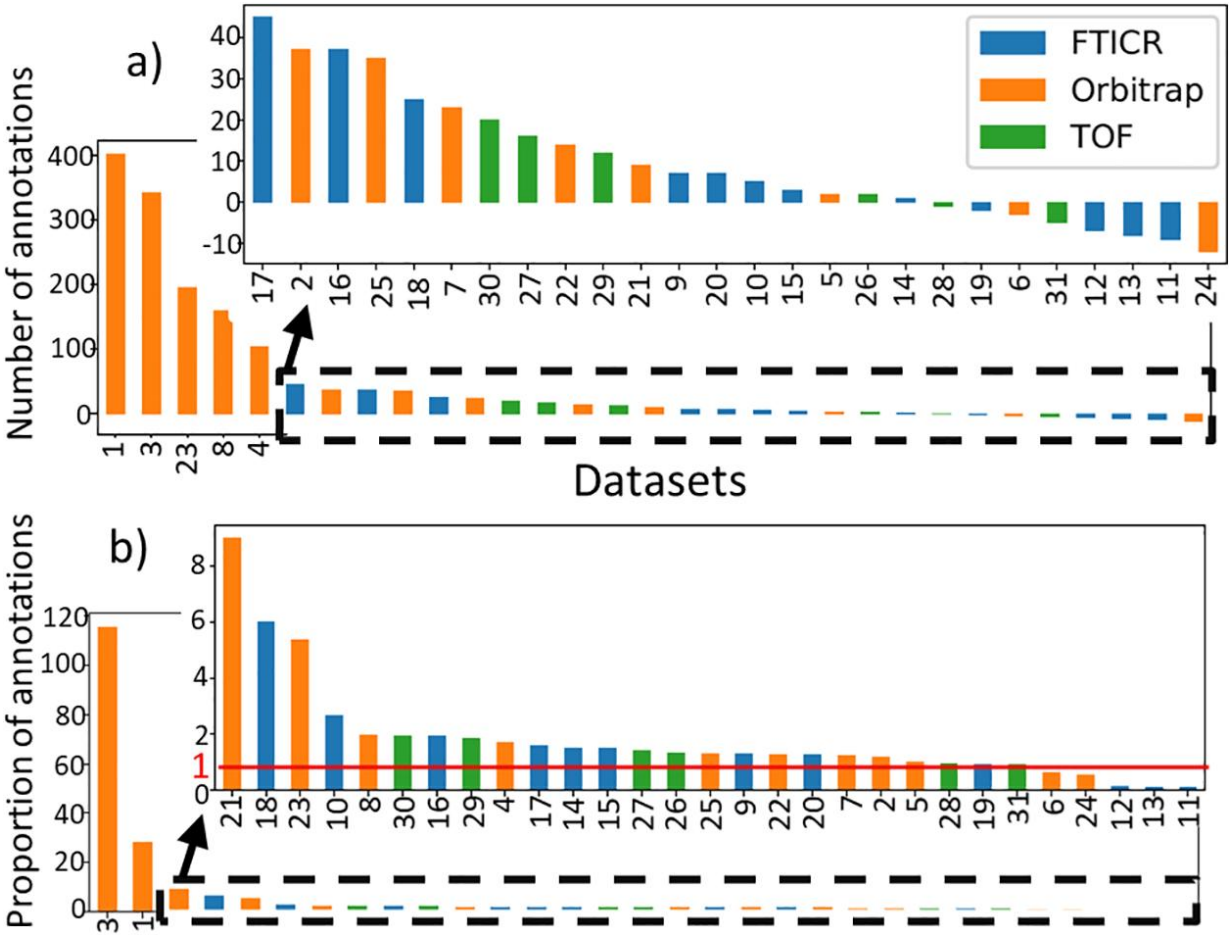


Figure 3.6. *Bar plot representing (a) gain in the number of METASPACE annotations (FDR 5%, 1 ppm tolerance) and (b) proportion of annotations gained for each of 31 public MSI data sets after recalibrating with our method.*

However, globally, it is essential to investigate certain recalibration conditions before applying the algorithm to other MSI data. These include (i) the assumption of the linear trend of the mass error as a function of *m/z* (Figure 3.3b), (ii) the amplitude of the slope of the error trend (Figure 3.3b), (iii) the maximum deviation of the mass shift T1, and (iv) the mass selection tolerance for hit selection T2. These may depend on the initial acquisition and sample heterogeneity, but also on the analyzer accuracy. Suggested quality plots (Figure 3.5 and Figures in Supporting Information, S3) could be considered

to validate the recalibration process on new MSI datasets beyond those evaluated here and to help select the correct algorithm parameters, in particular, T1 and T2 (see the Methods and "Recalibration Impact on Data Analysis" sections). Although the limitation of the number of hits seems to be the most crucial parameter to achieve good recalibration performance, this limitation will decrease in the future as the number of submitted data sets will increase. Indeed, enriching the database will cover more and more sample types and origins, thus increasing the quality of the m/z_{calib} lists generated. The sample signals will also likely be improved in the coming months and years, by increasing the ionization efficiency through the tuning of innovative ionization technologies such as laser-induced post-ionization (MALDI-2)[35] or by improving the transmission between the ion source and the MS analyzer.

3.2.5 Conclusions

Pixel-to-pixel mass shift decreases the interpretability of MSI by affecting the image reconstruction of identified ions and by decreasing the quality and the number of annotations in HRMS. Those have thus a crucial impact on biological interpretation, reducing the capacity to locate and annotate potential biomarkers on biological samples. Here, we report a new algorithm that recalibrates MSI data by adapting, at each pixel, the selection of endogenous signals for internal calibration. This method allows an optimal calibration of the different spatial regions of MSI data, decreasing the artefacts of mass shifts during data analysis.

An evaluation of the method was proposed by using the number of annotations obtained on the METASPACE platform for the original and recalibrated MSI data as an indicator of recalibration performances. Through this evaluation, we tested the performance and robustness of our recalibration algorithm on 31 different MSI data from various origins. An increase in the number of annotations, indicating better calibration, was observed after recalibration for most of the data investigated (75%). These results suggest that similar results could be obtained from different MSI acquired with similar MS analyzers. One aspect not demonstrated here is the recovery of spectral resolution in the image-averaged mass spectrum. Isobaric species, distinct ions with nearly identical *m/z* values, that are resolved on individual pixel spectra, can appear merged in the average, because pixel-to-pixel mass shifts broaden each peak during averaging. Recalibration should restore the separation of such isobaric pairs in the mean spectrum, an effect that will be characterized in future work.

However, this approach is not immune to certain limitations that could have an impact on recalibration performances. As discussed in the article, these limitations depend either on the signals of the sample or on the algorithm parameters. In the first case, it can be expected that in the near future the quality of the detected signal will increase

thanks to the advances and progress made in ionization efficiency, MS instrumentation, and in METASPACE annotations coverage. In the second case, inspection of the algorithm parameter values should be considered for images with low recalibration performances. To this end, the proposed quality plots provide a useful visualization tool to investigate the effects of the optimal parameters on the recalibrated MSI data.

Ultrahigh resolution instruments are not the only type of mass spectrometers that benefit from this method, as lower- resolution mass spectrometers may also suffer from mass shift. In this case, an increase of T2 will be expected as it is directly impacted by the accuracy of the mass analyzer. The reported recalibration strategy can also be applied to each type of MS data, provided that each acquisition scan contains enough peaks to properly estimate the true mass error, and that modeling hit errors with a linear model is sufficient. This is also true for other MS-based approaches involving hyphenated separation methods, as the MS signal is expected to vary a lot during the acquisition (high-performance liquid chromatography-mass spectrometry, capillary electrophoresis-mass spectrometry...). As for MSI acquisition, the fluctuation of ions during the acquisition requires internal calibration to effectively reduce the mass-shift effect. In addition, more and more instruments are integrating ion-mobility feature into mass spectrometry. The MS recalibration strategy proposed in this work could be extended in a broader perspective to calibrate ion mobility values (collision cross-section or mobility) through the use of mobility data from open sources.

3.2.6 Associated Content and Acknowledgments

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.0c05071.

Description of the 31 METASPACE MSI data according to the initial uploader; full detailed method; and plots for the 31 METASPACE MSI for the number of METASPACE annotations of the unprocessed (blue dots) and recalibrated (orange dots), for different mass tolerances (ppm) with the number of annotations shown for FDR ≤5, ≤10, ≤20, and ≤50% (PDF).

We thank the METASPACE submitters who shared their data for this work: James McKenzie (ICL), Zoltan Takats (ICL), Dhaka Bhandari (JLU Giessen), Mario Kompauer (JLU Giessen), Shane Ellis (M4I), Dusan Velickovic (PNNL), Jessica Lukowski (PNNL), Lavigne Regis (U Rennes 1), Annapurna Pamreddy (UTHSA), Guanshi Zhang (UTHSA), Bindesh Shrestha (Waters), and Emrys Jones (Waters). The authors acknowledge financial support from the European Union's Horizon 2020 research and innovation program under grant agreement no. 731077 (EU FT-ICR MS project, INFRAIA-02-2017) and from the European Union and Wallonia program FEDER BIOMED HUB Technology Support (no. 2.2.1/996) for the funding of the SolariX XR 9.4T. The authors also thank the European Union's Horizon 2020 program (EURLipids Interreg Euregio Meuse-Rhine project supported by the European Regional Development Fund (FEDER)), Horizon 2020 ERC project METACELL (no. 773089), and the ICT project CloudButton (no. 825184)

REFERENCES:

(1) Vaysse, P.-M.; Heeren, R. M. A.; Porta, T.; Balluff, B. Mass Spectrometry Imaging for Clinical Research – Latest Developments, Applications, and Current Limitations. *Analyst* **2017**, *142* (15), 2690–2712. https://doi.org/10.1039/C7AN00565B.

(2) Caprioli, R. M.; Farmer, T. B.; Gile, J. Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. *Anal. Chem.* **1997**, *69* (23), 4751–4760. https://doi.org/10.1021/ac970888i.

(3) Gode, D.; Volmer, D. A. Lipid Imaging by Mass Spectrometry – a Review. *Analyst* **2013**, *138* (5), 1289. https://doi.org/10.1039/c2an36337b.

(4) Khatib-Shahidi, S.; Andersson, M.; Herman, J. L.; Gillespie, T. A.; Caprioli, R. M. Direct Molecular Analysis of Whole-Body Animal Tissue Sections by Imaging MALDI Mass Spectrometry. *Anal. Chem.* **2006**, *78* (18), 6448–6456. https://doi.org/10.1021/ac060788p.

(5) Stoeckli, M.; Staab, D.; Schweitzer, A. Compound and Metabolite Distribution Measured by MALDI Mass Spectrometric Imaging in Whole-Body Tissue Sections. *International Journal of Mass Spectrometry* **2007**, *260* (2–3), 195–202. https://doi.org/10.1016/j.ijms.2006.10.007.

(6) Yang, J. Y.; Phelan, V. V.; Simkovsky, R.; Watrous, J. D.; Trial, R. M.; Fleming, T. C.; Wenter, R.; Moore, B. S.; Golden, S. S.; Pogliano, K.; Dorrestein, P. C. Primer on Agar-Based Microbial Imaging Mass Spectrometry. *J Bacteriol* **2012**, *194* (22), 6023–6028. https://doi.org/10.1128/JB.00823-12.

(7) Zaima, N.; Goto-Inoue, N.; Hayasaka, T.; Setou, M. Application of Imaging Mass Spectrometry for the Analysis of *Oryza Sativa* Rice. *Rapid Comm Mass Spectrometry* **2010**, *24* (18), 2723–2729. https://doi.org/10.1002/rcm.4693.

(8) Neumann, E. K.; Ellis, J. F.; Triplett, A. E.; Rubakhin, S. S.; Sweedler, J. V. Lipid Analysis of 30 000 Individual Rodent Cerebellar Cells Using High-Resolution Mass Spectrometry. *Anal. Chem.* **2019**, *91* (12), 7871–7878. https://doi.org/10.1021/acs.analchem.9b01689.

(9) Boggio, K. J.; Obasuyi, E.; Sugino, K.; Nelson, S. B.; Agar, N. Y.; Agar, J. N. Recent Advances in Single-Cell MALDI Mass Spectrometry Imaging and Potential Clinical Impact. *Expert Review of Proteomics* **2011**, *8* (5), 591–604. https://doi.org/10.1586/epr.11.53.

(10) Cooks, R. G.; Ouyang, Z.; Takats, Z.; Wiseman, J. M. Ambient Mass Spectrometry. *Science* **2006**, *311* (5767), 1566–1570. https://doi.org/10.1126/science.1119426.

(11) Brunelle, A.; Touboul, D.; Laprévote, O. Biological Tissue Imaging with Time-of-flight Secondary Ion Mass Spectrometry and Cluster Ion Sources. *J. Mass Spectrom.* **2005**, *40* (8), 985–999. https://doi.org/10.1002/jms.902.

(12) Alexandrov, T. MALDI Imaging Mass Spectrometry: Statistical Data Analysis and Current Computational Challenges. *BMC Bioinformatics* **2012**, *13* (S16), S11. https://doi.org/10.1186/1471-2105-13-S16-S11.

(13) Buchberger, A. R.; DeLaney, K.; Johnson, J.; Li, L. Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights. *Analytical Chemistry* **2018**, *90* (1), 240–265. https://doi.org/10.1021/acs.analchem.7b04733.

(14) Schwudke, D.; Schuhmann, K.; Herzog, R.; Bornstein, S. R.; Shevchenko, A. Shotgun Lipidomics on High Resolution Mass Spectrometers. *Cold Spring Harbor Perspectives in Biology* **2011**, *3* (9), a004614–a004614. https://doi.org/10.1101/cshperspect.a004614.

(15) Nikolaev, E. N.; Kostyukevich, Y. I.; Vladimirov, G. N. Fourier Transform Ion Cyclotron Resonance (FT ICR) Mass Spectrometry: Theory and Simulations. *Mass Spectrometry Reviews* **2016**, *35* (2), 219–258. https://doi.org/10.1002/mas.21422.
(16) Smith, D. F.; Kharchenko, A.; Konijnenburg, M.; Klinkert, I.; Paša-Tolić, L.; Heeren, R. M. A. Advanced Mass Calibration and Visualization for FT-ICR Mass Spectrometry Imaging. *Journal of the American Society for Mass Spectrometry* **2012**, *23* (11), 1865–1872. https://doi.org/10.1007/s13361-012-0464-1.
(17) Kozhinov, A. N.; Zhurov, K. O.; Tsybin, Y. O. Iterative Method for Mass Spectra Recalibration via Empirical Estimation of the Mass Calibration Function for Fourier Transform Mass Spectrometry-Based Petroleomics. *Anal. Chem.* **2013**, *85* (13), 6437–6445. https://doi.org/10.1021/ac400972y.
(18) Burton, R. D.; Matuszak, K. P.; Watson, C. H.; Eyler, J. R. Exact Mass Measurements Using a 7 Tesla Fourier Transform Ion Cyclotron Resonance Mass Spectrometer in a Good Laboratory Practices-Regulated Environment. *J. Am. Soc. Mass Spectrom.* **1999**, *10* (12), 1291–1297. https://doi.org/10.1016/S1044-0305(99)00106-3.
(19) Palmer, A.; Phapale, P.; Chernyavsky, I.; Lavigne, R.; Fay, D.; Tarasov, A.; Kovalev, V.; Fuchser, J.; Nikolenko, S.; Pineau, C.; Becker, M.; Alexandrov, T. FDR-Controlled Metabolite Annotation for High-Resolution Imaging Mass Spectrometry. *Nature Methods* **2017**, *14* (1), 57–60. https://doi.org/10.1038/nmeth.4072.
(20) Alexandrov, T. Spatial Metabolomics and Imaging Mass Spectrometry in the Age of Artificial Intelligence. *Annual Review of Biomedical Data Science* **2020**, *3* (1), 61–87. https://doi.org/10.1146/annurev-biodatasci-011420-031537.
(21) Gross, J. H. *Mass Spectrometry: A Textbook*; Springer Berlin Heidelberg: Berlin, Heidelberg, 2011. https://doi.org/10.1007/978-3-642-10711-5.
(22) Tiquet, M.; La Rocca, R.; Van Kruining, D.; Martinez-Martinez, P.; Eppe, G.; De Pauw, E.; Quinton, L.; Far, J. Mass Spectrometry Imaging Using Dynamically Harmonized FT-ICR at Million Resolving Power: Rationalizing and Optimizing Sample Preparation and Instrumental Parameters. Chemistry September 29, 2020. https://doi.org/10.26434/chemrxiv.13013900.v1.
(23) Kooijman, P. C.; Nagornov, K. O.; Kozhinov, A. N.; Kilgour, D. P. A.; Tsybin, Y. O.; Heeren, R. M. A.; Ellis, S. R. Increased Throughput and Ultra-High Mass Resolution in DESI FT-ICR MS Imaging through New-Generation External Data Acquisition System and Advanced Data Processing Approaches. *Sci Rep* **2019**, *9* (1), 8. https://doi.org/10.1038/s41598-018-36957-1.
(24) Boskamp, T.; Lachmund, D.; Casadonte, R.; Hauberg-Lotte, L.; Kobarg, J. H.; Kriegsmann, J.; Maass, P. Using the Chemical Noise Background in MALDI Mass Spectrometry Imaging for Mass Alignment and Calibration. *Anal. Chem.* **2020**, *92* (1), 1301–1308. https://doi.org/10.1021/acs.analchem.9b04473.
(25) Tracy, M. B.; Chen, H.; Weaver, D. M.; Malyarenko, D. I.; Sasinowski, M.; Cazares, L. H.; Drake, R. R.; Semmes, O. J.; Tracy, E. R.; Cooke, W. E. Precision Enhancement of MALDI-TOF MS Using High Resolution Peak Detection and Label-free Alignment. *Proteomics* **2008**, *8* (8), 1530–1538. https://doi.org/10.1002/pmic.200701146.
(26) Brochu, F.; Plante, P.-L.; Drouin, A.; Gagnon, D.; Richard, D.; Durocher, F.; Diorio, C.; Marchand, M.; Corbeil, J.; Laviolette, F. Mass Spectra Alignment Using Virtual Lock-Masses. *Sci Rep* **2019**, *9* (1), 8469. https://doi.org/10.1038/s41598-019-44923-8.

(27) Ràfols, P.; Castillo, E. D.; Yanes, O.; Brezmes, J.; Correig, X. Novel Automated Workflow for Spectral Alignment and Mass Calibration in MS Imaging Using a Sputtered Ag Nanolayer. *Analytica Chimica Acta* **2018**, *1022*, 61–69. https://doi.org/10.1016/j.aca.2018.03.031.
(28) Kulkarni, P.; Kaftan, F.; Kynast, P.; Svatoš, A.; Böcker, S. Correcting Mass Shifts: A Lock Mass-Free Recalibration Procedure for Mass Spectrometry Imaging Data. *Anal Bioanal Chem* **2015**, *407* (25), 7603–7613. https://doi.org/10.1007/s00216-015-8935-4.
(29) Schramm, T.; Hester, Z.; Klinkert, I.; Both, J.-P.; Heeren, R. M. A.; Brunelle, A.; Laprévote, O.; Desbenoit, N.; Robbe, M.-F.; Stoeckli, M.; Spengler, B.; Römpp, A. imzML — A Common Data Format for the Flexible Exchange and Processing of Mass Spectrometry Imaging Data. *Journal of Proteomics* **2012**, *75* (16), 5106–5110. https://doi.org/10.1016/j.jprot.2012.07.026.
(30) Ledl, T. Kernel Density Estimation: Theory and Application in Discriminant Analysis. *AJS* **2016**, *33* (3), 267–279. https://doi.org/10.17713/ajs.v33i3.441.
(31) Fischler, M. A.; Bolles, R. C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM* **1981**, *24* (6), 381–395. https://doi.org/10.1145/358669.358692.
(32) Sud, M.; Fahy, E.; Cotter, D.; Brown, A.; Dennis, E. A.; Glass, C. K.; Merrill, A. H.; Murphy, R. C.; Raetz, C. R. H.; Russell, D. W.; Subramaniam, S. LMSD: LIPID MAPS Structure Database. *Nucleic Acids Research* **2007**, *35* (Database), D527–D532. https://doi.org/10.1093/nar/gkl838.
(33) Pirro, V.; Guffey, S. C.; Sepúlveda, M. S.; Mahapatra, C. T.; Ferreira, C. R.; Jarmusch, A. K.; Cooks, R. G. Lipid Dynamics in Zebrafish Embryonic Development Observed by DESI-MS Imaging and Nanoelectrospray-MS. *Mol. BioSyst.* **2016**, *12* (7), 2069–2079. https://doi.org/10.1039/C6MB00168H.
(34) Alexandrov, T.; Ovchinnikova, K.; Palmer, A.; Kovalev, V.; Tarasov, A.; Stuart, L.; Nigmetzianov, R.; Fay, D.; Key METASPACE contributors; Gaudin, M.; Lopez, C. G.; Vetter, M.; Swales, J.; Bokhart, M.; Kompauer, M.; McKenzie, J.; Rappez, L.; Velickovic, D.; Lavigne, R.; Zhang, G.; Thinagaran, D.; Ruhland, E.; Sans, M.; Triana, S.; Sammour, D. A.; Aboulmagd, S.; Bagger, C.; Strittmatter, N.; Rigopoulos, A.; Gemperline, E.; Joensen, A. M.; Geier, B.; Quiason, C.; Weaver, E.; Prasad, M.; Balluff, B.; Nagornov, K.; Li, L.; Linscheid, M.; Hopf, C.; Heintz, D.; Liebeke, M.; Spengler, B.; Boughton, B.; Janfelt, C.; Sharma, K.; Pineau, C.; Anderton, C.; Ellis, S.; Becker, M.; Pánczél, J.; Violante, G. D.; Muddiman, D.; Goodwin, R.; Eberlin, L.; Takats, Z.; Shahidi-Latham, S. METASPACE: A Community-Populated Knowledge Base of Spatial Metabolomes in Health and Disease. Systems Biology February 3, 2019. https://doi.org/10.1101/539478.
(35) Soltwisch, J.; Kettling, H.; Vens-Cappell, S.; Wiegelmann, M.; Müthing, J.; Dreisewerd, K. Mass Spectrometry Imaging with Laser-Induced Postionization. *Science* **2015**, *348* (6231), 211–215. https://doi.org/10.1126/science.aaa1051.

CHAPTER FOUR

***

Mass, Mass defect, and Mass pattern as Complementary Information for Mass Spectrum Data Processing

Chapter 4: Contents

4.1 Kendrick Mass Defect as a filter for MS signal 129

4.1.1 Lack of spectral information in MSI annotation engine 129

4.1.2 Kendrick mass defect graphical representation 130

4.1.3 Rapid Visualization of Chemically Related Compounds Using Kendrick Mass Defect as a Filter in Mass Spectrometry Imaging 133

4.1.4 Limitations of MSKF 137

4.2 Expand Kendrick by adding a graph of mass differences 139

4.2.1 Material 145

4.2.2 MS spectrum as a graph of mass differences 146

4.2.3 Application to bacterial confrontation MSI 150

4.2.4 Discussion and Conclusion 158

4.2.5 Introducing machine learning for processing graphs of mass differences 160

4.1 Kendrick Mass Defect as a filter for MS signal

4.1.1 Lack of spectral information in MSI annotation engine

As discussed in the previous chapters, most signals detected in mass spectrometry imaging still need to be identified. Only a limited number of signals are usually recognized and annotated in METASPACE, with an estimation of 1% of MSI data[1]. This limitation stems from various sources, such as the scope of metabolite databases, insufficient knowledge of background signals, or, again, the scoring method used for annotation assignment. In Chapter 3, we saw how the mass shift can impact the interpretation and signal identification by decreasing the Metabolite-Signal Match (MSM) score of the METASPACE search engine. The MSM score uses different algorithms to determine what can be considered a good hit or not for a specific MS signal, such as the structure of the ion image, the signal correlation of the detected isotope distribution, and its theoretical isotope patterns.

The METASPACE search engine provides a limited view of spectral information, considering only the monoisotopic mass and its isotopic pattern for a given database hit[2]. However, given a common mass spectrum, MS peaks are rarely independent from each other. Indeed, a discrete molecule/compound can be detected as multiple mass peaks corresponding to cationization, adducts, and charge states, for example. Identifying the different signals corresponding to the same compound increases the confidence that those peaks come from that molecule. Let us consider a peak P1 identified as a molecule A, ionized by the means of an H^+ adduct, or as a molecule B cationized with a sodium ion Na^+. If two peaks, P2 and P3, detected in the same spectrum are identified as molecule A but with Na^+ and K^+ adducts, respectively, and no other peaks related to molecule B are detected, then P1 should be attributed to molecule A with higher confidence. This reasoning is also valid when multiple peaks are identified as different molecules from the same molecular family. Let us now consider a spectrum with only five peaks (P1 to P5) identified as triglycerides ionized with the

same adduct. Only P5 is, however, ambiguous, identified as a triglyceride or a non-lipid metabolite. In this case, P5 should be considered a triglyceride with higher confidence than the non-lipid metabolite annotation, as the triglyceride annotation is more frequent in this spectrum than the non-lipid metabolite.

4.1.2 Kendrick mass defect graphical representation

Finding a way to exploit the structural particularity of the MS signals, such as the molecular family to which a molecule belongs, would help confirm or invalidate hypothetical ion identifications. In this way, the Kendrick plot is a 2D representation of a mass spectrum that helps infer molecular family information from MS peaks without requiring a database of theoretical molecular formulae.

Kendrick's plots are constructed from Kendrick's mass defects (KMD) and are 2D maps that show the KMD as a function of the measured m/z[3]. Kendrick's mass defects rest on the change of the classical IUPAC atomic mass unit (amu) reference to another mass unit called *kb*. KMD is calculated from the KM, following these two equations:

$$KM = m/z \ \frac{\text{nominal mass (kb)}}{exact\ mass\ (kb)} \qquad (1)$$

$$KMD = round(KM) - KM \qquad (2)$$

In Figure 4.1, we illustrate the capacity of Kendrick representation to filter different molecular families from a prepared and designed complex sample. The sample comprises multiple MALDI spots, each containing a mix of the following diverse molecular families: lipids, lipopeptides, CH_3O-PEO-H polymers, tetraalkylammonium (TAA), and peptides of BSA tryptic digest, deposited in the 1,5-diaminonaphthalene MALDI matrix. All those different molecular families generate a complex mass spectrum that can be transformed into the Kendrick space.

In this example, we generate a Kendrick plot of this spectrum by changing the IUPAC amu reference (based on carbon atom, *kb* = *C*) to *kb* = CH_2. Therefore, in this new space, the nominal mass (kb) will be 14.0000 and the exact mass (kb) will be 14.01565. Applying formulae (1) and (2) to a list of *m/z* values allows for the identification of *m/z* values spaced apart by 14.01565, resulting in the same Kendrick Mass Defect (KMD). These *m/z* values can then be horizontally aligned on the KM axis of the Kendrick plot, making them easier to identify. Consequently, all the compounds differing only by a number of CH_2 will share the same KMD, and will therefore be automatically aligned in the map KMD=f(KM). In other words, the plot allows a decrease in the spectrum complexity by highlighting clusters of molecular families as horizontal sequences of points in the plot (Figure 4.1. A). In this example, the conversion is particularly useful for identifying which *m/z* values correspond to the tetraalkylammonium, lipid, and lipopeptide families (respectively highlighted in blue, green, and orange), as they now appear as horizontal lines in the plot (peaks(*m/z*) of each molecular family have the same KMD).

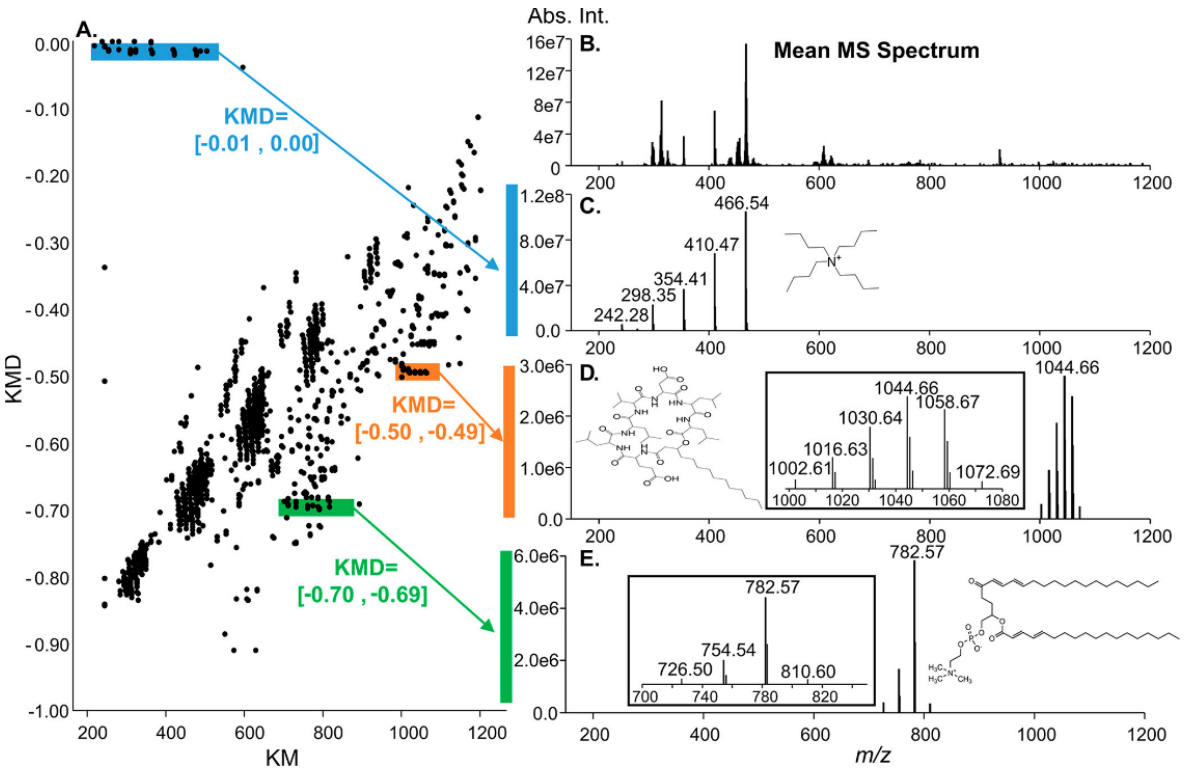


Figure 4.1. *Illustration of the selection of molecular family from a Kendrick plot. (A) Kendrick mass defect plot of the spectra obtained by FT-ICR MS. (B) Full MS spectrum when no filtration is applied. By selecting a specific KMD range and eventually a specific mass range, it is possible to filter the MS Spectra and show only the chemically related molecules such as TAA (C), lipopeptides: surfactins (D) or lipids: phosphatidylcholines (E).*

From these convoluted mass spectra, it is now possible to highlight a particular molecular family (target) by filtering using Kendrick analysis, based on the mass of the target family's main repeating unit. To horizontally align and highlight other molecular families, the kb value can be easily adjusted to the mass of the repeating unit that characterizes the targeted molecular family. For example, *m/z* values corresponding to the polymer family CH_3O-PEO-H (target) can be aligned by changing the *kb* value to the mass of the ethylene glycol group (C_2H_4O), the main repeating unit of the polymer polyethylene oxide.

However, Kendrick alone is not enough to isolate a molecular family completely. It is important to note that a shared KMD does not by itself imply chemical relatedness: because the KMD encodes only the mass defect relative to the chosen base unit (kb), chemically unrelated families can fall on the same horizontal sequence at different *m/z*. Given the mass spectrometer resolution and the size of the selected KMD and KM ranges, a single KMD range may therefore contain several distinct molecular families, which is precisely why additional criteria are needed to confirm family membership. Moreover, the number of clusters (horizontal sequences of points) remains high, limiting exploration of spectral signals to only obvious, large horizontal sequences, leaving other clusters unexplored. Thus, Kendrick's plots primarily serve to visualize raw spectral information, helping manually select peaks of a particular molecular family by interpreting mass defects and patterns in raw MS signals.

4.1.3 Rapid Visualization of Chemically Related Compounds Using Kendrick Mass Defect as a Filter in Mass Spectrometry Imaging

Kendrick plot filtering is often used to analyze complex samples not typically related to mass spectrometry imaging; however, it appears particularly useful for imaging signals for two main reasons.

(i) As a filtering method to remove background signals (matrix adducts, chemical noise…) localized in different KMD ranges.

(ii) To efficiently annotate the chemical composition of a large set of spectra (typically found in MSI) by identifying molecular family clusters, highlighted through their alignment in the Kendrick plot.

For these reasons, a Python-based program to filter/select MS signals based on their plotted KM and KMD has been developed in the MS-Lab. The program called Mass Spectrometry Kendrick Mass Filter (MSKF) (graphical interface developed by Christopher Kune and reimplemented in this thesis in the *deltamsi* package) allows MS signal selection in a targeted or untargeted manner by selecting a particular KM/KMD region or by letting the program parse all the KMD regions. Since multiple molecular families can be found within the same KM/KMD regions, we introduce a new filter that clusters MS signals of the same KMD if they are spaced by a specified $exact\ mass\ (kb)$.

In Figure 4.1, the *m/z* values corresponding to each molecular family have been extracted with a great efficiency thanks to MSKF program. A Kendrick plot was first generated from a raw MSI spectrum, using a kb value (CH_2), resulting in a plot in which horizontal sequences of points correspond to ions of the same family. MSKF is then used to manually select a region (rectangular selection) around those horizontal lines, defining a KMD and KM range for each selection (Figure 4.1, left: blue, green, and orange rectangles). For each selection (rectangle), we cluster the *m/z* that are spaced by kb and generate a filtered spectrum resulting in a clean selection of ions from

tetraalkylammonium (Figure 4.1. C), surfactins (lipopeptides family, Figure 4.1. D) and phosphatidylcholines (lipid family, Figure 4.1. E).

When applied to an MS image, MSKF can, in addition to the filtered spectrum, visualize the combined ion images of all selected MS signals within the filtered regions. Indeed, MSKF was successfully applied to 2 different MSI experiments: a MALDI imaging of bacterial culture (Figure 4.2) and a MALDI imaging of a mouse brain section (Figure 4.3). In the bacterial dataset, the objective was to detect different families of lipopeptides secreted by two bacterial strains (*Bacillus velezensis GA1* and *Pseudomonas sessilinigenes* CMR12a), grown together in coculture. To identify such lipopeptides using MSKF, *kb* =CH_2 was chosen to take advantage of the property that lipopeptides contain a variable hydrophobic lipid chain that varies by the number of -CH_2-. The MS signal is then filtered in an untargeted manner, revealing different groups of peaks at different KMDs (Figure 4.2, left). Those peaks are then identified by their exact masses as iturins, surfactins, and sessilin and show different spatial distributions according to their lipopeptide families (Figure 4.2). Iturins and surfactins are known to be secreted by *Bacillus sp.*, as evidenced by their detection on the colony and around it (Figure 4.2 (B) and (C)), with clear diffusion of the compounds into the medium. In contrast, sessilins (Figure 4.2 (F)) are only detected on *Pseudomonas sp.* as expected, and with no apparent diffusion.

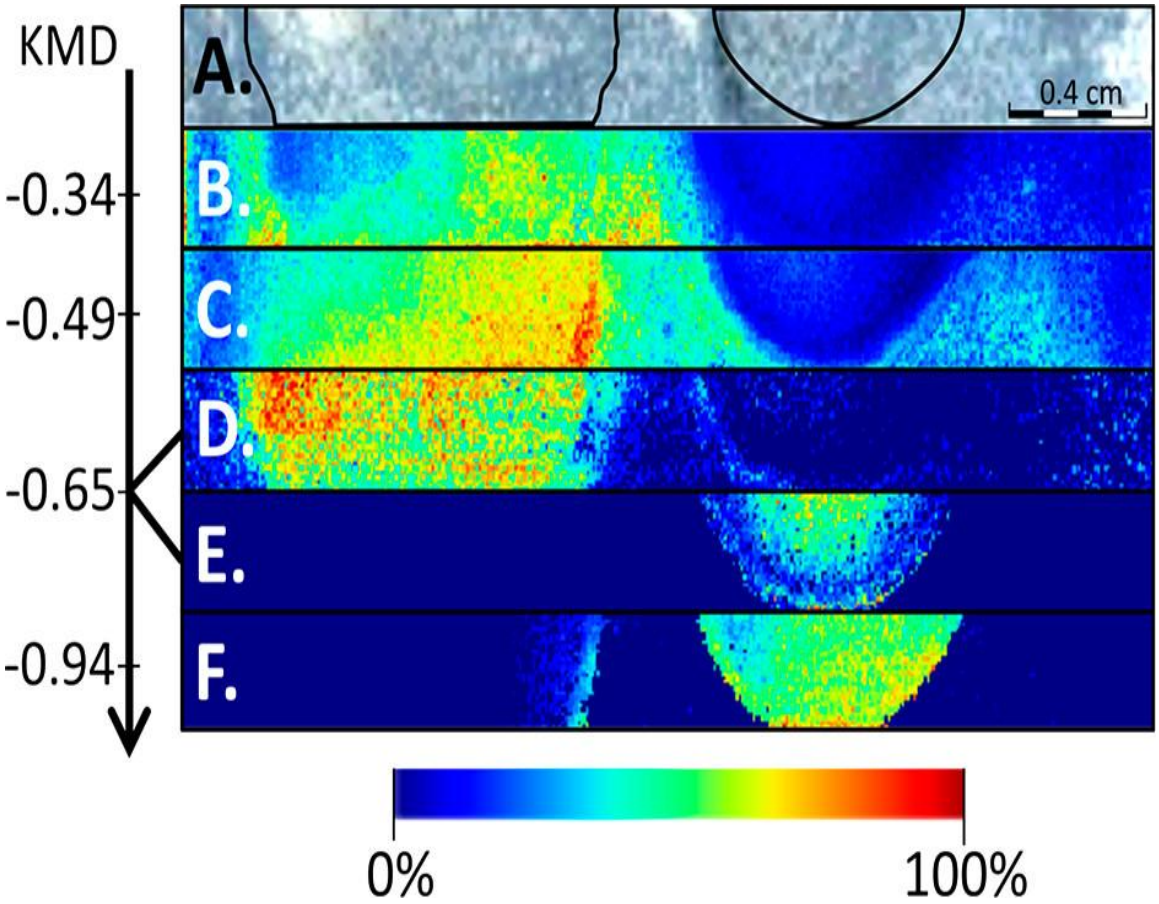


Figure 4.2. *KMD filtering was applied to MALDI imaging of bacterial culture. MALDI mass spectrometry images of dried microbial colonies grown on agar and sprayed with HCCA were obtained using an FT-ICR mass spectrometer at a spatial resolution of 100 µm[4]. (A) Optical image of the sample covered with matrix, Bacillus sp. is circled on the left side, and Pseudomonas sp. is circled on the right side. The identified families are presented in order of their KMD values. (B) Iturins A (sodium adducts) at the KMD range −0.33 to −0.35 (C) Surfactins (sodium adducts) at the KMD range −0.49 to −0.51. (D) and (E) identification of two different lipid classes at the KMD range −0.65 to −0.66, (F) Sessilins (sodium adducts) at the KMD range −0.93 to −0.95. The color scale represents the normalized ion intensity.*

In the second dataset, the program was used with the same *kb* to cluster MS signals corresponding to lipids in the mouse brain. Indeed, lipids are composed of various families that differ mostly in the polarity of their polar heads and the lengths of their fatty chains, which are extended by the addition of multiple CH_2 groups. Once again, detected peaks are confirmed by exact mass matching with the LipidMaps database. MSKF allows the detection of lipids of different classes and their localizations (Figure

4.3). In Figure 4.3, we can see that the spatial localization differs from each lipid class. Moreover, within the same lipid class and molecular family, the spatial distribution varies with the degree of GPC lipid unsaturation.

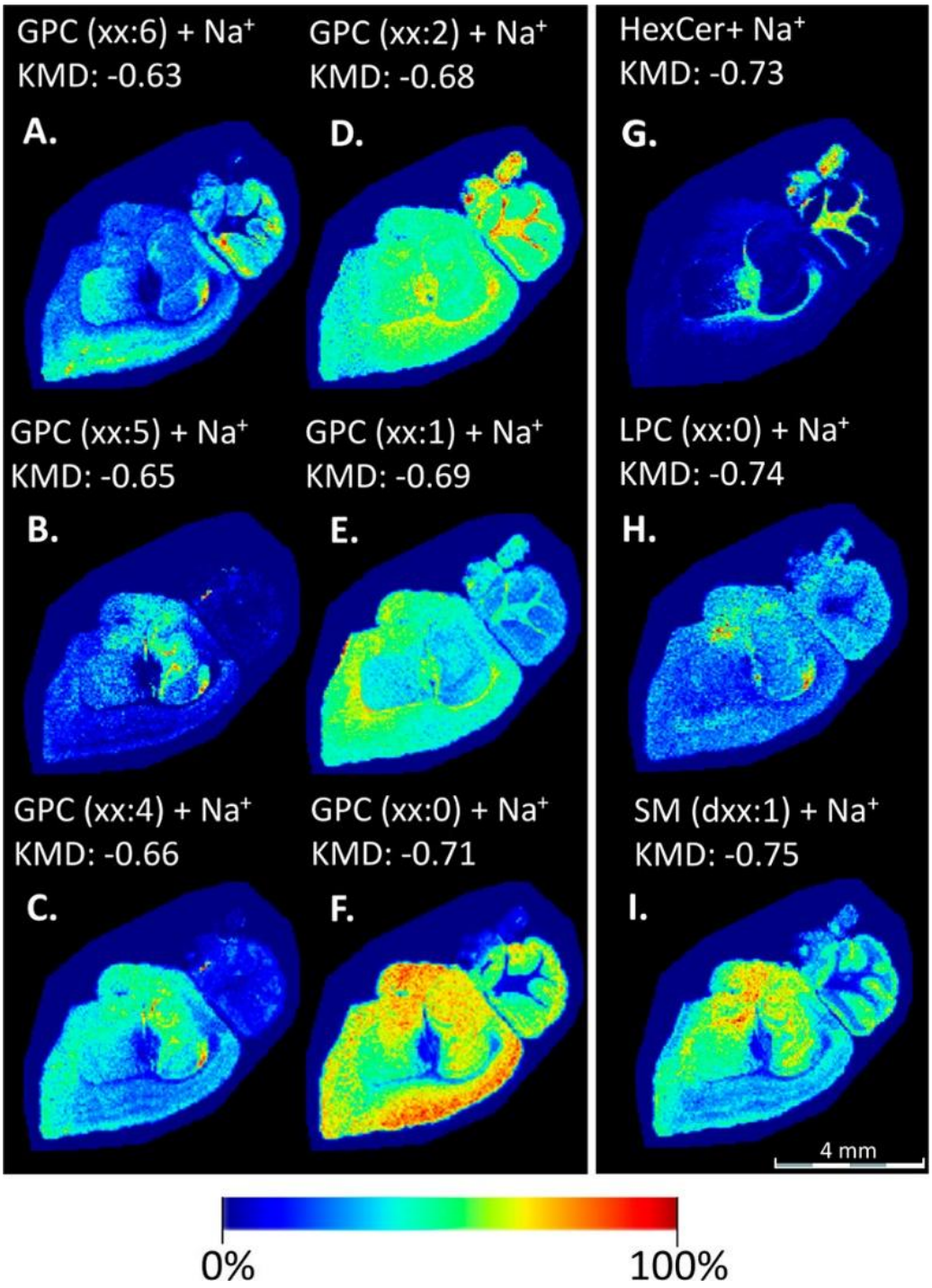


Figure 4.3. *KMD filtering is applied on MALDI imaging of a sagittal slice of a mouse brain. MALDI mass spectrometry images of mouse brain tissues sectioned using a Cryostat into 14 µm thick sagittal slices and sprayed with HCCA were obtained using an FT-ICR mass spectrometer at a spatial resolution of 100 µm[4]. Lipid classes (GPCs, SMs, LPCs) are attributed to each cluster, and according to their unsaturation (xx:n), "xx" represents the number of carbon, and "n" corresponds to the number of unsaturation. (A) to (F) Image of sodiated glycerophosphocholines, $[GPCs+Na]^+$ with 6, 5, 4, 2, 1, and 0 unsaturations, respectively. (G) Image of sodiated hexosylceramides, $[HexCers+Na]^+$. (H) Image of sodiated lysophoshocholines, $[LPCs+Na]^+$. (I) Image of sodiated sphingomyelins, $[SMs+Na]^+$. The color scale represents the normalized ion intensity.*

These works mainly refer to the following publications, where the detailed parameters of image acquisition and data analysis are presented: *"Rapid Visualization of Chemically Related Compounds Using Kendrick Mass Defect As a Filter in Mass Spectrometry Imaging"* written by Christopher Kune, Andréa McCann, **Raphaël La Rocca**, Anthony Arguelles Arias, Mathieu Tiquet, Daan Van Kruining, Pilar Martinez Martinez, Marc Ongena, Gauthier Eppe, Loïc Quinton, Johann Far, and Edwin De Pauw, and published in Analytical Chemistry. **2019**; 91, 13112−13118.

Using Kendrick's mass-defect-based methodology for MSI datasets is useful for processing and analyzing complex MSI data by focusing the analysis on essential signals without limiting us to signals known in advance. For example, lipopeptides can be found by database matching without using Kendrick filtering. However, we are limited to lipopeptides already present in the database, rather than those that could be discovered by visualizing their behavior on Kendrick's plot.

4.1.4 Limitations of MSKF

The MSKF method relies on a manual selection of a Kendrick region, followed by clustering peaks using the *kb* value to limit the introduction of unrelated signals in the cluster. Kendrick's representation offers advantages over classical database matching, such as annotation propagation and highlighting the *m/z* of interest from raw data without a database. Exploiting these advantages depends mainly on the preliminary

user's knowledge, which guides the selection of the *kb* and the KMD area to explore through manual selection.

However, this method introduces different limitations and biases for MSI analysis. In MSKF, only a single mass difference (*kb*) is used for peak clustering. This restriction results from the need to manually select clusters, as confirmed by visual inspection of a horizontal sequence of points in the plot (which is only possible for a single *kb* value). For example, lipids can be clustered by kb=CH_2, corresponding to the increase of the fatty chain lengths, and by kb= H_2, corresponding to the variation of the degree of unsaturation. Using a single mass-difference selection is an important limitation of the MSKF program for analyzing MALDI MSI signals related to metabolites, since multiple patterns (mass differences) can be observed within a molecular family, and different kb should be used depending on the nature of the molecules present in the sample. For example, different adducts, isotopes, and diverse metabolite modifications such as methylation, oxidation, and unsaturation could be identified from mass differences. A second limitation of MSKF is that the spatial distributions of the ions are not considered during the clustering. Therefore, two ions spaced by kb with different/opposite ion images can still be clustered, increasing the risk of grouping unrelated images. Lastly, a Kendrick region is selected/judged attractive based on visual inspection of horizontal trends. Due to this subjective inspection, it is impossible to quantify the confidence this selection provides in identifying and selecting an ion. As stated previously, specific mass patterns between ions could increase the confidence of an identification. This information can be integrated into the METASPACE scoring system to improve the confidence of ion identification.

4.2 Expand Kendrick by adding a graph of mass differences

Building on the limitations of Kendrick-based filtering described above, we introduce here an alternative representation of a mass spectrum designed to improve the automated selection of groups of peaks (and, in MSI, their corresponding ion images) that are likely to originate from chemically related molecules. Kendrick mass defect (KMD) plots are highly effective for visualizing homologous series: after choosing a repeating unit (e.g., CH_2), ions that differ by this unit tend to align and form horizontal sequences. However, translating this visual structure into a robust, automated clustering strategy is not straightforward.

In practice, the apparent proximity of points in (KM, KMD) space does not reliably define chemical relatedness: the separation between series depends on the chosen Kendrick base, multiple molecular families can overlap within the same KM/KMD region, and series can extend as elongated patterns with non-uniform density. As a consequence, applying generic distance-based clustering directly on the Kendrick coordinates (e.g., k-means or hierarchical clustering) is often sensitive to parameterization (scaling, bandwidths, cut thresholds) and may merge unrelated patterns or fragment a single family into many small clusters. These limitations motivate a representation that explicitly encodes chemical relations, not as geometric proximity in a 2D plot, but as connections defined by multiple expected mass differences.

A complementary way to represent high-resolution MS data (including MSI peak lists) is to treat detected features, typically measured *m/z* values, as nodes in a graph, and to connect two nodes when their Δm matches a chemically interpretable transformation. This mass-difference network concept was introduced and demonstrated by Rainer Breitling and coworkers on FT-ICR-MS metabolomics data from the protozoan parasite *Trypanosoma brucei*, where they profiled parasites grown in vivo (rats) and in vitro

(serum culture) and compared these to the corresponding environmental profiles (rat serum and culture medium). Using the resulting peak lists (hundreds of accurate masses), they reconstructed ab initio networks in two complementary ways: (i) an untargeted strategy that identifies frequently recurring Δm among all pairwise mass differences (interpreted as non-random, chemically grounded transformations), and (ii) a semi-targeted strategy matching Δm values to a curated list of common biochemical transformations. Applying this to the T. brucei data, they showed that the inferred network is highly non-random and that a small number of Δm values account for a large fraction of observed connections (e.g., dominant differences consistent with hydrogenation/dehydrogenation and small hydrocarbon additions), providing a data-driven way to summarize which transformations most structure the measured metabolome. They further illustrated how such connectivity can support family-level interpretation and putative annotation by extracting subnetworks around known metabolites (e.g., a phosphatidylcholine feature whose neighborhood links a large fraction of strongly covarying masses)[5].

In a mass-difference network, each edge encodes a specific Δm that is consistent with the gain or loss of atoms (or functional groups), i.e., an elemental change that yields a predictable difference in exact mass between two related compounds. Practically, edges are created by matching all pairwise differences between detected features (often *m/z* features) to a transformation list, a catalog of expected Δm values associated with plausible chemical changes. The transformation list can be defined a priori from biochemical knowledge, or it can be guided by recurring Δm values observed in the data. For example, Traquete et al. built sample mass-difference networks using a deliberately small, a priori list of 15 mass-difference "building blocks" (MDBs) selected to represent ubiquitous reactions; they curated this list by searching the BRENDA enzyme database and constrained it to neutral transformations below 80 Da, that is, mass differences corresponding to the net gain or loss of a neutral group of atoms in a

biochemical reaction (e.g., addition of CH2 or O), as opposed to charge- or adduct-related differences, adding extra MDBs only when motivated by the biology (e.g., glycation-related transformations for one dataset)[6]. Note that this restriction to neutral biochemical transformations differs from the list adopted here (Table 4.1), which also incorporates isotope and adduct differences specific to MALDI MSI. In contrast, the review by Amara et al. emphasizes that transformation lists can also be derived systematically from reaction resources such as KEGG (and related metabolic-reaction databases), and illustrates the idea with explicit examples like dephosphorylation represented as the loss of HPO_3 ($\Delta m \approx 79.966$)[7]. Finally, "data-driven" strategies start from the Δm distribution itself: Breitling et al. highlighted that ultra-high mass accuracy makes the network structure strongly non-random and allows dominant Δm patterns to emerge[8], while TMDS-type approaches explicitly compute and rank frequently occurring Δm values as candidates for interpretation[9].

A key advantage of mass-difference networks is that, when edges correspond to well-defined elemental differences, they can support annotation (molecular-formula) propagation: once a node in a connected component is assigned a reliable elemental formula, formulas for neighboring nodes can be inferred by applying the edge's difference formula, provided that chemical plausibility constraints are enforced (elemental ratio bounds, admissible element sets, ion-type consistency, etc.). This "seed-and-propagate" idea is explicitly described in the metabolomics network review[7] and is repeatedly cited as a practical route for feature-level formula assignment in mass-difference graphs. It has also been used as a basis for elemental composition assignment in ultra-high-resolution FT-ICR-MS through mass-difference / Kendrick-analogous network analyses[10] and is discussed as a strategy even for lower-resolution LC-MS when propagation starts from high-confidence annotated metabolites[11].

However, this approach depends strongly on the chosen transformation list: if a relevant Δm is absent, chemically related features will remain disconnected, and missing intermediates (undetected metabolites) can fragment reaction series and create gaps in the reconstructed network[7]. A pragmatic workaround, explicitly stated in that same review, is to include composite transformations that represent multiple successive steps, e.g. adding C2H4 (2×CH2) or C4H8 (2×C2H4) to bridge gaps in alkyl-chain series, at the cost of reduced mechanistic specificity[7].

Several tools implement mass-difference networks. MetaNetter builds and visualizes transformation-based networks for high-accuracy FT-MS data and can incorporate feature correlations across perturbations to support metabolite relationships; MetaNetter 2 further adds adduct and isotope networks plus cross-sample summaries[12,13]. MetNet follows a similar transformation-based strategy but also prioritizes edges using orthogonal evidence such as retention-time shifts and quantitative associations across samples[14].

In many workflows, feature groups are defined as connected components of the mass-difference network, meaning sets of nodes linked by paths of allowed Δm edges. These components are determined directly by graph construction and filtering, rather than by community-detection methods such as Louvain or Leiden[15]. In practice, isolated nodes are often removed, leaving one large connected component and several smaller ones, which can serve as an operational definition of molecular families or transformation-linked groups[6]. To improve specificity, graph construction may be restricted by transformation type, retention-time consistency, or correlation thresholds, reducing spurious links in the resulting components[7].

In mass spectrometry imaging, graph-based representations built from spatial similarity between ion images are increasingly used to group *m/z* features into data-driven modules. A common construction is an ion-image similarity graph, where each node is an *m/z* image and edge weights quantify spatial similarity (often Pearson correlation or cosine similarity), followed by graph clustering (community detection) or classical image clustering. For example, Wüllems et al. construct a similarity matrix between selected *m/z* images, sparsify it by thresholding, and then apply community detection via modularity optimization (spectral approach) to identify "molecular communities" of laterally co-localized ions; they demonstrate the approach on barley seed and human glioblastoma MSI data and provide an interactive exploration tool (GRINE) for inspecting communities and substructures[16].

Spatial correlation can also be exploited for chemical deconvolution/identification, not only for "functional modules". Dong et al. introduce PICA, where for a chosen "bait" ion they compute pixel-wise Pearson correlations to all other *m/z* features; they show that highly co-localized ions (PCC ≥ 0.9) are predominantly in-source fragments, isotopes, adducts, or multimers of the same metabolite, enabling the construction of "pseudo-MS/MS spectra" directly from MSI data for improved metabolite identification, without requiring an explicit mass-difference constraint during grouping[17].

Earlier, Alexandrov et al. proposed to cluster all *m/z* images according to spatial similarity as a way to summarize what "all ion images look like" in a dataset, illustrating the approach on MALDI-MSI rat brain and human larynx carcinoma data[18]. While this work is not framed as a network method, it established spatial co-localization as a powerful unsupervised signal for grouping ions in MSI.

Although ion-image colocalization graphs have been used to help interpret MSI signals, several of these approaches rely only on spatial similarity and do not impose explicit mass-difference constraints on the edges. In such a purely spatial graph, distinct molecular families that share the same tissue localization (e.g., co-localized lipids and matrix-related ions) may be grouped together, because spatial correlation alone cannot distinguish whether co-varying peaks are related by chemistry or merely co-present in the same region. Introducing mass-difference relationships between *m/z* features provides the missing chemical context: it becomes possible to interpret how co-localized signals differ in composition (e.g., isotope spacing, adduct shifts, oxidation, homologous series), and therefore to separate families that overlap spatially but differ in their Δm structure.

In this section, we combine the mass-difference and spatial-coherence ideas introduced above by formalizing them as a graph-construction problem. Starting from a user-defined set of Δm transformations, we build a mass-difference graph over the detected peaks. We then retain only the edges that are also supported by high ion-image similarity, so that the resulting groups of peaks are both chemically consistent and spatially co-localized. We demonstrate the approach on a bacterial co-culture MSI dataset (*Pseudomonas–Bacillus*) and show that it recovers several interpretable molecular families, which can be inspected through their associated ion images and within-graph structure. This method is implemented in the Python package *deltamsi*, described in chapter 2 (https://github.com/LaRoccaRaphael/deltamsi).

4.2.1 Material

Indium tin oxide glass slides (Bruker, Bremen, Germany) are coated, with 1/5 strength CAA agar medium (1/5 casamino acids (Becton, Dickinson and Company) 1 g/L, $MgSO_4.7H_2O$ 0.05 g/L, K_2HPO_4 0.2 g/L, agar (Bacto TM) 1.5 g/L). Slides are deposited in sterile petri dishes on top of a humidified filter paper disc. *Pseudomonas sessilinigenes* CMR12a and *Bacillus velezensis* GA1. The slide was divided virtually into 2 parts, with one strain inoculated on the first and the other on the second to create a confrontation line between the 2 strains. The inoculated slides are cultured at 18°C for 16-18h in the dark, before being dried in a vacuum chamber for 48 hours at room temperature. The slides are then covered by the HCCA matrix and imaged on a SolariX XR 9.4T (Bruker). A single image is selected and converted to centroided imzML format, a standard MSI data format, using FlexImaging 5.0.

The MSI imzML is then processed to obtain ion images of the 3,919 most intense MSI total-ion-spectrum peaks. Ion images are generated by summing mass signals within 0.01 Da of the average mass spectrum peaks' centroids. This value is selected to capture the peak width in the average spectrum at *m/z* 1,000 (resolution of 100,000 at *m/z* 1,000). All ion images are first standardized, and then pixel values below the 0.5 quantile of the image are set to 0 for each ion image. Then, we apply a 2D median filter to each ion image. Finally, the image similarities between pairs of ion images are computed using cosine similarity.

The similarity is computed pixel-wise: both images are flattened into vectors with a fixed pixel ordering, so a given index refers to the same physical location in both, and the score sums the products of the intensities at each matching pixel, normalized by the image magnitudes. Cosine similarity and Pearson correlation are identical except that Pearson first subtracts each image's mean. This matters for ion images, which are non-negative and dominated by near-zero background: with Pearson, mean-centering turns these shared background pixels into matched negative deviations that inflate the

correlation, so two unrelated but equally sparse ions can appear strongly correlated simply because they share an empty background. Cosine ignores this background, and following hotspot thresholding and median filtering it was found to best reproduce expert assessments of ion-image co-localization[19].

Although this measure uses no coordinates or neighborhood relations, it still reflects spatial co-localization: because the pixel ordering is shared, a high score requires both ions to be intense at the same locations. Co-localized ions have their intense pixels at coinciding indices and score highly; spatially distinct ions score near zero. The only step using local spatial structure is the median filter, which suppresses single-pixel noise.

4.2.2 MS spectrum as a graph of mass differences

In this section, we describe how to process a mass spectrum into a graph of mass differences and how to apply this representation to MSI. A mass spectrum is usually presented as a list of peaks, each representing an ion with a specific *m/z* value and intensity.

Given a list of allowed mass differences D, we can represent a mass spectrum as a graph G with edges E and nodes V.

$$D = d_1, d_2, d_3, \dots, d_l$$
$$\text{Nodes}: V = v_1, v_2, \dots, v_n$$
$$\text{Edges}: E = (v_i, v_j): \mid\mid m_i - m_j \mid - d \mid \leq t_1, d \in D \quad (3)$$

Each mass spectrum peak is represented by a node v_i where i ranges from 1 to n, the total number of peaks in the spectrum. The *m/z* value of the peak v_i is denoted by m_i. The peak intensity is not conserved in this representation. An edge connects two nodes v_i and v_j when the absolute difference between their *m/z* values matches any element of D, within the mass tolerance t_1 (Figure 4.4 a-b).

Such a graph representation can be applied to an MSI experiment. In this case, the nodes include the *m/z* of the corresponding peak and its ion image. I_i and I_j are the ion images associated with nodes v_i and v_j , respectively. It is then possible to add another condition to create an edge, such as:

$$\text{Edges: } E = (v_i, v_j) : || \, m_i - m_j \, | - d \, | \leq t_1, d \in D, \;\; \text{cosine}_{(I_i, I_j)} > t_2 \quad (4)$$

In practice, edges are created in two steps. A pair of nodes (v_i, v_j) first qualifies as a candidate edge when its *m/z* difference matches an allowed d ∈ *D* within t1 (Equation 3); the cosine similarity between the two ion images is then computed, and the edge is retained only if it exceeds t_2 (Equation 4). The similarity, therefore, acts as a binary gate that confirms or rejects mass-difference candidates rather than creating edges itself, and retained edges are unweighted; the cosine value is not stored, so every edge contributes equally.

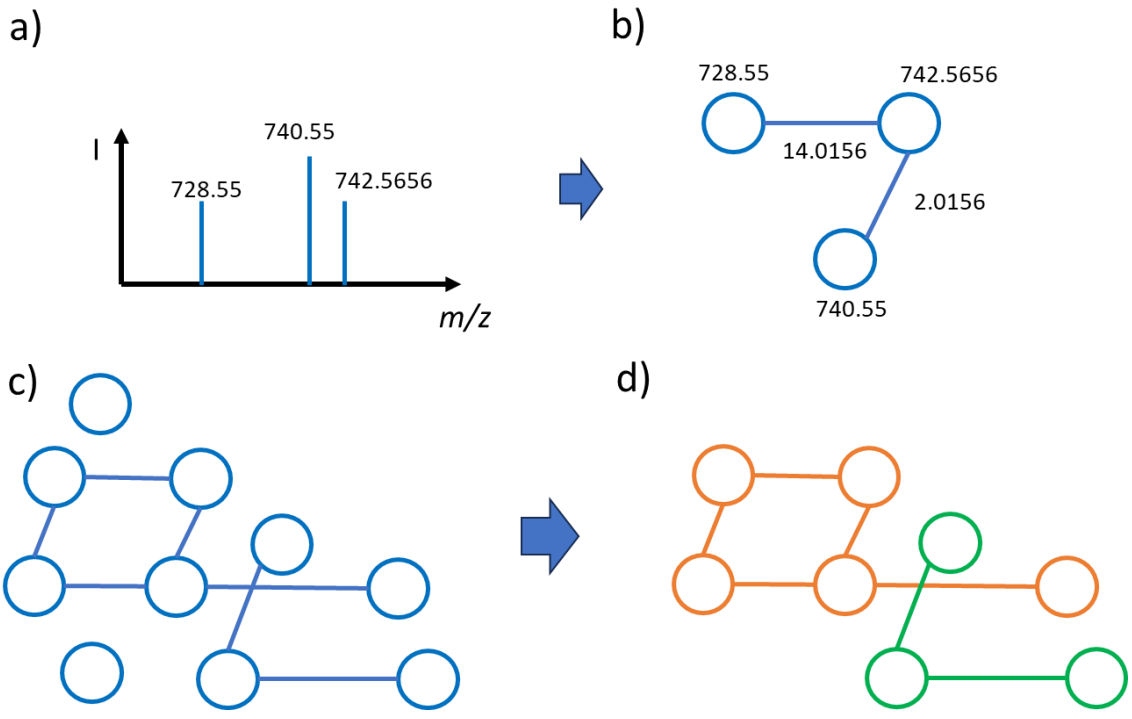


Figure 4.4. *a-b generation of a graph of mass differences from a mass spectrum containing three peaks. This example illustrates 3 nodes v_1, v_2, v_3 that correspond to 3 peaks with the following masses $m_1 = 728.5500, m_2 = 742.5656, m_3 = 740.5500$. A graph of mass difference is generated from a list D of allowed mass differences $D = d_1 = 14.01565, d_2 = 2.01565$. c)-d) Filtration of nodes based on their cluster sizes. The graph in c) contains ten nodes. In d), clusters of size 1 are removed from the graph, leaving a graph of size 8 with 2 clusters of size 5 (orange) and 3 (green).*

In this analysis, t_1 and t_2 have been selected to 0.0025 Da and 0.3, respectively. t_1 should be adapted to the mass accuracy of the mass spectrometer in the range of *m/z* of interest. t_1 corresponds to an error of 2.5 ppm at *m/z* 1,000. For mass spectrometry imaging, we should adapt t_1 to the mass accuracy of the average spectrum as the presence of mass shift in individual spectrum impacts the average mass accuracy (*see Chapters 2 and 3*). Recalibration, or, to a lesser extent, alignment, should be applied to the MSI before creating the edges. In practice, t_2 is set as a compromise: too low a value connects coincidentally similar ions, while too high a value fragments genuine families whose members have overlapping but non-identical distributions (e.g. the GPC unsaturation series in Figure 4.3). It is chosen by inspecting the distribution of pairwise cosine values together with representative image pairs across that range, and placing the threshold where pairs visually transition from co-localized to unrelated. Pairs with known relationships provide useful anchors: isotopes or adducts of the same ion are necessarily co-localized and set a lower bound on the similarity expected for true edges, whereas pairs from clearly different regions indicate the level produced by chance. This inspection led to t_2 = 0.3, low enough to preserve within-family variation while still excluding spatially unrelated ions, and the value can be confirmed by checking that the recovered clusters remain stable when t_2 is varied.

The selection of list D is crucial since a too-large set of mass differences increases the risk of connecting unrelated compounds. We have selected the following 13 mass differences in table 4.1, all those values are the labels of the mass differences in D. The mass differences listed in Table 4.1 are frequently observed in MALDI MSI experiments.

Mass Difference *(Da)*	Formula	Reason
1.00335	$^{13}C - ^{12}C$	Presence of one ^{13}C isotope
2.00670	$^{13}C_2 - ^{12}C_2$	Presence of two ^{13}C isotopes
2.00425	$^{18}O - ^{16}O$	Presence of one ^{18}O isotope
2.01565	H_2	Addition or reduction of a double bond
13.97925	O – 2H	O – 2H
14.01565	CH_2	Addition of a CH_2 group
15.97394	K – Na	Difference between Na and K adducts
15.99491	O	Oxidation
21.98194	Na – H	Difference between Na and H adducts
28.03130	C_2H_4	Addition of two CH_2 groups
37.95588	K – H	Difference between K and H adducts
56.06260	C_4H_8	Addition of four CH_2 groups
189.04259	$C_{10}H_7NO_3$	Presence of a matrix adduct (HCCA)

Table 4.1. *Table of mass differences D considered for transforming the spectra into a graph of mass differences.*

An edge E between 2 nodes v_i, v_j can have multiple labels if multiple d meet the condition : $|| m_i - m_j | - d | \leq t_1$. For example, the mass differences between $^{13}C_2 - ^{12}C_2$ and $^{18}O - ^{16}O$ are both equal to 0.0024. This value is lower than the threshold t_1, which suggests that edges may exist as they simultaneously match both $^{13}C_2 - ^{12}C_2$ and $^{18}O - ^{16}O$. This situation creates ambiguous edge annotations. Since edge labels are not used during graph creation, they do not affect the clustering results. $C_{10}H_7NO_3$ has also been added to this list as it corresponds to the molecular formula of HCCA, and HCCA adducts are often observed in MALDI for various compounds [20,21].

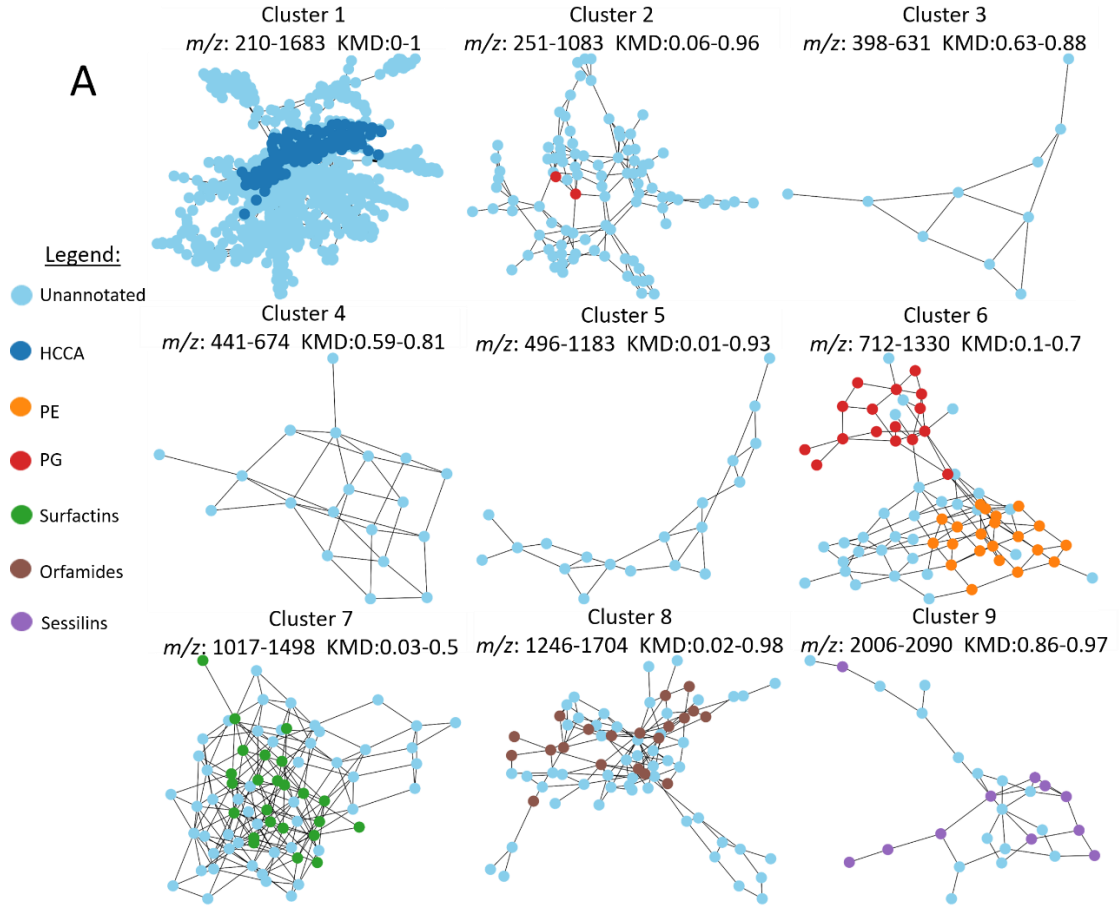
Once the graph of mass differences G is created, analysis begins by identifying all disconnected subgraphs in G. A disconnected subgraph is a subset of nodes connected only within the subset, forming its own connected component. These nodes are, by definition, completely disconnected from all other nodes of G. In other words, a

disconnected subgraph consists of a group of nodes isolated from the rest of the graph. Here, disconnected subgraphs are called clusters and should ideally correspond to different molecular families. We measure a cluster's importance (prioritize it) by simply counting the number of nodes it contains. For example, in Figure 4.4, c-d, we have one graph. This graph identifies two clusters (disconnected subgraphs), highlighted in orange and green. We can then sort the clusters by their node counts and filter out those that are too small.

4.2.3 Application to bacterial confrontation MSI

In this section, we applied our data processing method to an MSI of co-cultured bacteria (see Materials and Methods section). We evaluate our method based on its ability to retain informative ions and cluster molecular families. In our approach, we first analyzed the method's effectiveness at filtering out noise signals. After MSI pre-processing, we detect 3,919 peaks in the MSI average spectrum. We evaluate the quality of each peak ion image by utilizing the structural measure developed in Chapter 2. This spatial measure quantifies the randomness of an ion image's spatial distribution, yielding a score between 0 and 1. An ion image with a score of 0.6 or higher is considered structured. This threshold has been selected so we can still perceive the morphology of bacterial colonies in the ion images with scores above it. Once the graph of mass differences is generated, it is cleaned by removing all clusters with fewer than 10 nodes. The nine clusters of interest that emerged from this cleaning step are represented in Figure 4.5.A. Combining 3,831 edges and 1,655 nodes, they contain 1,655 out of the 3,919 peaks, *i.e.* 42% of the global MSI peaks. As a comparison, Figure 4.5.B shows the annotated Kendrick plot (*kb*= CH_2) of all the 3,919 peaks of the MSI, whereas Figure 4.5.C shows the same plot but drawn after the filtration step, resulting in a cleaner Kendrick plot. This revised plot provides better separation between molecular families, indicating the effectiveness of the noise-removal step. Indeed, the 9 clusters retain 80% (1,133/1,406) of the peaks with a spatial structure over 0.6 in the raw image, 91% of the

MSI's total signal (sum of all peak raw intensities). The method enables the efficient filtering of the noisiest signals from the MSI.



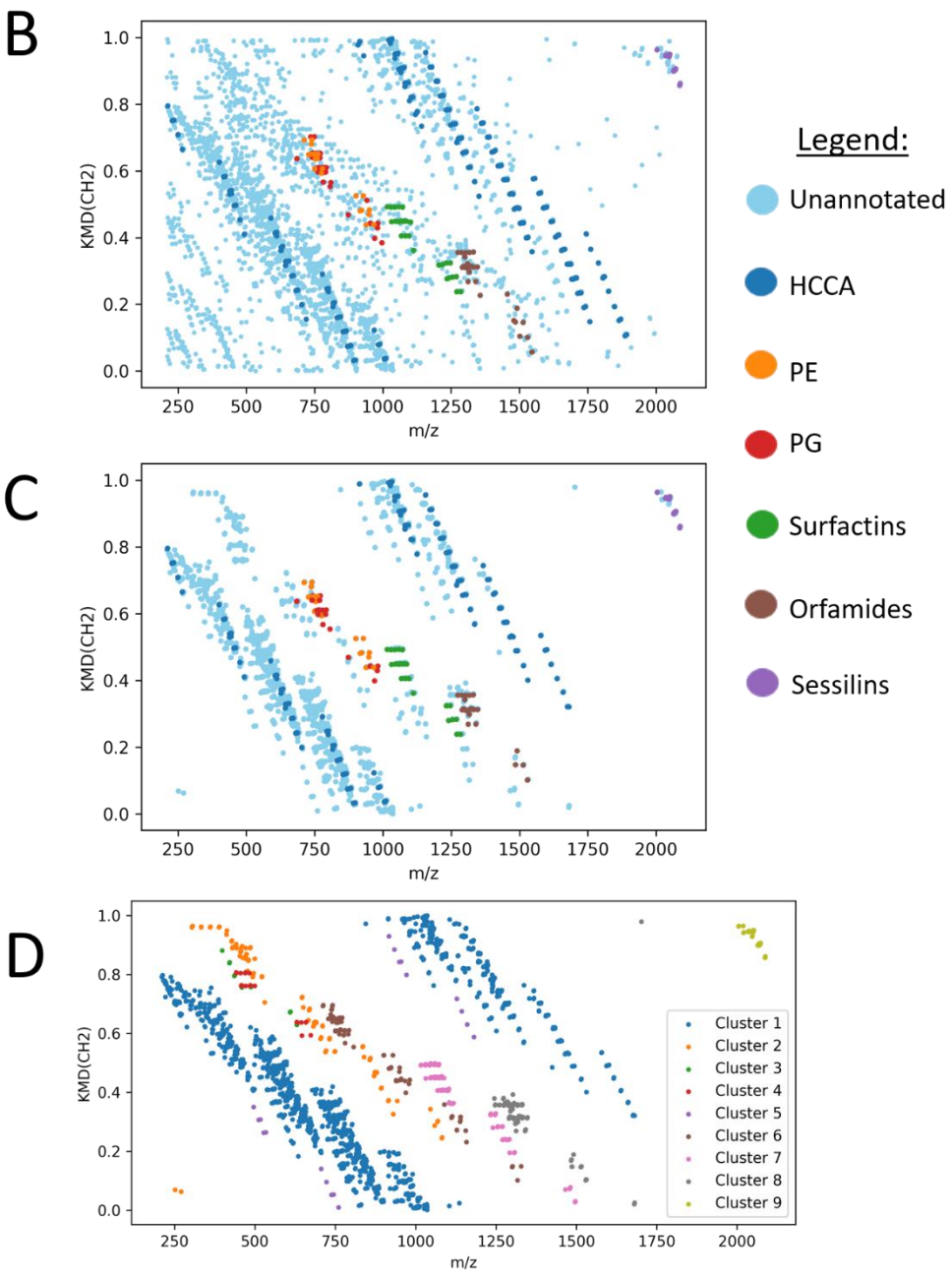


Figure 4.5. *Graph's cluster analysis of a bacterial confrontation, MSI. A) Representation of the 9 most significant clusters after cluster filtration. B) Kendrick plot base CH_2 of all 3,919 peaks in the raw MSI average spectrum. C) Kendrick plot base CH_2 of all node m/z values after cluster filtration. Annotated peaks/nodes on plot A have different colors according to their associated molecular family. D) Kendrick plot base CH_2 of each node from each cluster. Cluster size (number of nodes) 1-9 (1269, 97, 10, 19, 19, 67, 75, 72, 27) (86%, 2.1%, 0.1%, 0.3%, 0.1%, 0.1%, 0.4%, 1.3%, 0.5%).*

This second section evaluates the method's capacity to isolate molecular families. Many molecular families are known to be produced and detected from this bacterial confrontation, such as:

- surfactins produced by *Bacillus velezensis GA1*,
- sessilins and orfamide produced by *Pseudomonas sessilinigenes* CMR12a ,
- phospholipids present in the bacterial membranes,
- HCCA cluster ions from the MSI sample preparation.

Ions corresponding to each family are annotated and highlighted in the legend of Figure 4.5. Ions are annotated based on a list of exact masses of different molecular formulae with a tolerance of +/- 0.005 Da. The list was built by combining the masses of ions corresponding to HCCA clusters, various lipopeptides (orfamides, sessilins), and phospholipids (PE, PC) known to be produced by this *Pseudomonas* strain[22], each time computed with different adducts and matrix adducts. The phospholipid signals from *Bacillus velezensis* were too weak to detect or annotate. Only two nodes were identified and annotated as PG lipids on the *Bacillus* side. However, those annotations are uncertain due to the low number of annotated peaks (see discussion below). The only *Bacillus* signals that can be confidently identified are the nodes annotated as surfactins. The weaker signal from *Bacillus* lipids could be explained by fewer cells per pixel, since their colonies are smaller than those of *Pseudomonas.*

As recently discussed (Figure 4.5), the clustering method shows a high capacity to separate molecular families, with each cluster containing a single family. Figure 4.6 summarizes each cluster by a single average ion image, obtained as the pixel-wise mean of the standardized ion images of all nodes in the cluster. This collapses an entire molecular family into one representative map, so a coherent cluster yields a sharp, family-specific distribution rather than a blurred one. Each panel is one cluster, oriented as in the confrontation assay: the left half ($x < 120$) carries the *Bacillus velezensis* GA1 colonies and the right half ($x > 120$) the *Pseudomonas sessilinigenes* CMR12a colonies,

with the color scale running from low to high normalized intensity. Read against the family assignments of Figure 4.5, the panels show directly where each family localizes.

Figure 4.6 shows that the method successfully clustered all the main molecular families in the six most significant clusters. Indeed, Cluster 1 is the largest, accounting for 86% of the raw signal and 1,269 nodes. It contains, unsurprisingly, all the matrix-associated ions and clusters (HCCA). The remaining unannotated nodes are not related to the bacteria (media), as their average ion image (Figure 4.6-Cluster 1) shows a spatial distribution around the bacterial colonies. A possible origin of those peaks could be background ions (molecules from the agar media) combined with HCCA adducts. The graph structure of Cluster 1 (Figure 4.5.A) reveals that nodes annotated as HCCA are closely connected, forming a single community. However, there are indications of other node communities, as other groups of nodes are closely connected in *Cluster 1*. Further exploration of community separation may reveal additional molecular families. Clusters 2 and 3 consist of nodes that are not annotated, except for two of them in cluster 2, annotated as phosphatidylglycerol lipids (PG). The cluster's spatial distribution indicates a prevalence of ions detected on the *Bacillus* GA1 side (Figure 4.6 – Cluster 2,3). Since surfactins are the most intense signals observed on the *Bacillus* side, and since the *Cluster 2 and 3* have the same spatial distribution as the surfactins cluster (Cluster 7, see below) and contains ions with lower masses than surfactins (Figure 4.5.D), we can hypothesize that this cluster contains surfactin fragment ions. Importantly, fragment ions are not expected to connect to their parent cluster in the graph, since the mass losses produced by fragmentation are not part of D; fragments therefore form their own connected components rather than joining the surfactin cluster. This hypothesis could be obviously confirmed by extracting the media and analyzing the extract by LC-MS/MS, looking for these specific *m/z* ions. Clusters 4 and 5 contain only unannotated nodes. From their spatial distributions, Cluster 4 and Cluster 5 mostly contain background ions. Clusters 7,8, and 9 are the most interesting ones as

they capture each a different family of lipopeptides. In Figure 4.6, the ion image of each cluster corresponds to what was already observed for this kind of molecule:

- Cluster 7 contains ions detected on the *Bacillus* GA1 side and annotated as surfactins (Figure 4.6 – Cluster 7).
- *Cluster 8* contains ions correlated only to *Pseudomonas* CMR12a and annotated as orfamides. These molecules diffuse out of the bacterial colonies (Figure 4.6 – Cluster 8).
- Cluster 9 contains ions that correlate only to *Pseudomonas* CMR12a and are annotated as sessilin ions. Sessilins visibly diffuse less than orfamides (Figure 4.6 – Cluster 9).

Some unannotated nodes within each cluster correspond to isotopes (only the monoisotopic peak is annotated and colored). Further investigation is required to determine whether non-isotopic unannotated nodes correspond to metabolite variants of the cluster's main annotation. For example, should we consider all the unannotated nodes of cluster 9 to be sessilin ions?

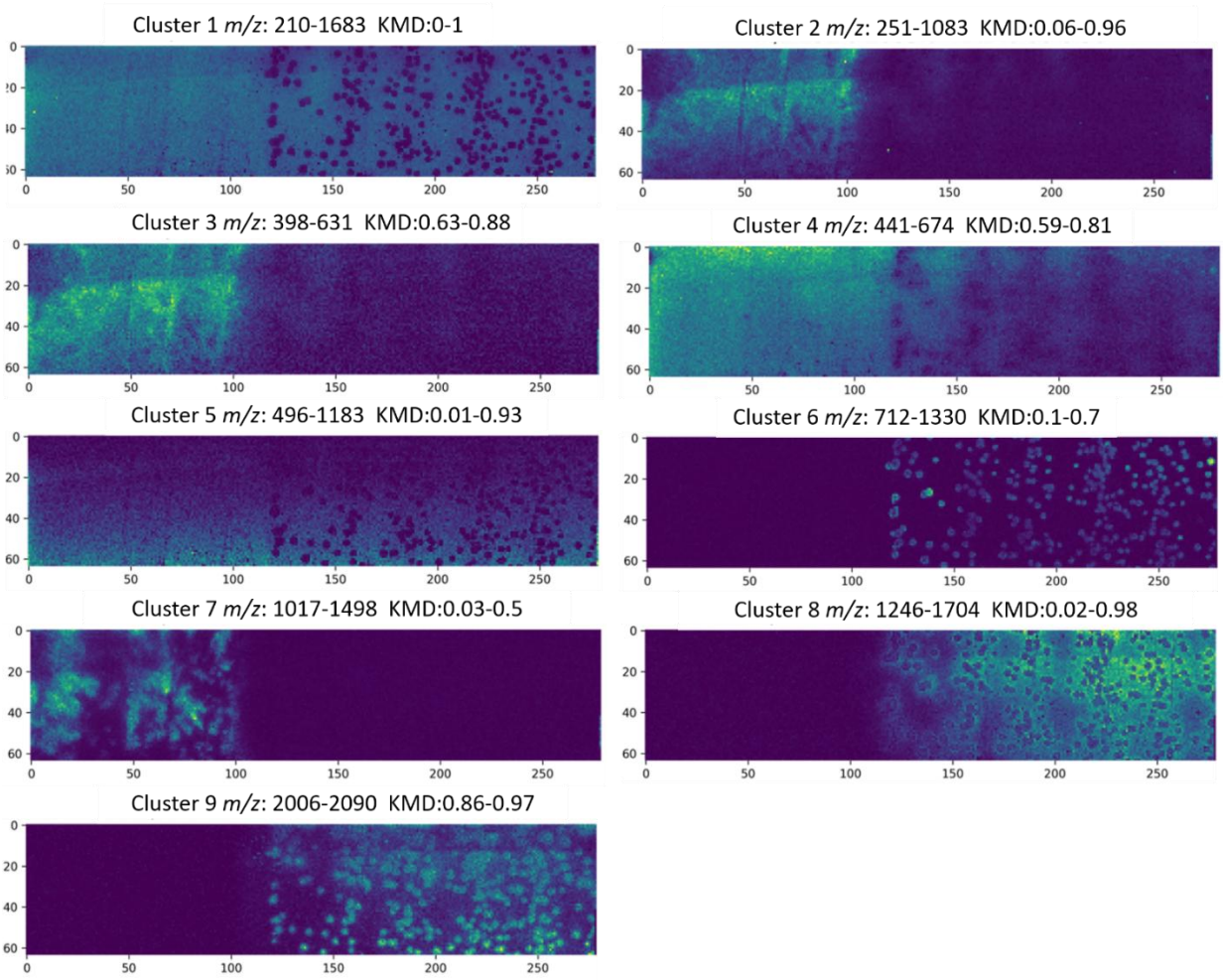


Figure 4.6. *Average ion image of each of the 9 clusters obtained from the graph of mass differences of the bacterial confrontation MSI. For each cluster, the average ion image is the pixel-wise mean of the standardized ion images of all its nodes, giving a single map of the family's spatial distribution. In every panel, the left side (x < 120) shows the Bacillus velezensis GA1 colonies and the right side (x > 120) the Pseudomonas sessilinigenes CMR12a colonies; the color scale represents normalized ion intensity. Cluster identities: 1 = matrix (HCCA) and media background; 2 and 3 = putative surfactin fragments (Bacillus); 4–5 = background; 6 = phospholipids PE/PG (Pseudomonas); 7 = surfactins (Bacillus); 8 = orfamides (Pseudomonas); 9 = sessilins (Pseudomonas).*

Unlike the MSKF approach, which yields multiple clusters for a single molecular family (i.e., multiple horizontal sequences of points), this process further reduces complexity by producing a single cluster per family. For example, our method yields a single cluster containing all surfactin-annotated ions. If we used the MSKF method with *kb*=CH_2, more clusters would be created for the same molecular family. For instance, *Cluster 7* includes

more than 20 MSKF clusters (10 without isotopes) (see Figure 4.5.D pink dots). MSKF divides surfactin-annotated ions into 7 MSKF clusters (Figure 4.5.C green points). As a result, our approach simplifies the manual investigation by reducing the number of clusters per molecular family.

The only cluster that contains annotation labels with two different molecular families is Cluster 6, which combines phospholipids PE and PG. In Figure 4.6, the ion image of that cluster is well correlated with *Pseudomonas CMR12a* (*Bacillus* signal is too weak), confirming the phospholipid annotations, since those ions are primarily detected in bacterial membranes. However, even if PE and PG are found in the same cluster, they are well separated in the graph and form 2 communities, one for each family (Figure 4.5, A). A community in a graph is a set of nodes that are highly connected among themselves but poorly connected to other sets of nodes. Here, a community differs from a cluster since edges can connect multiple communities. In this cluster, the 2 communities are linked by 2 edges; one connects 2 unannotated nodes, and the other connects a node annotated as PE [$C_{39}H_{76}NO_8P$ +Na+K-H] and another as PG [$C_{39}H_{75}O_{10}P$ +2Na-H]. An edge exists between those two nodes since their mass difference is precisely 1.0046, where our method considers an ^{13}C Isotope mass difference. It is possible to make the creation of such bad edges more robust by detecting communities within each cluster.

Community detection could be applied to further exploit graph topology and separate densely connected substructures corresponding to more refined molecular families. In this context, Leiden clustering could be used to identify groups of nodes that are strongly connected internally while remaining weakly connected to the rest of the graph, thereby improving the robustness of the analysis against false-positive edges[15]. This strategy could help reveal more intricate molecular families within a cluster and is proposed in the Python package *deltamsi*.

4.2.4 Discussion and Conclusion

Representing MSI as a graph has multiple advantages over classical Kendrick analysis and MSKF. MSKF combines manual selection followed by peak clustering for a single mass difference, and it allows a robust selection of molecular families by avoiding clustering unrelated compounds through manual inspection of the Kendrick plot. In the present method, we removed the manual selection and replaced it with a ranking of clusters based on the cluster size. The robustness of the selection is ensured by two new conditions: 1) comparing the spatial distributions of two nodes and 2) using multiple mass differences. Those conditions result in the highest signal compression, meaning that most spectral information is grouped into fewer clusters. This compression allows the raw MSI signal to be inspected more rapidly and thoroughly.

The method can be compared to existing compression methods, such as PCA (Principal Component Analysis). PCA is often used in MSI to reduce image complexity by grouping spectral information into components that capture most of the spatial signal variation in an MSI experiment, with the first components accounting for most of the MSI signal variance. Then, the first components can be used to explore the most common ion spatial distribution instead of manually inspecting all the MSI's ion images. Similarly, our method allows us to observe the most interesting signals not only based on their spatial distribution (as in PCA) but also on their spectral similarity.

We demonstrate those advantages by analyzing a bacterial confrontation MSI. The method successfully clustered all the main molecular families into the six largest clusters. Therefore, it is sufficient to inspect the spatial distribution of 6 ion images to understand the dataset's overall chemical composition. MSKF produces less-compressed information and more clusters within the same family, since it cannot cluster ions separated by multiple mass differences. Our new method simultaneously performs adduct and isotope deconvolution.

While this method is a real improvement for investigating complex MSI data, it still has limitations, mainly in its reliance on edge importance and false-positive mass-difference matching. False positive mass difference matching can be defined as a mass difference between 2 nodes found in *D,* but where the corresponding molecular formula label in *D* differs from the real molecular formula difference. On the other hand, each edge has the same weight/importance in the graph. For example, in this analysis, a node can be linked by a mass difference of the ^{13}C isotope and the $^{13}C_2$ isotope. However, according to the molecular formula of the node, we are not expecting a node with an edge of $^{13}C_2$ isotope mass difference without detecting the single ^{13}C isotope ion. Here, a pair of nodes linked by ^{13}C isotope or $^{13}C_2$ isotopes has the same importance, but ideally, we should prioritize the ^{13}C isotope over the $^{13}C_2$ isotopes. An example of prioritization could be sorting cluster size by counting only nodes with at least ^{13}C Isotope edges. A second disadvantage is that we cannot adapt D (the set of mass differences) according to the node's *m/z*. For example, we do not expect H_2 and CH_2 mass differences within nodes of matrix ions (HCCA). However, those edges are found to exhibit potentially false-positive mass differences, leading to heterogeneous clusters. Finally, we should adjust the importance of edges based on their sequence in the graph. For example, a cluster containing five nodes connected only by CH_2 should be more important than one containing five nodes connected only by Na-H mass differences, since the latter is likely due to false-positive difference matches. As already discussed for *Cluster 6* (Figure 4.5), community detection should also be applied to exploit graph topology, to be more robust against false-positive mass-difference matches, and to reveal communities containing different molecular families.

4.2.5 Introducing machine learning for processing graphs of mass differences

The graph-of-mass-differences framework developed in this chapter organizes peaks into chemically meaningful families by connecting ions related through specific Δm patterns (e.g., isotopes, adducts, homologous series). While this unsupervised representation is powerful for compressing spectral complexity and guiding interpretation, it still relies on hand-crafted rules: the set of allowed relations is fixed, all edge types are treated equally, and spurious peak matchings can propagate through the graph. A natural next step is therefore to replace heuristic processing with a learned model that can exploit graph topology and weight relations rather than treating all edge types as equivalent.

In MSI, supervised classification and segmentation are often performed on vectorized spectra, where each feature corresponds to the intensity of a peak or a binned *m/z* channel after alignment. In this common setting, the model can learn associations between feature intensities and classes, but it does not explicitly leverage the structure of high-resolution mass information, for example, that a known mass difference may link two peaks, or that repeated relations form characteristic motifs such as CH_2 chains. As a result, chemically related peaks are considered independent variables, and the model may depend on fragile, dataset-specific feature combinations rather than more consistent molecular-family patterns.

In Chapter 5, we build on the graph representation introduced here and propose a supervised graph neural network (GNN) framework. Rather than grouping peaks into molecular families in an unsupervised manner, as done in this chapter, we apply GNNs directly to graphs of mass differences to perform a different task: the robust classification of mass spectra. By learning relation-specific weights and aggregating information across connected nodes, the GNN integrates both peak attributes and

graph topology, allowing it to recognize informative motifs such as homologous series, down-weight relations that frequently produce false-positive matches, and thereby improve robustness to peak dropout and small mass shifts for mass spectrum classification. Crucially, although the model is trained at the spectrum level, its learned representation can be repurposed to interpret which signals drive each prediction: by assigning a class-relevance score to every node, we extract the *m/z* values that the model associates with each class. This effectively recasts the spectrum-level classification as an *m/z*-level classification, recovering molecular families linked to each biological class or condition, the same level of information targeted by the unsupervised analysis of this chapter, but now obtained in a supervised, task-driven manner.

REFERENCES:

(1) Alexandrov, T. Spatial Metabolomics and Imaging Mass Spectrometry in the Age of Artificial Intelligence. *Annual Review of Biomedical Data Science* **2020**, *3* (1), 61–87. https://doi.org/10.1146/annurev-biodatasci-011420-031537.

(2) Palmer, A.; Phapale, P.; Chernyavsky, I.; Lavigne, R.; Fay, D.; Tarasov, A.; Kovalev, V.; Fuchser, J.; Nikolenko, S.; Pineau, C.; Becker, M.; Alexandrov, T. FDR-Controlled Metabolite Annotation for High-Resolution Imaging Mass Spectrometry. *Nature Methods* **2017**, *14* (1), 57–60. https://doi.org/10.1038/nmeth.4072.

(3) Kendrick, Edward. A Mass Scale Based on CH_2 = 14.0000 for High Resolution Mass Spectrometry of Organic Compounds. *Anal. Chem.* **1963**, *35* (13), 2146–2154. https://doi.org/10.1021/ac60206a048.

(4) Kune, C.; McCann, A.; Raphaël, L. R.; Arias, A. A.; Tiquet, M.; Van Kruining, D.; Martinez, P. M.; Ongena, M.; Eppe, G.; Quinton, L.; Far, J.; De Pauw, E. Rapid Visualization of Chemically Related Compounds Using Kendrick Mass Defect As a Filter in Mass Spectrometry Imaging. *Anal. Chem.* **2019**, *91* (20), 13112–13118. https://doi.org/10.1021/acs.analchem.9b03333.

(5) Breitling, R.; Ritchie, S.; Goodenowe, D.; Stewart, M. L.; Barrett, M. P. Ab Initio Prediction of Metabolic Networks Using Fourier Transform Mass Spectrometry Data. *Metabolomics* **2006**, *2* (3), 155–164. https://doi.org/10.1007/s11306-006-0029-z.

(6) Traquete, F.; Luz, J.; Cordeiro, C.; Sousa Silva, M.; Ferreira, A. E. N. Graph Properties of Mass-Difference Networks for Profiling and Discrimination in Untargeted Metabolomics. *Front. Mol. Biosci.* **2022**, *9*, 917911. https://doi.org/10.3389/fmolb.2022.917911.

(7) Amara, A.; Frainay, C.; Jourdan, F.; Naake, T.; Neumann, S.; Novoa-del-Toro, E. M.; Salek, R. M.; Salzer, L.; Scharfenberg, S.; Witting, M. Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation. *Front. Mol. Biosci.* **2022**, *9*, 841373. https://doi.org/10.3389/fmolb.2022.841373.

(8) Breitling, R.; Ritchie, S.; Goodenowe, D.; Stewart, M. L.; Barrett, M. P. Ab Initio Prediction of Metabolic Networks Using Fourier Transform Mass Spectrometry Data. *Metabolomics* **2006**, *2* (3), 155–164. https://doi.org/10.1007/s11306-006-0029-z.

(9) Kunenkov, E. V.; Kononikhin, A. S.; Perminova, I. V.; Hertkorn, N.; Gaspar, A.; Schmitt-Kopplin, P.; Popov, I. A.; Garmash, A. V.; Nikolaev, E. N. Total Mass Difference Statistics Algorithm: A New Approach to Identification of High-Mass Building Blocks in Electrospray Ionization Fourier Transform Ion Cyclotron Mass Spectrometry Data of Natural Organic Matter. *Anal. Chem.* **2009**, *81* (24), 10106–10115. https://doi.org/10.1021/ac901476u.

(10) Tziotis, D.; Hertkorn, N.; Schmitt-Kopplin, Ph. Kendrick-Analogous Network Visualisation of Ion Cyclotron Resonance Fourier Transform Mass Spectra: Improved Options for the Assignment of Elemental Compositions and the Classification of Organic Molecular Complexity. *Eur J Mass Spectrom (Chichester)* **2011**, *17* (4), 415–421. https://doi.org/10.1255/ejms.1135.

(11) Forcisi, S.; Moritz, F.; Lucio, M.; Lehmann, R.; Stefan, N.; Schmitt-Kopplin, P. Solutions for Low and High Accuracy Mass Spectrometric Data Matching: A Data-Driven Annotation Strategy in Nontargeted Metabolomics. *Anal. Chem.* **2015**, *87* (17), 8917–8924. https://doi.org/10.1021/acs.analchem.5b02049.

(12) Burgess, K. E. V.; Borutzki, Y.; Rankin, N.; Daly, R.; Jourdan, F. MetaNetter 2: A Cytoscape Plugin for Ab Initio Network Analysis and Metabolite Feature Classification. *Journal of Chromatography B* **2017**, *1071*, 68–74. https://doi.org/10.1016/j.jchromb.2017.08.015.

(13) Jourdan, F.; Breitling, R.; Barrett, M. P.; Gilbert, D. MetaNetter: Inference and Visualization of High-Resolution Metabolomic Networks. *Bioinformatics* **2008**, *24* (1), 143–145. https://doi.org/10.1093/bioinformatics/btm536.
(14) Naake, T.; Fernie, A. R. MetNet: Metabolite Network Prediction from High-Resolution Mass Spectrometry Data in R Aiding Metabolite Annotation. *Anal. Chem.* **2019**, *91* (3), 1768–1772. https://doi.org/10.1021/acs.analchem.8b04096.
(15) Traag, V. A.; Waltman, L.; Van Eck, N. J. From Louvain to Leiden: Guaranteeing Well-Connected Communities. *Sci Rep* **2019**, *9* (1), 5233. https://doi.org/10.1038/s41598-019-41695-z.
(16) Wüllems, K.; Kölling, J.; Bednarz, H.; Niehaus, K.; Hans, V. H.; Nattkemper, T. W. Detection and Visualization of Communities in Mass Spectrometry Imaging Data. *BMC Bioinformatics* **2019**, *20* (1), 303. https://doi.org/10.1186/s12859-019-2890-6.
(17) Dong, Y.; Shachaf, N.; Feldberg, L.; Rogachev, I.; Heinig, U.; Aharoni, A. PICA: Pixel Intensity Correlation Analysis for Deconvolution and Metabolite Identification in Mass Spectrometry Imaging. *Anal. Chem.* **2023**, acs.analchem.2c04778. https://doi.org/10.1021/acs.analchem.2c04778.
(18) Alexandrov, T.; Chernyavsky, I.; Becker, M.; Von Eggeling, F.; Nikolenko, S. Analysis and Interpretation of Imaging Mass Spectrometry Data by Clustering Mass-to-Charge Images According to Their Spatial Similarity. *Anal. Chem.* **2013**, *85* (23), 11189–11195. https://doi.org/10.1021/ac401420z.
(19) Ovchinnikova, K.; Stuart, L.; Rakhlin, A.; Nikolenko, S.; Alexandrov, T. ColocML: Machine Learning Quantifies Co-Localization between Mass Spectrometry Images. *Bioinformatics* **2020**, *36* (10), 3215–3224. https://doi.org/10.1093/bioinformatics/btaa085.
(20) Burguet, P.; La Rocca, R.; Kune, C.; Tellatin, D.; Stulanovic, N.; Rigolet, A.; Far, J.; Ongena, M.; Rigali, S.; Quinton, L. Exploiting Differential Signal Filtering (DSF) and Image Structure Filtering (ISF) Methods for Untargeted Mass Spectrometry Imaging of Bacterial Metabolites. *J. Am. Soc. Mass Spectrom.* **2024**, *35* (8), 1743–1755. https://doi.org/10.1021/jasms.4c00129.
(21) Janda, M.; Seah, B. K. B.; Jakob, D.; Beckmann, J.; Geier, B.; Liebeke, M. Determination of Abundant Metabolite Matrix Adducts Illuminates the Dark Metabolome of MALDI-Mass Spectrometry Imaging Datasets. *Anal. Chem.* **2021**, *93* (24), 8399–8407. https://doi.org/10.1021/acs.analchem.0c04720.
(22) Andric, S.; Meyer, T.; Rigolet, A.; Prigent-Combaret, C.; Höfte, M.; Balleux, G.; Steels, S.; Hoff, G.; De Mot, R.; McCann, A.; De Pauw, E.; Arias, A. A.; Ongena, M. Lipopeptide Interplay Mediates Molecular Interactions between Soil Bacilli and Pseudomonads. *Microbiol. Spectrum* **2021**, *9* (3), 17. https://doi.org/10.1128/spectrum.02038-21.

CHAPTER FIVE

***

Relational Graph Convolutional Network for Robust Mass Spectrum Classification

Chapter 5: Content

5.1 Preface 167

5.2 Relational Graph Convolutional Network for Robust Mass Spectrum Classification 168

5.2.1 Abstract 168

5.2.2 Introduction 170

5.2.3 Experimental section 173

5.2.4 Method 175

5.2.4.1 General approach 175

5.2.4.2 Graph representation of the spectrum 175

5.2.4.3 Relational Graph Convolutional Network 177

5.2.4.4 Class activation maps for network interpretability 177

5.2.4.5 Code availability 178

5.2.5 Results and discussion 179

5.2.5.1 Robustness toward signal variation 179

5.2.5.2 R-GCN signal interpretability 188

5.2.5.3 Relationship to Chapter 4 199

5.2.6 Conclusion and perspectives 202

5.1 Preface

Chapter 4 introduced graphs of mass differences as an unsupervised representation to group centroid peaks into chemically meaningful families using Δm relations (isotopes, adducts, and homologous series). In this chapter, we apply the same graph representation to a different supervised goal: rather than clustering a dataset into molecular families, we use it for robust MSI spectrum classification and then investigate how the model can be used to interpret class-associated molecular families.

This chapter reproduces our published work: "*Relational Graph Convolutional Network for Robust Mass Spectrum Classification*" written by **Raphaël La Rocca**, Anthony Cioppa, Enrico Ferrarini, Monica Höfte, Marc Van Droogenbroeck, Edwin De Pauw, Gauthier Eppe, and Loïc Quinton, and published in the Journal of the American Society for Mass Spectrometry, in **2025**, 36, 2036–2047 (DOI: 10.1021/jasms.5c00055).

In brief, spectra are modeled as relational graphs where peaks are nodes and Δm relations are typed edges; an R-GCN predicts spectrum labels, and CAM-style attribution highlights peaks and molecular families contributing to each prediction.

The demonstrations in this chapter rely on high-resolution MALDI data (high mass accuracy, e.g., FT-ICR), on which the mass defects and sub-millidalton mass differences exploited by the model are reliably measured. As the chapter notes, the method's dependence on mass precision restricts its direct application to data of comparable resolution; transfer to lower-resolution instruments, to other acquisition modes, and to non-imaging MS therefore remains to be established.

5.2 Relational Graph Convolutional Network for Robust Mass Spectrum Classification

5.2.1 Abstract

Supervised machine learning methods have shown impressive performance in interpreting mass signals and automatically segmenting spatially meaningful regions in Mass Spectrometry Imaging (MSI). Such segmentation generates maps that provide researchers with valuable insights into sample composition and serve as a foundation for downstream statistical analyses. However, these models often require data set-specific preprocessing and do not fully exploit the rich mass features available in high-resolution mass spectrometry (HRMS). Unlike low-resolution mass spectrometry, HRMS reveals additional features such as mass defects and repeated mass differences that carry important chemical information. In this work, we propose a novel deep learning architecture based on a Relational Graph Convolutional Network (R-GCN) that captures and leverages those HRMS mass features. Our model explicitly encodes structural features such as mass defects and known mass differences to represent each spectrum as a graph, enabling the learning of associations between chemically related ion families. To the best of our knowledge, no existing deep learning models for MSI classification incorporate this level of chemically informed mass structure. Most existing methods treat spectra as flat vectors or image-like inputs, thereby ignoring the underlying mass relationships. We evaluate our R-GCN approach against several conventional machine learning and deep learning baselines across diverse MSI data sets, demonstrating its robustness to common signal variations (e.g., mass shift, ion loss). Finally, we integrate Class Activation Mapping (CAM) to enhance model interpretability, enabling the identification of ion families that are relevant to specific biological or spatial regions.

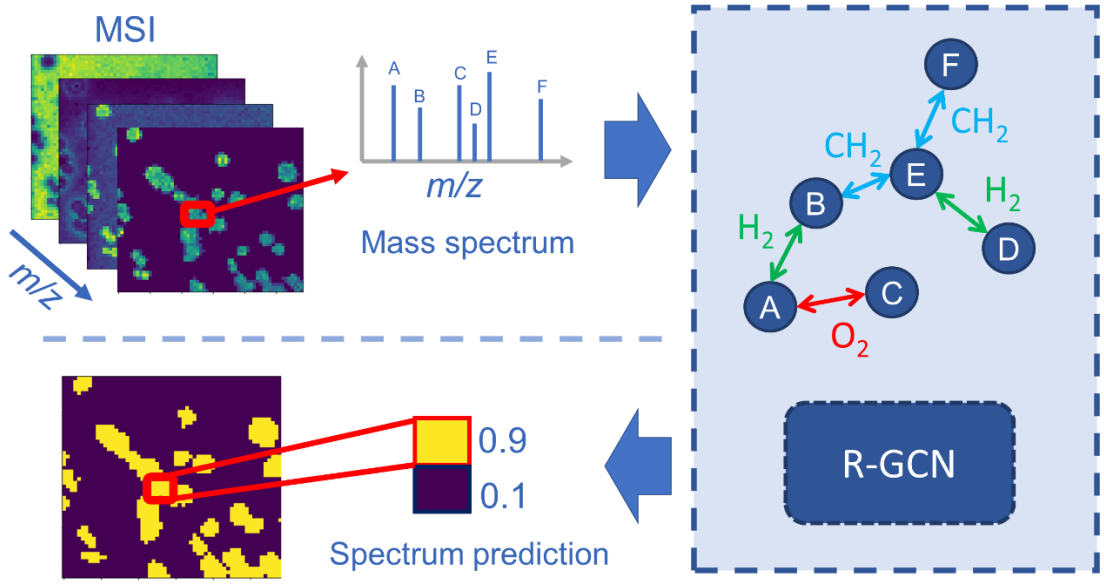


Figure 5.1 *Overview of the relational mass-difference graph pipeline for MSI spectrum classification. For each MSI pixel, the centroid mass spectrum is extracted and converted into a relational graph where nodes represent detected peaks (A–F), and edges encode predefined Δm relations (illustrated here with examples such as H_2, CH_2, and O_2). A Relational Graph Convolutional Network (R-GCN) processes this graph to predict the spectrum's class as probabilities (e.g., 0.9 vs 0.1). Repeating this operation for all pixels yields a spatial prediction/segmentation map over the MSI image.*

5.2.2 Introduction

Mass Spectrometry Imaging (MSI) is widely used to detect biological regions of interest from spatially resolved spectral signatures, notably in clinical and microbiological applications[1–5]. As introduced in Chapter 1, MSI acquisitions yield a mass spectrum at each pixel, and high-resolution mass spectrometry (HRMS) provides high mass accuracy and isotopic signatures that can support molecular formula hypotheses and downstream interpretation[6,7].

The large number of spectra and signals generated by HRMS-MSI requires specialized computational approaches for exploratory analysis and for automated classification[8]. Unsupervised methods can reveal regions with similar spectral signatures, whereas supervised learning is particularly useful when annotated classes are available (e.g., tissue types, conditions), enabling spectrum classification and pixel-wise segmentation[9–13]. In conventional supervised pipelines, spectra are typically transformed into feature vectors by matching peaks (or bins) across spectra within a tolerance and using the corresponding intensities as inputs to standard classifiers (e.g., LDA/SVM/RF) or other learning algorithms[14,15]. When coupled with interpretation tools, such models can highlight class-associated *m/z* values and facilitate biological interpretation[11,13].

In practice, MSI spectra exhibit substantial variability across pixels and samples due to sample preparation, intrinsic molecular heterogeneity, and instrumental effects[16,17]. These variations can manifest as intensity noise, peak dropout, and, importantly, small *m/z* shifts, which directly affect peak matching in HRMS and can degrade both feature quality and classification performance[18–21]. As discussed in Chapter 3, mitigating such effects often requires careful preprocessing (e.g., recalibration, denoising, filtering) guided by expert knowledge and parameter choices[20–22].

To reduce reliance on manual feature engineering, Convolutional Neural Networks (CNNs) have been proposed for spectrum classification with limited preprocessing by learning local patterns from binned spectra[10,23–25]. However, CNNs are not well adapted to HRMS: preserving the benefits of high resolution typically requires very fine binning, which greatly increases input dimensionality and model size, making training and deployment less practical on common GPU hardware. Moreover, binning-based representations do not explicitly encode chemically meaningful relations between peaks that are central to HRMS interpretability.

A complementary HRMS-specific representation is based on mass defect and repetitive mass differences, commonly visualized with Kendrick plots[26,27]. As detailed in Chapter 4, mass patterns can be defined as sets of peaks separated by characteristic Δm values corresponding to isotopes, adducts, fragmentations, and chemical transformations[28,29], and these relations can be encoded as edges in a graph representation (Figure 5.2). Such structured representations naturally group peaks into molecular families that are informative for both discrimination and interpretation.

In this work, we propose a supervised deep learning model based on a Relational Graph Convolutional Network (R-GCN) that represents each spectrum as a relational graph to efficiently model HRMS structure: exact mass, intensity, mass defect, and mass patterns. We evaluate the method against state-of-the-art approaches in two steps: (i) classification of background versus foreground spectra from multiple HRMS-MSI images and (ii) classification of spectra from homeostatic and non-inflammatory Non-Alcoholic Fatty Liver Disease (NAFLD), with an emphasis on robustness to signal variations. Finally, we combine the learned representation with class activation maps (CAM)[30] to highlight class-associated peaks and molecular families and to support interpretation across MSI datasets.

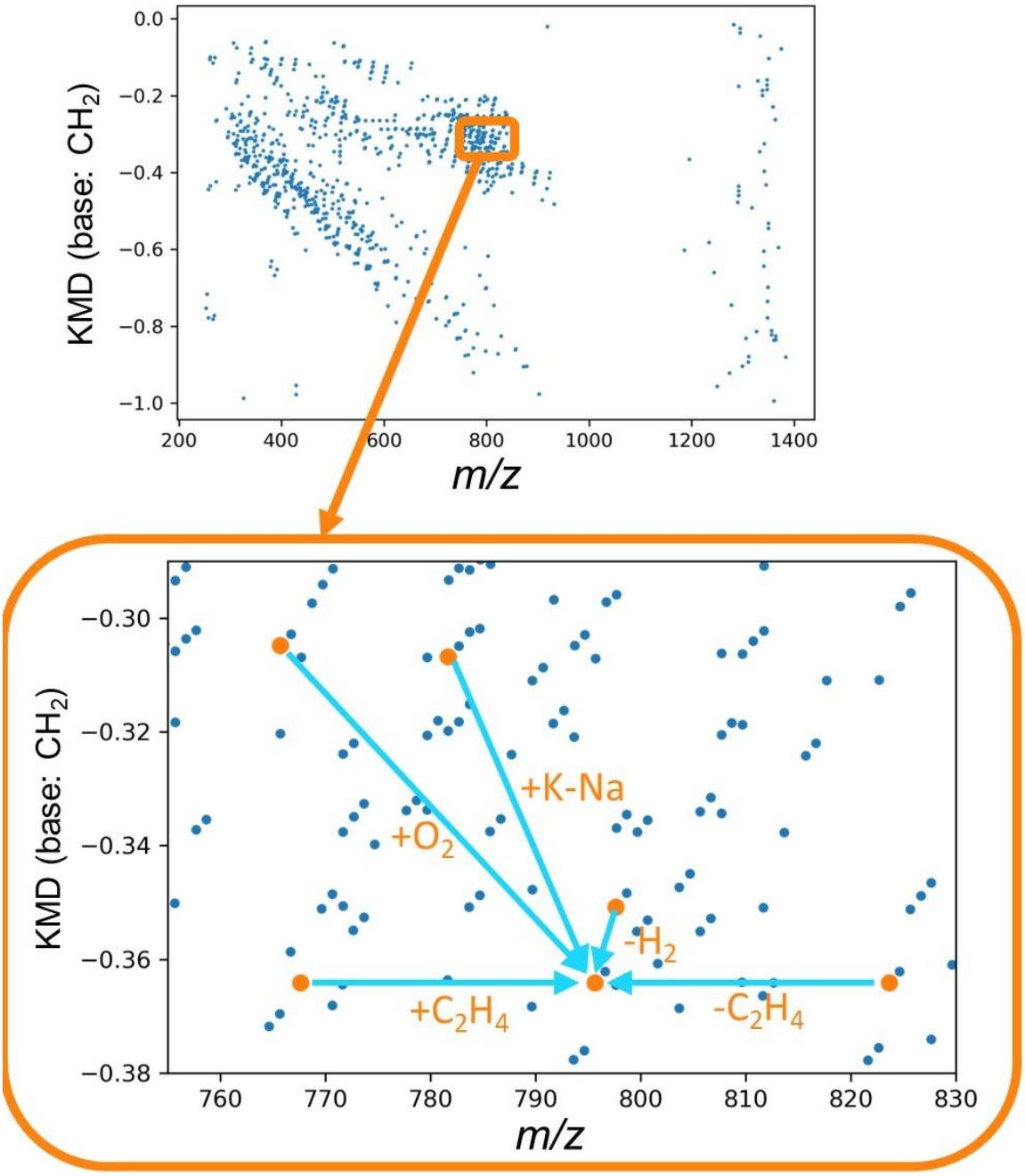


Figure 5.2. *Representation of HRMS mass features by a Kendrick plot in CH_2 base of a MALDI MSI of a human lung biopsy. Blue points represent mass peaks. Light blue arrows represent specific mass differences corresponding to a particular difference in atomic composition.*

5.2.3 Experimental section

In this study, we collect several MALDI MSI data sets from METASPACE (see Supporting Information, S1), a hepatocyte data set including homeostatic state (Control) and non-inflammatory Non-Alcoholic Fatty Liver Disease (NAFLD) hepatocytes from Rappez et al.[5], and we provide an additional data set from our laboratory, related to *Pseudomonas sessilinigenes* CMR12a[31] bacterial colonies grown on agar. We include two additional MALDI-MSI data sets from Burguet et al.[32] These data sets were designed to compare culture conditions that either activate or repress specialized metabolite production *in Streptomyces scabiei*. Sample preparation and spatial annotation are detailed in Supporting Information S6 for all data sets.

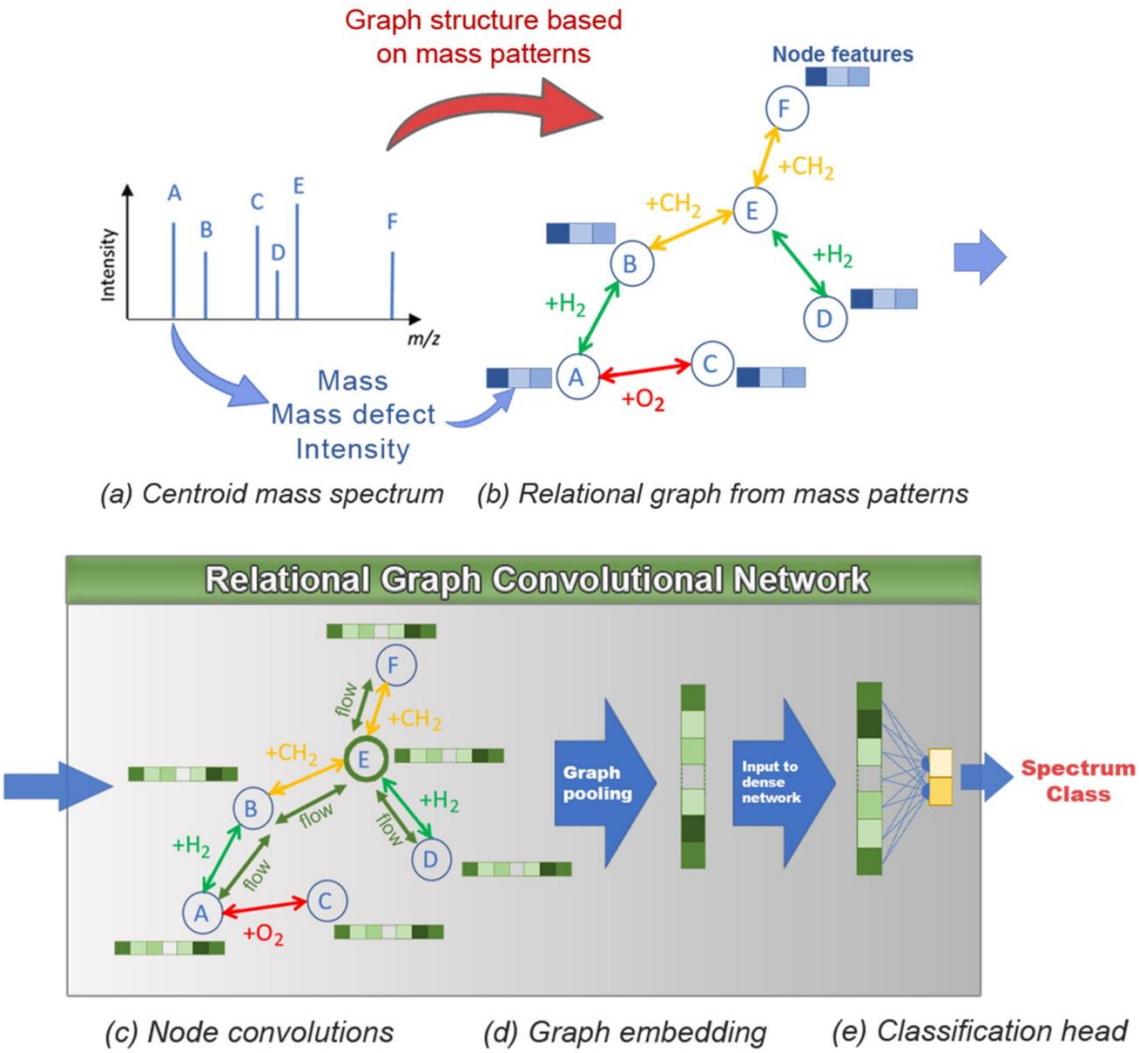


Figure 5.3. *Pipeline of our R-GCN for HRMS classification. a) A centroid mass spectrum is transformed into a graph containing b) the HRMS features of each peak. c) graph convolution layers create a representation for each node, then d) aggregate it into a single vector representation by a pooling layer passed to e) a classification head.*

5.2.4 Method

5.2.4.1 General approach

In HRMS, the mass peak intensity values are not the only information available. Masses, mass defects, and mass differences between peaks contain important information from which a model can learn and extract essential patterns for data interpretation (Figure 5.2). Currently, those features are systematically overlooked by automatic state-of-the-art methods. However, this information is often used for manual analysis and provides a good estimation of the molecular composition. In this work, we develop a new deep learning algorithm for MSI's automatic mass spectrum classification to efficiently incorporate those extra mass features. Crucially, the model operates on a single spectrum at a time: no spatial context, such as the pixel coordinates or the spectra of neighboring pixels, is incorporated into the graph representation or the classification decision. Each pixel is classified independently, based solely on the mass features of its own spectrum. To accomplish this, we select a graph representation for the spectrum described below.

5.2.4.2 Graph representation of the spectrum

As a first step, each spectrum of each pixel (see Figure 5.3a) is converted into a graph. To this end, each centroid in the spectrum is represented as a node in the graph. For each node, three features are associated: the mass of the centroid, its intensity, and the computed mass defect according to IUPAC as follows:

$$Mass\ Defect = Mass - \lceil Mass \rceil \qquad (1)$$

Mass Defect is defined as the difference between the measured Mass (the ion mass in atomic mass units) and its ceiling Mass ($\lceil Mass \rceil$, the smallest integer greater than or equal to the mass). Note that the mass defect and mass values are scaled between -

0.5 and 0.5, and the centroid intensities are log-transformed and then Z-score normalized. Then, two nodes are connected by a bidirectional edge if the absolute difference between their associated mass corresponds, within a certain mass tolerance (e.g., set at ± 0.001 Da), to a predefined set of mass differences. Spatial coordinates of the pixel from which the spectrum was acquired are deliberately excluded from the graph: nodes and edges encode only mass-domain information (*m/z*, mass defect, intensity, and Δm relation type), making the representation fully decoupled from the acquisition geometry.

Note that mass defect carries chemically meaningful information beyond peak identification: oxygen-rich compounds tend to exhibit more negative mass defects, while lipids and other hydrogen-rich species cluster at less negative values, making mass defect a useful proxy for broad elemental composition class and molecular family distinction.

For all experiments, we selected the same set composed of 102 mass differences. Those differences describe the most frequent mass patterns in MALDI-MSI experiments[28](see Supporting Information, S3). For instance, an element in the set may represent the mass difference of a change of adducts, isotopic patterns, the masses of common chemical transformations such as the addition of H_2, CH_2, and O_2, and mass differences from DHB and HCCA. As an extra feature, we provide this label to the bidirectional edge connecting the two nodes. Therefore, our spectrum representation is a bidirectional relational graph (see Figure 5.3b). Different methods for selecting mass differences have been explored (see Supporting Information, S8).

5.2.4.3 Relational Graph Convolutional Network

Graph Convolutional Network (GCN) is a particular network architecture that applies convolution operations on node features in a graph by aggregating the information on adjacent nodes. Performing multiple consecutive convolutions allows for the aggregation of node information further and further away in the graph. Relational-GCN (R-GCN) is a particular type of GCN where the graph edges contain labels. This allows us to model different relation types between the nodes. In this case, the convolution aggregates adjacent node representations connected by similar edge types[33].

We used the PyTorch Geometric library[34] to build our own R-GCN network architecture. Our R-GCN takes as input a graph of mass differences and encodes the graph's node features with two graph convolution layers. The first convolution layer converts the initial three features of each node into eight features, and the second one transforms the eight features of each node into 16 features (see Figure 5.3c). Next, average pooling is applied to all nodes, feature by feature. This pooling operation allows the reduction of all the node's representation into a single vector of 16 features (see Figure 5.3d). This representation is then passed to a single fully connected layer that outputs two values. All layers are activated with the ReLU (rectified linear unit) function except the output layer, which is activated by a Softmax Activation function to produce the predictions for each class (see Figure 5.3e). A more detailed description of the network architecture can be found in Supporting Information S7). The impact of the method's parameters is also explored (see Supporting Information, S8).

5.2.4.4 Class activation maps for network interpretability

Classification activation map (CAM)[30] is a method for interpreting, visualizing, and debugging CNN networks. We propose to adapt CAM to interpret learned mass features of our R-GCN. The CAM operation consists of applying the classification head of the network directly to each node representation before the pooling layer. This gives each

node present in the graph two features corresponding to the contribution of the node to each of the two classes. This allows us to interpret which node (and therefore which peak in the spectrum) is characteristic to a class. We consider a peak to be characteristic of a particular class if its CAM feature, i.e., its contribution to that class, is greater than its contribution to the other class.

5.2.4.5 Code availability

The R-GCN method implementation and the implementations of various classification methods are available on GitHub at https://github.com/LaRoccaRaphael/Massconvnet. Training time and model complexity for each method on the background/foreground classification task are provided in the Supporting Information (S11.1).

	Background/Foreground	Control and non-inflammatory NAFLD human HepaRG hepatocytes
	Balanced Accuracy	*Balanced Accuracy*
R-GCN (this work)	0.874 ± 0.001	0.991 ± 0.002
1D-CNN	0.860 ± 0.002	0.996 ± 0.001
RF	0.871 ± 0.001	0.985 ± 0.001
SVM	0.873 ± 0.0	0.999 ± 0.0
LDA	0.861 ± 0.0	0.999 ± 0.0
XGBoost	0.863 ± 0.0	0.996 ± 0.0

Table 5.1. *Performance comparison between our R-GCN method against different methods using balanced accuracies on the BG/FG classification task of the METASPACE MSI datasets and the classification of control and non-inflammatory NAFLD human HepaRG hepatocytes.*

5.2.5 Results and discussion

As explained in the introduction, the current work proposes a new method for mass spectrum classification that takes advantage of the features provided by HRMS (exact masses, mass defects, mass patterns). We characterize our new R-GCN method for mass spectrum classification into two sections: (1) Its robustness toward signal variation and (2) how it can be used for *m/z* signal interpretation (R-GCN signal interpretability). To support our experiments, we compare our method against different spectrum classification methods: Random Forests (RF), Linear discriminant analysis (LDA), Support vector machine (SVM), eXtreme Gradient Boosting (XGBoost), and 1D Convolutional Neural Networks (1D-CNNs). Each preprocessing method is described in Supporting Information S9.

5.2.5.1 Robustness toward signal variation

Experimental setup

For this first experiment, we evaluate the method's robustness toward signal variation on two different classification tasks: classification of foreground (FG) and background (BG) spectra from 4 different MSI and classification of human HepaRG hepatocytes cells into control and non-inflammatory NAFLD states. Each spectrum corresponds to a single MSI pixel; no aggregation or averaging is performed. For this purpose, we create:

- A training set and a validation set from the four METASPACE MSI datasets, gathered by randomly sampling individual pixel spectra: 64% and 16% of all available pixels, respectively (6,400- and 1,600-pixel spectra). The remaining 20% of pixel spectra (2,000) are kept as a variation-free test set.
- A training set and a validation set from 2 images of each class in the hepatocyte dataset, gathered by randomly sampling individual pixel spectra: 80% and 20%, respectively (4,000- and 1,000-pixel spectra). The remaining four images (2 per class) are kept as a variation-free test set (5,000-pixel spectra), ensuring that the test pixels originate from entirely unseen MSI acquisitions.

From each data set variation-free test set, we generate four variation test sets by simulating diverse signal variations on the data, which are constant and random mass shifts, a decrease in ion detection intensity, and a mix of signal classes (see below). All six methods (RF, SVM, LDA, XGBoost, 1D-CNN, and R-GCN) are trained on the same training set without any mass variation, except for signal mixing variation. Then, the models are evaluated on the four variation test sets to study their behavior when faced with these variations. For the evaluation metric, we choose balanced accuracy since the background/foreground data sets are highly imbalanced toward the background (see Supporting Information S11.1). Balanced accuracy is defined by

$$balanced\ accuracy = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \quad (2)$$

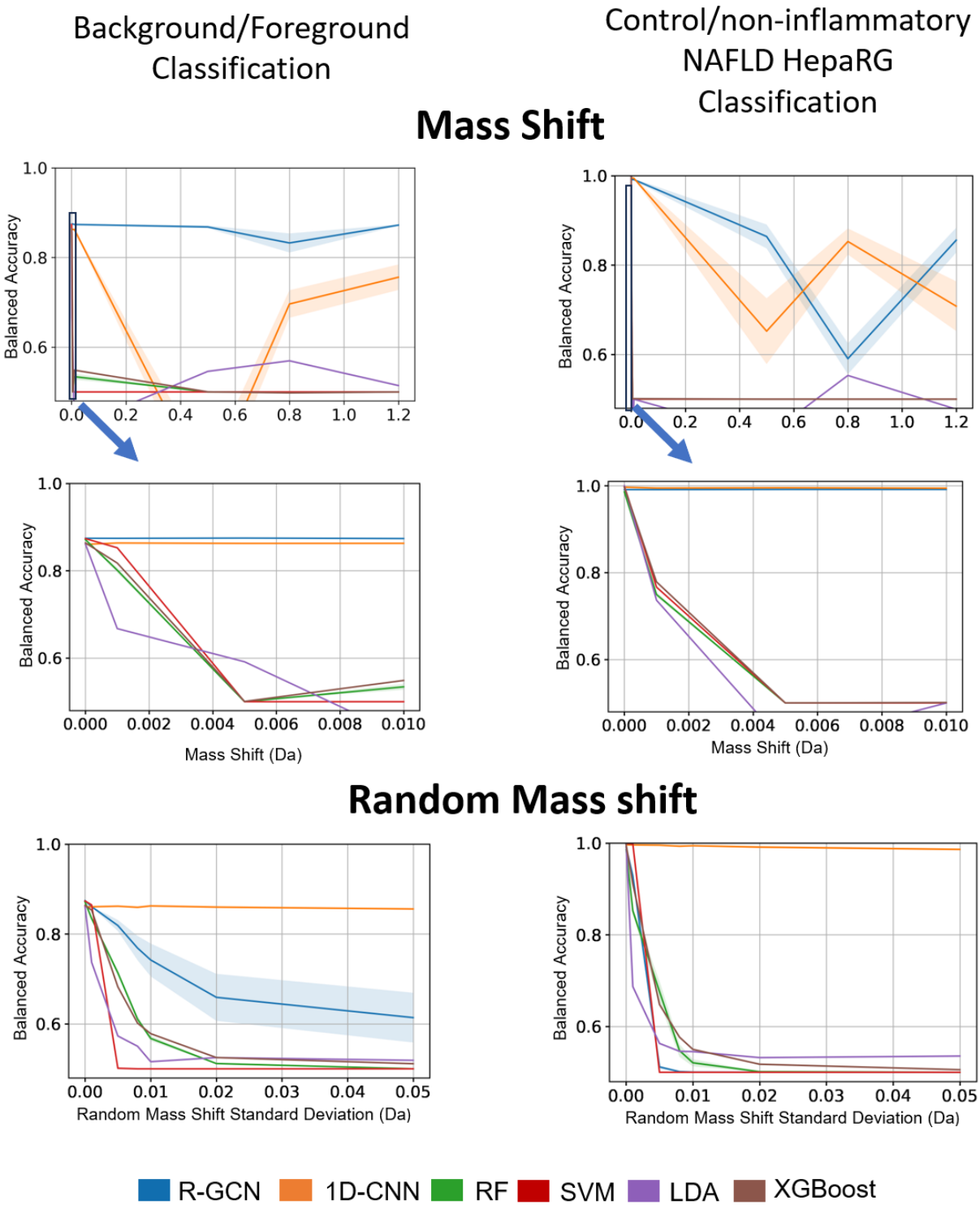
where TP − true positive, TN − true negative, FP – false positive, FN − false negative. The four signal variations simulated in the test sets are the following.

- Test 1. A *mass shift* was simulated by adding a constant value to every signal in each spectrum.
- Test 2. A *random mass shift* was simulated by adding a random value drawn from a normal distribution N(0,σ) to every signal in each spectrum. An increase in the standard deviation of this distribution corresponds to a decrease in mass precision.
- Test 3. A *decrease in ion detection sensitivity* was imposed by removing a proportion of the lowest intensity peaks in each spectrum.
- Test 4. We simulate signal mixing by modifying the test and train sets. For each set, we randomly take 50% of the spectra and add a spectrum from the other class to each spectrum. The added spectrum is multiplied by a value between 0 and 1, where a value of 1 indicates an equal mix of the two class signals.

Note that each of the signal variations described above often co-occurs in real cases; simulation allows us to isolate them, which is essential to study their impact on the

various classification methods. As a benchmark, we evaluate all methods on the original test set without any signal variation. To account for the reproducibility of models with random weight initialization, we train 15 models for each method by varying the order in which the training samples are presented to the model during the training phase (called random seeds in the following). We provide the mean and standard deviation of the balanced accuracy across those 15 repetitions. Since RF, R-GCN, and 1D-CNN have random initialization and the order in which the training samples are presented matters in their optimization, the standard deviation may be greater than 0. As can be seen, all models achieve similar performances (see Table 5.1).

The differences in input representation across methods are the primary driver of the robustness differences observed below, and are worth making explicit before interpreting the results. RF, SVM, LDA, and XGBoost receive a fixed-length vector of 2,000 intensity values matched against a reference peak list built from the training set mean spectrum; they have no access to exact *m/z* values, mass defects, or mass difference relationships. The 1D-CNN receives a binned spectrum at 0.1 Da resolution and therefore retains only a coarse positional encoding of *m/z*, with no explicit mass defect or mass difference information. It operates directly on the 2,000 most intense centroid peaks of each individual spectrum. The R-GCN is the only method that explicitly encodes exact *m/z* values, mass defects, and inter-peak mass difference relationships as typed graph edges, operating also on the 2,000 most intense centroid peaks of each individual spectrum. No method incorporates spatial information.


Background/Foreground Classification
Control/non-inflammatory NAFLD HepaRG Classification
Mass Shift
Balanced Accuracy
Mass Shift (Da)
Random Mass shift
Random Mass Shift Standard Deviation (Da)
R-GCN
1D-CNN
RF
SVM
LDA
XGBoost

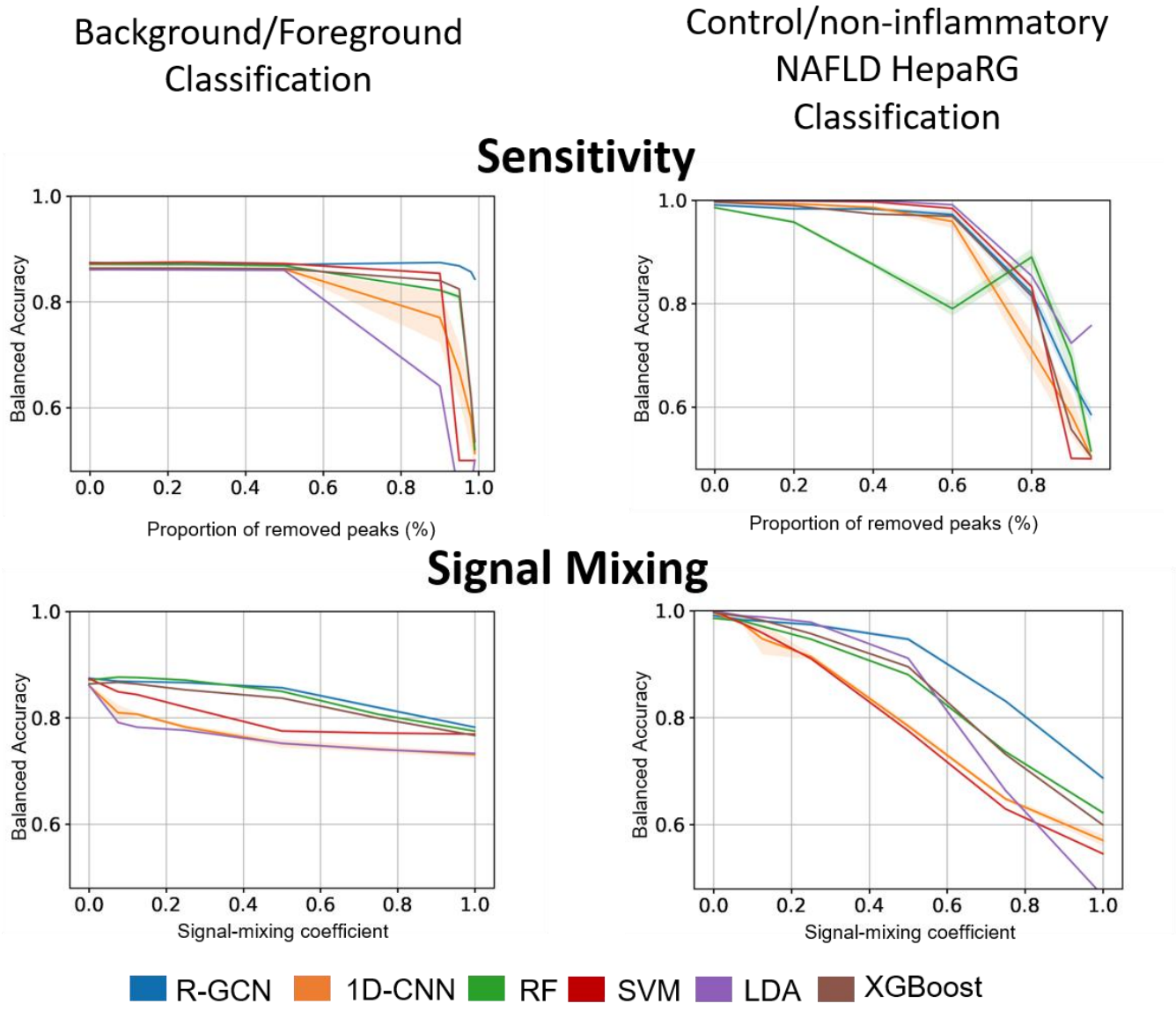


***Figure 5.4**. Performance of the binary classification Background/Foreground (BG/FG) and control/non-inflammatory NAFLD human HepaRG hepatocytes for RF, SVM, LDA, XGBoost, 1D-CNN, and R-GCN on four simulated signal variations: Mass shift, Random mass shift, Decrease in ion sensitivity (Lowest peak removal), Signal mixing. The error bars correspond to the standard deviation of the prediction evaluation for 15 models trained with random seeds.*

Classification results with mass variations

The performances over different signal variation ranges (*x*-axis of the plots) for each type of variation are shown in Figure 5.4. The models are trained on the same training set without signal variations for all six methods. To account for the variability that can arise during training, we train 15 models for each algorithm by shuffling each time the order of the samples in the training phase and the network initialization parameters when relevant. As explained above in the benchmark, only RF, R-GCN, and 1D-CNN can exhibit nonzero standard deviations.

Our R-GCN method shows excellent robustness to mass shift variation compared to other methods (Figure 5.4). Note that simulations with exaggerated shifts (mass shift >0.01 Da) are designed to examine the model's reliance on mass features and are not used to demonstrate typical instrument mass deviations. Accordingly, we also present mass shift values on a more realistic scale in the second row of Figure 5.4.

In fact, it outperforms the other methods over the entire mass shift range except the 1D-CNN on the hepatocyte data set. The R-GCN and 1D-CNN's performances vary much more with mass shift compared to the other methods. For the 1D-CNN, this nonmonotonic behavior is primarily due to the fixed binning strategy used in the preprocessing of input spectra. When mass shifts cause ions to misalign with their expected bins, the input representation leads to a drop in classification accuracy. Interestingly, as the shift approaches 1 Da, performance begins to recover. This is because the shifted ion starts to align with the first isotopic peak (all ions are singly charged due to the MALDI process). This alignment effect partly compensates for the misalignment caused by the shift, resulting in the performance rebound observed in Figure 5.4 after a shift of 0.8 Da.

In contrast, the RF, SVM, LDA, and XGBoost models fail completely when they reach a mass shift of 0.001 Da. In fact, those methods are not robust to mass shift because their input representation is built by matching peaks to a fixed reference list within a tolerance window and using the resulting intensities as features, discarding the actual *m/z* values entirely. As a result, any shift that causes a peak to fall outside its expected matching window directly corrupts the feature vector, with no mechanism to recover the correct signal. In contrast, the R-GCN and 1D-CNN retain explicit mass information in their input representation, as node features and bin positions, respectively, allowing them to remain partially functional even when peaks are displaced.

For the R-GCN, when we shift all masses by a constant value, the graph structure remains unchanged, and only the node features (*m*/*z* and mass defect) are impacted. Indeed, extremely large changes in the values of these features (e.g., 0.8 Da) impact the model performance with a bigger impact on the hepatocyte's data set, which means that mass defect is an important feature for spectra classification in this data set. This perfectly shows the robustness of our model toward this type of variation in the signal; moreover, the robustness toward implausible mass shifts (0.8 Da) demonstrates how the model depends on mass defect values for classification.

Notably, both the R-GCN and 1D-CNN exhibit a partial performance recovery as the mass shift approaches 1 Da. As discussed above, this can be partly attributed to a cyclical realignment effect, where shifted ions approach mass defect or bin values of chemically related species. For the R-GCN specifically, however, the precise origin of this recovery is difficult to attribute unambiguously, as the model simultaneously combines graph topology, exact mass, and mass defect information whose relative contributions vary depending on the dataset.

A random mass shift, which corresponds to a decrease in precision, has a different impact on the models, as seen in Figure 5.4. Here, the 1D-CNN is insensitive to random shifts at this scale, as previously discussed. In the case of our method, the graph

structure is subjected to changes, and, as a result, the performance drops slowly in the BG/FG data set and more rapidly in the hepatocyte data set. The changes are mainly caused by the graph's edge removal as the distance between the mass centroids becomes larger than the method's predefined mass tolerance. In fact, our R-GCN is sensitive to a change in the mass distance between the centroids (see Supporting Information, S8, Table S3), which means that the mass precision of the MSI in the train set restricts its application to MSI with related mass precision. This problem can be avoided by correctly setting the mass tolerance according to the original MSI precision of the test set.

For the signal variation of a decrease in ion detection sensitivity, the different methods demonstrate similar performances for 0−50% removal, with lower performances for RF and LDA. For higher proportions, our R-GCN consistently outperforms the other two on the BG/FG data set. However, the differences in performance between data sets are visible. Those differences can be attributed to the underlying complexity of the data set. The background/foreground data set includes MSI from four diverse biological sources (e.g., tissue, bacteria), leading to stronger, more generalizable class distinctions. Background spectra often contain abundant matrix-related peaks (e.g., HCCA), whereas foreground spectra feature biologically relevant ions such as endogenous lipids and metabolites. These signals remain even after significant peak removal, allowing the R-GCN to preserve classification accuracy by leveraging residual graph structure.

In contrast, the hepatocyte data set involves more subtle metabolic differences between two similar cell states. Most discriminative peaks belong to chemically similar lipid species that are easily lost under high peak dropout, which weakens graph connectivity and model performance. This explains why the R-GCN shows greater resilience in the BG/FG task but converges toward other methods in the hepatocyte setting.

Signal mixing is a common phenomenon in MSI, as mass spectra often span over two different biological regions due to an inadequate spatial resolution. Therefore, the resulting mass spectra contain mass information from the two regions. The lower performance of the 1D-CNN demonstrates the importance of maintaining high spectral resolution when classifying mass spectra with mixed biological origins (Figure 5.4). The 1D-CNN's coarser resolution, noted above, greatly impacts model performance here. The 1D-CNN's bin size (0.1 Da) compared to the mass tolerance of the other methods (0.001 Da) greatly impacts the model performance. The R-GCN consistently outperforms the other methods for high-signal mixing, especially in the hepatocyte data set, where the two classes have similar signals. Attempting to increase the number of bins in the 1D-CNN to match the resolution used by the R-GCN (e.g., 0.001 Da tolerance) would dramatically increase the CNN's complexity, from approximately 4.8 million to 480 million parameters. This scale of dimensionality imposes significant computational and memory demands and substantially increases the risk of overfitting, especially given the relatively small training sizes typical of MSI data sets. In contrast, the R-GCN architecture is far more scalable and robust. It requires only 15,746 trainable parameters, and its complexity remains constant regardless of the number of bins or peaks. Instead, it depends on the structure of the graph defined by mass differences, making it inherently better suited for HRMS data where rich mass patterns and high resolution are key to molecular interpretation.

5.2.5.2 R-GCN signal interpretability

As shown in the previous sections, R-GCN exhibits improved robustness to variations in the signal and to predictions on unseen data sets with distinct signals (see Supporting Information, S10, Figure S3). These improvements can be attributed to better representation of HRMS features, such as mass patterns, masses, and mass defects. These new features allow the method to use mass information more flexibly than 1D-CNN. Building on the R-GCN's ability to capture chemically meaningful mass relationships, we now explore whether its learned feature representations can also aid in interpreting MSI data. Specifically, we investigate whether R-GCN can help identify ions of interest from a list of *m*/*z* values by highlighting those that the model associates with spatial regions of interest in various case studies. Our approach consists of training an R-GCN model for binary classification of MSI spectra and then applying the trained model to a representative spectrum from each dataset. Rather than classifying spectra, we repurpose the model to assign class relevance scores to individual ions, thereby identifying which *m*/*z* values contribute most to distinguishing between the two classes. To achieve this, we integrate a Class Activation Map (CAM) strategy into the R-GCN architecture and apply it across multiple datasets to extract *m*/*z* values associated with different spatial regions or experimental conditions, such as foreground (FG) vs background (BG), or control vs treated samples. The resulting CAM scores allow us to refine the mean spectrum of each MSI dataset by filtering for class-relevant ions, highlighting signals corresponding to molecules produced under specific biological conditions. This not only improves interpretability but also simplifies downstream analysis by focusing on biologically meaningful signals.

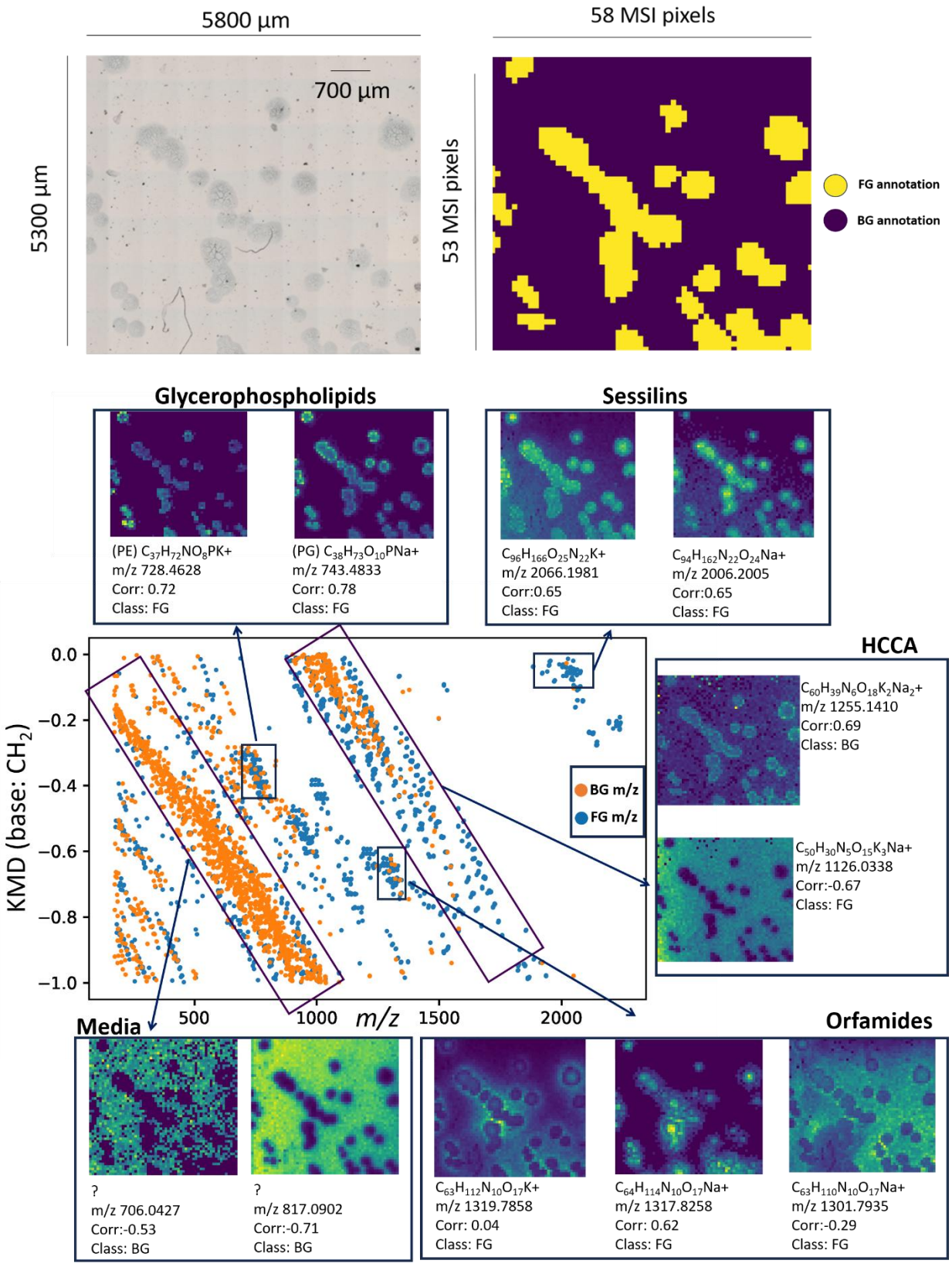

5800 µm
58 MSI pixels
5300 µm
700 µm
53 MSI pixels
FG annotation
BG annotation
Glycerophospholipids
(PE) $C_{37}H_{72}NO_8PK$+
m/z 728.4628
Corr: 0.72
Class: FG
(PG) $C_{38}H_{73}O_{10}PNa$+
m/z 743.4833
Corr: 0.78
Class: FG
Sessilins
$C_{96}H_{166}O_{25}N_{22}K$+
m/z 2066.1981
Corr:0.65
Class: FG
$C_{94}H_{162}N_{22}O_{24}Na$+
m/z 2006.2005
Corr:0.65
Class: FG
HCCA
$C_{60}H_{39}N_6O_{18}K_2Na_2$+
m/z 1255.1410
Corr:0.69
Class: BG
$C_{50}H_{30}N_5O_{15}K_3Na$+
m/z 1126.0338
Corr:-0.67
Class: FG
KMD (base: CH_2)
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
500
1000
m/z
1500
2000
BG m/z
FG m/z
Media
?
m/z 706.0427
Corr:-0.53
Class: BG
?
m/z 817.0902
Corr:-0.71
Class: BG
Orfamides
$C_{63}H_{112}N_{10}O_{17}K$+
m/z 1319.7858
Corr: 0.04
Class: FG
$C_{64}H_{114}N_{10}O_{17}Na$+
m/z 1317.8258
Corr: 0.62
Class: FG
$C_{63}H_{110}N_{10}O_{17}Na$+
m/z 1301.7935
Corr:-0.29
Class: FG

Figure 5.5. *On the top, bright-field microscope image of the colonies (left) and the corresponding annotation of the BG and FG in the MSI (right). On the bottom, Kendrick plot in CH2 base of the most specific signals selected by CAM for each class FG (orange) and BG (dark blue) of the Pseudomonas colonies MSI. Different molecular families are highlighted by rectangle HCCA, glycerophospholipids, and lipopeptides such as orfamides and sessilins. Around the Kendrick plot, we show different ion images as an illustration of the different molecular families.*

Experimental setup

In contrast to the experiments in Section 5.2.5.1, here we modify the R-GCN input by setting all node intensities to 1 during both training and inference. Since all 2,000 most intense centroid peaks of each spectrum are always retained as nodes regardless of their intensity, fixing node intensities to 1 does not change which peaks are represented; all are included as before. Rather, it removes intensity as a discriminative feature, so that the model relies exclusively on exact *m/z* values, mass defects, and graph topology defined by mass difference relationships. The resulting representation is best understood as an unweighted structural encoding of the spectrum: classification is driven entirely by which peaks are present and how they are connected through chemically meaningful mass differences, rather than by how abundant those peaks are. This design isolates the contribution of mass-structural relationships to classification, making the subsequent CAM scores interpretable as reflecting chemical family membership rather than intensity dominance. To improve robustness to small deviations in *m*/*z* values, the tolerance for graph generation is set to 0.003 Da across all data sets. For each data set, a representative spectrum is constructed by computing the mean spectrum across all

individual spectra of the dataset. Once the model is trained on each dataset, we apply the CAM strategy to the representative spectrum and its associated relational graph. The CAM produces two activation values per node (i.e., per *m*/*z* value), corresponding to the contribution of each ion to the two respective classes. This allows us to identify *m*/*z* values that are most characteristic of each class and effectively classify or associate ions with specific biological or experimental conditions (see Methods for details). We

evaluate this interpretability strategy across three distinct case studies, each representing a different biological context and classification task.

- *Case Study 1: Spatially Resolved Ion Detection in P. sessilinigenes Colonies.*
  The first case study focuses on identifying ions expressed in distinct spatial regions within a single MSI data set. We apply this to a newly generated MSI of *P. sessilinigenes* colonies grown on agar (fifth data set in Supporting Information S6, Figure S1). An R-GCN model is trained to classify spectra from the colony regions (foreground) versus the surrounding medium (background). This task illustrates how the model can differentiate microbial metabolic signals from those of the media (Figure 5.5 left part).
- *Case Study 2: Natural Product Discovery in Streptomyces scabiei.*
  The second case study explores the use of R-GCN for natural product discovery, using two MALDIMSI data sets from Burguet et al[32]. These data sets were designed to compare culture conditions that either activate or repress specialized metabolite production in Streptomyces scabiei:
  - *Data set 1: Cellobiose-Induced Pathogenicity.* Here, we compare spectra from *S. scabiei* cultures exposed to cellobiose (which induces virulence-related metabolite production) against negative controls treated with water or sugar (Supporting Information S6). An R-GCN model is trained to classify spectra from the induced versus control conditions.
  - *Data set 2: Iron-Regulated Siderophore Production.* In this data set, *S. scabiei* is grown under iron-depleted versus iron-supplemented conditions to study siderophore biosynthesis. The R-GCN model is again trained to distinguish between spectra from the iron-starved (experimental) and iron-rich (control) cultures.

- *Case Study 3: Lipid Remodeling in a Human Hepatocyte NAFLD Model.*
  The third case study uses a data set from the SpaceM study (Rappez et al.[5]) (Supporting Information S6), also evaluated in earlier sections of this paper. It involves an MSI of human hepatocytes (dHepaRG) under homeostatic conditions (control) and under stimulation with oleic and palmitic fatty acids (+FA) to model non-inflammatory NAFLD. SpaceM analysis previously revealed metabolic remodeling in this model, with an increase in triglyceride and diglyceride expression in NAFLD hepatocytes and glycerophospholipid enrichment in controls. We train an R-GCN model to classify spectra from + FA (NAFLD) versus control hepatocytes, enabling interpretability in the context of lipid dysregulation.

Finally, we annotate selected *m*/*z* values in each data set to evaluate biological relevance:

In the *P. sessilinigenes* dataset, the *m*/*z* values from the representative spectrum are matched against a curated list of exact masses corresponding to known molecular formulas, using a tolerance of ±3 ppm (see Supporting Information, S4 and S5). Those molecular formulas correspond to HCCA ion masses and various lipopeptides (orfamides, sessilins) and phospholipids (PE, PC) known to be produced by this *P. sessilinigenes* strain[35]. Examples of annotated *m*/*z* values are shown in Figure 5.5, along with ion images and bacterial colony regions. To further support these assignments, Spearman correlations were computed between ion images and annotation masks to assess spatial colocalization.

In the *S. scabiei* cultures, we reference known ion identities from Burguet et al[32]. For the cellobiose-induced data set, Thaxtomin A and B are expected in the experimental condition. For the iron-repressed data set, siderophores such as Desferrioxamine B, Desferrioxamine E, and Scabichelin are reported (see Supporting Information S4).

In the Hepatocyte (SpaceM) dataset, annotations are retrieved from the original METASPACE results (as described in the SpaceM publication). The key lipid classes are reported in our data and include diglycerides, triglycerides, and glycerophospholipids, which reflect the reported metabolic remodeling in the non-inflammatory NAFLD model (see Supporting Information S4).

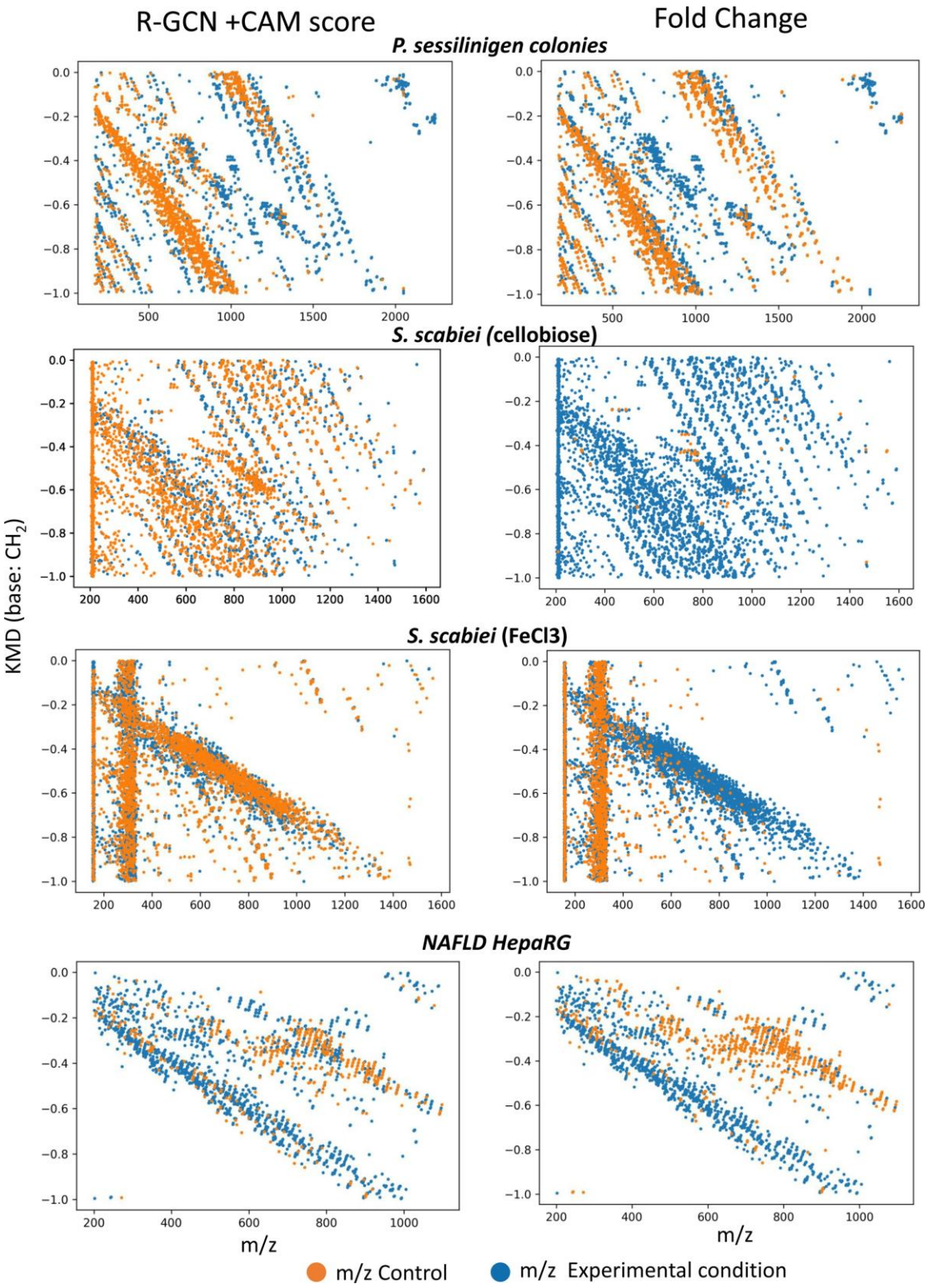


Figure 5.6. *This figure presents the comparison of m/z values extracted across four key datasets: Pseudomonas, Streptomyces scabiei grown on cellobiose, Streptomyces scabiei treated with FeCl3, and human HepaRG hepatocytes in a NAFLD context. The data*

were analyzed using two distinct methods: the CAM score and fold change calculations to find discriminant m/z across various conditions. In the Kendrick plot, m/z values are categorized based on their association with experimental conditions. Values related to the control condition are marked in orange, while those associated with experimental conditions are represented in blue. This visual representation allows for swift identification of m/z values linked to specific biological conditions.

Signal filtering and identification in *P. sessilinigenes* Colonies

In this section, we examine how our method filters *m*/*z* values in the *P. sessilinigenes* data set using the CAM-derived scores. As shown in Figure 5.5, the Kendrick plot highlights distinct molecular families associated with each class**.**

Background-specific (BG) ions cluster tightly in a specific Kendrick region (labeled as the "Media" below 1200 *m*/*z* in Figure 5.5). While these ions remain unannotated, their high oxygen content (suggested by mass defect) probably indicates they originate from agar components, particularly dextrose. Spatially, these ions are most abundant around, but not within, the bacterial colonies. In contrast, foreground-specific (FG) ions span several molecular families. These include: HCCA- derived ions from the MALDI matrix, phospholipids, typically found within colony interiors, and Lipopeptides, such as orfamides and sessilins, known to be produced by *Pseudomonas* strains. Interestingly, lipopeptide ions exhibit diverse spatial behaviors: some (e.g., orfamides) diffuse outward from the colonies, while others (e.g., sessilins) appear more localized within the colony area. Despite this spatial heterogeneity, all lipopeptide ions are still classified as FG-specific by the model. This highlights an important strength of our approach: it can associate ions with biological origin even when their spatial distribution is only weakly correlated with the class labels (e.g., Spearman correlation near zero or even negative).

Similarly, HCCA ions present across the entire MSI surface show varied localization depending on their adduct type and molecular composition. Some are more abundant in the colonies, while others dominate the background.

Overall, the R-GCN + CAM approach enables the filtering of background ions (e.g., media-related) and the classification of biologically relevant molecular families. Importantly, the method can identify class-associated ions even when their spatial correlation is weak, reflecting a better understanding of molecular similarity rather than relying solely on spatial colocalization.

Analysis of *m/z* classification

To assess the consistency of *m*/*z* classification across different model initializations, we trained five R-GCN models with distinct random seeds and computed pairwise accuracies between their predicted class labels for each *m*/*z* value. For the *Pseudomonas* dataset, the models showed strong agreement, with an average accuracy of 0.98 (see Supporting Information S11.3, Table S5). Similar levels of consistency were observed in the other data sets. These results confirm that our ion-level predictions are robust to random initialization. Based on this consistency, we selected a representative model per data set for downstream interpretability analyses.

The *m*/*z* values classified by R-GCN + CAM are visualized in Figure 5.6 (left column) using Kendrick plots. For comparison, we also applied a fold-change (FC) analysis, where FC of an *m*/*z* is defined as the ratio between the mean intensity of that *m*/*z* in the experimental condition and the control. An *m*/*z* is considered specific to the experimental condition if FC > 1, or to the control if FC < 1. These results are shown in Figure 5.6 (right column). We compare the outputs of R-GCN + CAM and FC across all three data sets (Supporting Information S4).

- *P. sessilinigenes Data Set.* In the Kendrick plot, notable differences are observed between the two methods. The orfamide region, a known lipopeptide family produced by *P. sessilinigenes*, is correctly associated with the foreground (colony) class by R-GCN, whereas FC misclassifies many of these ions (orange points in the orfamide region).

- *S. scabiei Data Sets (Cellobiose and Iron-Depleted Conditions).* In both data sets, R-GCN + CAM identifies a greater number of ions associated with the control condition than FC does, as indicated by the higher number of orange points. While it may appear that CAM removes too many experimental-condition ions (e.g., fewer blue points), it successfully retains key *m*/*z* values corresponding to thaxtomin A/B and siderophores (desferrioxamines, scabichelin) identified in the original study (see Supporting Information S4).

- *Hepatocyte (NAFLD) Data Set.* In this dataset, R-GCN and FC produce broadly similar results. Both methods associate triglycerides (TG, 800−1000 *m*/*z*, KMD −0.2 to 0) and diglycerides (DG, 550−700 *m*/*z*, KMD −0.2 to 0) with the experimental condition (+FA), consistent with findings from the original SpaceM paper. One difference is that R-GCN classifies a broader range of DG (19/20 DG) ions as experimental-specific, whereas FC shows more variability across the same region (13/20 DG).

Summary and Implications for MSI Interpretation. Overall, these results demonstrate that R-GCN + CAM not only provides interpretable and chemically informed ion-level classification but also effectively removes media-associated ions and misclassified ions that are often retained by fold-change-based methods. This highlights the value of incorporating structural mass relationships into MSI interpretation beyond traditional statistical comparisons.

Fold-change analysis and the R-GCN + CAM approach rely on fundamentally different principles. Fold change considers only intensity differences between conditions, without regard to mass structure. In contrast, our R-GCN model is configured in this experiment to ignore intensities by setting all node values to rely solely on mass-based features such as mass defect and mass differences. This structural perspective enables

the model to identify entire molecular families that share characteristic mass patterns, even if not all members are overexpressed in the experimental condition.
For example, in the *P. sessilinigenes* data set, the model correctly classifies the orfamide lipopeptide family as foreground-specific, even though some orfamide ions are diffusely distributed and weakly spatially correlated with the colony region. Similarly, in the hepatocyte data set, the model associates most diglycerides (DG) and triglycerides (TG) with the NAFLD condition, despite not all members of these lipid families being overexpressed individually.

One of the advantages of the R-GCN + CAM strategy lies in its ability to cluster and identify ions from the same molecular family, even when their spatial behavior differs. This stems from the graph representation: during graph convolutions, node features are propagated and aggregated from neighboring nodes based on their connectivity and edge types. As a result, CAM scores reflect not only a node's own characteristics but also those of chemically related ions nearby in the graph. This makes it possible for a weakly class-specific ion to be correctly classified when embedded within a chemically coherent neighborhood.

While these preliminary results suggest that R-GCN + CAM can aid manual interpretation and discovery of ions of interest, we acknowledge that the clear advantage of our approach was most apparent in the *P. sessilinigenes* data set, where the model uniquely highlighted orfamides. In the other two case studies, both R-GCN and FC identified the major ions described by the original authors, showing comparable performance on known signals. However, the R-GCN occasionally highlights weakly spatially correlated ions not previously reported molecules that could merit further investigation, but for which we currently lack ground truth validation.

As noted above, intensity was set to 1 only to separate the R-GCN's mass-based contribution from the signal already captured by fold change. Reintroducing it alongside graph topology is a natural next step to test whether differentially expressed molecular families can still be recovered while also weighting ions by abundance, though this would require careful choice of the representative spectrum on which CAM is computed, since intensity-dependent scores are sensitive to it and the mean spectrum may not reflect condition-specific expression.

5.2.5.3 Relationship to Chapter 4

The R-GCN's core component is a graph where edges denote chemically significant mass differences, similar to the unsupervised clustering method discussed in Chapter 4, but the two were not built for the same purpose, and making this explicit clarifies what each contributes.

The Chapter 4 method is a dataset-level data-reduction tool: it groups the peaks of an MSI experiment into connected components that ideally correspond to a single molecular family, so that the chemistry of the sample can be inspected through a few representative ion images. Reaching that representation requires the full preprocessing pipeline, average spectrum, peak detection and integration, and one ion image per peak, because edges are filtered by ion-image colocalization. The R-GCN’s preprocessing occupies the opposite end of this spectrum. It is a supervised classifier that takes as input a single centroid spectrum directly, without any of that preprocessing, and it uses the graph not to compress a dataset but to gain the robustness to mass shift, peak dropout, and signal mixing demonstrated in Section 5.2.5.1. The two approaches are therefore complementary rather than competing: one reorganizes a dataset into molecular families for exploration, the other classifies individual spectra robustly and at scale.

This distinction carries over to the way each method associates *m/z* values with molecular families. In Chapter 4, the grouping is intrinsic to the graph and class-agnostic: a cluster is constructed to hold one family. In the R-GCN the grouping is a by-product of a supervised decision; CAM propagates a spectrum-level class assignment down to the nodes (Section 5.2.5.2), so the resulting set of class-relevant ions is not one family but the ensemble of families that characterize a class, several of which Chapter 4 would have separated into distinct clusters. The graph contributes here not by isolating families but by keeping them intact across the class boundary: because convolution lets a node inherit the behavior of its chemically connected neighbors, the members of a family tend to receive a coherent assignment and to be attributed, as a whole, to the region where the family is most expressed, even when some members are only weakly or negatively correlated spatially. This is the family-level coherence that fold change, scoring each *m/z* independently, cannot reproduce.

That coherence is, however, a direct consequence of the fact that, unlike Chapter 4, the R-GCN does not use colocalization. Adding a colocalization constraint to the R-GCN graph would remove precisely the edges through which a spatially diffuse family member borrows the class assignment of its relatives, suppressing exactly the ions the present approach is able to recover. Importing spatial filtering would thus trade family-level recovery for tighter spatial specificity; which behavior is preferable is application-dependent, and we regard it as a question to be studied rather than a settled choice.

It is also not straightforward to add spatial information to the R-GCN in the way Chapter 4 does, and the obstacle is structural rather than practical. In Chapter 4, colocalization is an edge condition, a similarity between two ion images, that is only definable once the dataset has been reduced to a peak list with one image per peak. The R-GCN builds its graph from a single spectrum, where ion images do not exist at the node level; the same filter, therefore, cannot be attached without abandoning the per-pixel, preprocessing-free formulation that gives the classifier its robustness and scalability.

Two routes could nonetheless give the R-GCN access to spatial information, and they motivate the perspectives below. The first keeps the per-pixel classifier but augments it with spatial context—adding spatially derived node features, or coupling the GNN to a component that reasons across neighboring pixels (for instance, a graph in which adjacent spectra exchange messages). This would require changing the network type, since the current architecture has no mechanism to look beyond a single pixel.

We can consider integrating spatial information into the R-GCN, with nodes defined as ion images rather than peaks of one spectrum; colocalization then re-enters as in Chapter 4, but as a relation within a supervised model rather than a hand-set threshold. In either case, spatial information would add a distinct interpretive axis: it would separate ions that are class-relevant by chemistry from those that are class-relevant merely by shared localization, and would give an independent spatial check on the family-level assignments that CAM currently derives from mass structure alone.

This dataset-level, spatially aware formulation is also the natural setting for the direct *m/z*-level model outlined in Section 5.2.6. In this chapter, the R-GCN was deliberately benchmarked against single-spectrum classifiers to test the relational representation for classification; its success is what motivates repurposing it for interpretation, but the ion-level scores remain a post hoc reading of a spectrum classifier and should be taken as a perspective. Training the network to classify *m/z* directly, nodes carrying their own ion images, edges combining Δm with colocalization, would let chemical relatedness and spatial coherence be learned jointly. The limiting factor is supervision: *m/z*-level training requires molecularly annotated MSI, or at least reliably labeled spectra, which is the same ground-truth gap that prevented us from validating the weakly correlated, previously unreported ions surfaced by the model here.

5.2.6 Conclusion and perspectives

This work is pioneering because it exploits HRMS mass features to classify spectra. To achieve this, we integrate HRMS mass features using a Relational Graph Convolutional Network (R-GCN), representing a mass spectrum as a graph. We show that our method outperforms diverse classification methods when subjected to signal variations, especially in the case of mass shift and signal mixing, which is often the case in MSI experiments.

Finally, we propose to integrate a CAM strategy into our model to interpret how each *m*/*z* value is used for predictions. We demonstrate that our model combines ions belonging to different known molecular families in a biologically relevant way. Based on these results discussed through concrete and meaningful examples, we recommend the scientific community explore using our R-GCN CAM as a new tool for extracting biologically meaningful *m*/*z* from MSI mass spectra. Finally, we would like to put the novel approach proposed in this work into perspective. A key feature of our model is using a set of common mass differences. This set is composed of mass patterns frequently identified on MALDI MSI data. The mass differences between HCCA and DHB are already coded in this list. However, by adding the corresponding mass difference to the list, one can easily adapt the list to other mass patterns (mass differences), such as other MALDI matrices or other chemical transformations.

The flexibility of our approach also opens new opportunities for integration with other types of mass spectrometry data. For instance, information from electrophoretic or chromatographic techniques (e.g., retention time), as well as from ion mobility spectrometry (e.g., drift time or collision cross section), could be incorporated seamlessly as additional node features alongside *m*/*z*, mass defect, and intensity. This would extend the applicability of R-GCNs beyond MSI, offering value to researchers working with LC-MS or ion mobility-MS for tasks such as sample classification or signal interpretation, although this remains to be demonstrated.

Furthermore, we demonstrate that our model, when combined with Class Activation Mapping (CAM), can extend beyond spectrum-level classification to interpret and annotate individual *m*/*z* features. While we currently use a model trained for spectrum classification and apply CAM post hoc to identify informative ions, the same architecture could be adapted for direct *m*/*z*-level classification. This could be achieved by removing the pooling layer and applying the classification head directly to the ion embeddings after the graph convolution layers.

Such a setup could, for example, be used to train a model to identify contaminant ions. By training on a small curated list of known contaminants, the R-GCN could potentially generalize to recognize other, previously unseen contaminant *m*/*z* values that share similar structural properties. This would provide a powerful tool for automated noise filtering and data cleaning in MSI and other MS-based workflows.

Overall, our results establish the foundation for a new generation of deep learning models in mass spectrometry models that do not just classify but also interpret, generalize, and extend to new tasks in chemically meaningful ways. How broadly the approach generalizes across instruments, resolutions, and tasks will require further evaluation.

REFERENCES:

(1) Seeley, E. H.; Caprioli, R. M. MALDI Imaging Mass Spectrometry of Human Tissue: Method Challenges and Clinical Perspectives. *Trends Biotechnol.* **2011**, *29* (3), 136–143. https://doi.org/10.1016/j.tibtech.2010.12.002.

(2) Holzlechner, M.; Eugenin, E.; Prideaux, B. Mass Spectrometry Imaging to Detect Lipid Biomarkers and Disease Signatures in Cancer. *Cancer Rep.* **2019**, *2* (6), e1229. https://doi.org/10.1002/cnr2.1229.

(3) Watrous, J. D.; Dorrestein, P. C. Imaging Mass Spectrometry in Microbiology. *Nat. Rev. Microbiol.* **2011**, *9* (9), 683–694. https://doi.org/10.1038/nrmicro2634.

(4) Hu, H.; Helminiak, D.; Yang, M.; Unsihuay, D.; Hilger, R. T.; Ye, D. H.; Laskin, J. High-Throughput Mass Spectrometry Imaging with Dynamic Sparse Sampling. *ACS Meas. Sci. Au* **2022**, *2* (5), 466–474. https://doi.org/10.1021/acsmeasuresciau.2c00031.

(5) Rappez, L.; Stadler, M.; Triana, S.; Gathungu, R. M.; Ovchinnikova, K.; Phapale, P.; Heikenwalder, M.; Alexandrov, T. SpaceM Reveals Metabolic States of Single Cells. *Nat. Methods* **2021**, *18* (7), 799–805. https://doi.org/10.1038/s41592-021-01198-0.

(6) Buchberger, A. R.; DeLaney, K.; Johnson, J.; Li, L. Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights. *Anal. Chem.* **2018**, *90* (1), 240–265. https://doi.org/10.1021/acs.analchem.7b04733.

(7) Palmer, A.; Phapale, P.; Chernyavsky, I.; Lavigne, R.; Fay, D.; Tarasov, A.; Kovalev, V.; Fuchser, J.; Nikolenko, S.; Pineau, C.; Becker, M.; Alexandrov, T. FDR-Controlled Metabolite Annotation for High-Resolution Imaging Mass Spectrometry. *Nat. Methods* **2017**, *14* (1), 57–60. https://doi.org/10.1038/nmeth.4072.

(8) Alexandrov, T. MALDI Imaging Mass Spectrometry: Statistical Data Analysis and Current Computational Challenges. *BMC Bioinformatics* **2012**, *13* (S16), S11. https://doi.org/10.1186/1471-2105-13-S16-S11.

(9) Verbeeck, N.; Caprioli, R. M.; Van de Plas, R. Unsupervised Machine Learning for Exploratory Data Analysis in Imaging Mass Spectrometry. *Mass Spectrom. Rev.* **2020**, *39* (3), 245–291. https://doi.org/10.1002/mas.21602.

(10) Behrmann, J.; Etmann, C.; Boskamp, T.; Casadonte, R.; Kriegsmann, J.; Maaβ, P. Deep Learning for Tumor Classification in Imaging Mass Spectrometry. *Bioinformatics* **2018**, *34* (7), 1215–1223. https://doi.org/10.1093/bioinformatics/btx724.

(11) Xie, Y. R.; Castro, D. C.; Bell, S. E.; Rubakhin, S. S.; Sweedler, J. V. Single-Cell Classification Using Mass Spectrometry through Interpretable Machine Learning. *Anal. Chem.* **2020**, *92* (13), 9338–9347. https://doi.org/10.1021/acs.analchem.0c01660.

(12) Hanselmann, M.; Köthe, U.; Kirchner, M.; Renard, B. Y.; Amstalden, E. R.; Glunde, K.; Heeren, R. M. A.; Hamprecht, F. A. Toward Digital Staining Using Imaging Mass Spectrometry and Random Forests. *J. Proteome Res.* **2009**, *8* (7), 3558–3567. https://doi.org/10.1021/pr900253y.

(13) Tideman, L. E. M.; Migas, L. G.; Djambazova, K. V.; Patterson, N. H.; Caprioli, R. M.; Spraggins, J. M.; Van De Plas, R. Automated Biomarker Candidate Discovery in Imaging Mass Spectrometry Data through Spatially Localized Shapley Additive Explanations. *Anal. Chim. Acta* **2021**, *1177*, 338522. https://doi.org/10.1016/j.aca.2021.338522.

(14) Yang, C.; He, Z.; Yu, W. Comparison of Public Peak Detection Algorithms for MALDI Mass Spectrometry Data Analysis. *BMC Bioinformatics* **2009**, *10* (1), 4. https://doi.org/10.1186/1471-2105-10-4.

(15) Balluff, B.; Elsner, M.; Kowarsch, A.; Rauser, S.; Meding, S.; Schuhmacher, C.; Feith, M.; Herrmann, K.; Röcken, C.; Schmid, R. M.; Höfler, H.; Walch, A.; Ebert, M. P. Classification of HER2/Neu Status in Gastric Cancer Using a Breast-Cancer Derived Proteome Classifier. *J. Proteome Res.* **2010**, *9* (12), 6317–6322. https://doi.org/10.1021/pr100573s.

(16) Shimma, S.; Sugiura, Y. Effective Sample Preparations in Imaging Mass Spectrometry. *Mass Spectrom.* **2014**, *3* (Special_Issue), S0029–S0029. https://doi.org/10.5702/massspectrometry.S0029.

(17) Balluff, B.; Hopf, C.; Porta Siegel, T.; Grabsch, H. I.; Heeren, R. M. A. Batch Effects in MALDI Mass Spectrometry Imaging. *J. Am. Soc. Mass Spectrom.* **2021**, *32* (3), 628–635. https://doi.org/10.1021/jasms.0c00393.

(18) Veselkov, K.; Sleeman, J.; Claude, E.; Vissers, J. P. C.; Galea, D.; Mroz, A.; Laponogov, I.; Towers, M.; Tonge, R.; Mirnezami, R.; Takats, Z.; Nicholson, J. K.; Langridge, J. I. BASIS: High-Performance Bioinformatics Platform for Processing of Large-Scale Mass Spectrometry Imaging Data in Chemically Augmented Histology. *Sci. Rep.* **2018**, *8* (1), 4053. https://doi.org/10.1038/s41598-018-22499-z.

(19) Deininger, S.-O.; Cornett, D. S.; Paape, R.; Becker, M.; Pineau, C.; Rauser, S.; Walch, A.; Wolski, E. Normalization in MALDI-TOF Imaging Datasets of Proteins: Practical Considerations. *Anal. Bioanal. Chem.* **2011**, *401* (1), 167–181. https://doi.org/10.1007/s00216-011-4929-z.

(20) La Rocca, R.; Kune, C.; Tiquet, M.; Stuart, L.; Eppe, G.; Alexandrov, T.; De Pauw, E.; Quinton, L. Adaptive Pixel Mass Recalibration for Mass Spectrometry Imaging Based on Locally Endogenous Biological Signals. *Anal. Chem.* **2021**, *93* (8), 4066–4074. https://doi.org/10.1021/acs.analchem.0c05071.

(21) Tiquet, M.; La Rocca, R.; Kirnbauer, S.; Zoratto, S.; Van Kruining, D.; Quinton, L.; Eppe, G.; Martinez-Martinez, P.; Marchetti-Deschmann, M.; De Pauw, E.; Far, J. FT-ICR Mass Spectrometry Imaging at Extreme Mass Resolving Power Using a Dynamically Harmonized ICR Cell with 1ω or 2ω Detection. *Anal. Chem.* **2022**, *94* (26), 9316–9326. https://doi.org/10.1021/acs.analchem.2c00754.

(22) Inglese, P.; Correia, G.; Takats, Z.; Nicholson, J. K.; Glen, R. C. SPUTNIK: An R Package for Filtering of Spatially Related Peaks in Mass Spectrometry Imaging Data. *Bioinformatics* **2019**, *35* (1), 178–180. https://doi.org/10.1093/bioinformatics/bty622.

(23) Seddiki, K.; Saudemont, P.; Precioso, F.; Ogrinc, N.; Wisztorski, M.; Salzet, M.; Fournier, I.; Droit, A. Cumulative Learning Enables Convolutional Neural Network Representations for Small Mass Spectrometry Data Classification. *Nat. Commun.* **2020**, *11* (1), 5595. https://doi.org/10.1038/s41467-020-19354-z.

(24) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521* (7553), 436–444. https://doi.org/10.1038/nature14539.

(25) Liu, J.; Osadchy, M.; Ashton, L.; Foster, M.; Solomon, C. J.; Gibson, S. J. Deep Convolutional Neural Networks for Raman Spectrum Recognition: A Unified Solution. *The Analyst* **2017**, *142* (21), 4067–4074. https://doi.org/10.1039/C7AN01371J.

(26) Kune, C.; McCann, A.; Raphaël, L. R.; Arias, A. A.; Tiquet, M.; Van Kruining, D.; Martinez, P. M.; Ongena, M.; Eppe, G.; Quinton, L.; Far, J.; De Pauw, E. Rapid Visualization of Chemically Related Compounds Using Kendrick Mass Defect As a Filter in Mass Spectrometry Imaging. *Anal. Chem.* **2019**, *91* (20), 13112–13118. https://doi.org/10.1021/acs.analchem.9b03333.

(27) Richardson, L. T.; Neumann, E. K.; Caprioli, R. M.; Spraggins, J. M.; Solouki, T. Referenced Kendrick Mass Defect Annotation and Class-Based Filtering of Imaging MS Lipidomics Experiments. *Anal. Chem.* **2022**, *94* (14), 5504–5513. https://doi.org/10.1021/acs.analchem.1c03715.

(28) Janda, M.; Seah, B. K. B.; Jakob, D.; Beckmann, J.; Geier, B.; Liebeke, M. Determination of Abundant Metabolite Matrix Adducts Illuminates the Dark Metabolome of MALDI-Mass Spectrometry Imaging Datasets. *Anal. Chem.* **2021**, *93* (24), 8399–8407. https://doi.org/10.1021/acs.analchem.0c04720.

(29) Simon, C.; Dührkop, K.; Petras, D.; Roth, V.-N.; Böcker, S.; Dorrestein, P. C.; Gleixner, G. Mass Difference Matching Unfolds Hidden Molecular Structures of Dissolved Organic Matter. *Environ. Sci. Technol.* **2022**, *56* (15), 11027–11040. https://doi.org/10.1021/acs.est.2c01332.

(30) Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Las Vegas, NV, USA, 2016; pp 2921–2929. https://doi.org/10.1109/CVPR.2016.319.

(31) Perneel, M.; Heyrman, J.; Adiobo, A.; De Maeyer, K.; Raaijmakers, J. M.; De Vos, P.; Höfte, M. Characterization of CMR5c and CMR12a, Novel Fluorescent *Pseudomonas* Strains from the Cocoyam Rhizosphere with Biocontrol Activity. *J. Appl. Microbiol.* **2007**, *103* (4), 1007–1020. https://doi.org/10.1111/j.1365-2672.2007.03345.x.

(32) Burguet, P.; La Rocca, R.; Kune, C.; Tellatin, D.; Stulanovic, N.; Rigolet, A.; Far, J.; Ongena, M.; Rigali, S.; Quinton, L. Exploiting Differential Signal Filtering (DSF) and Image Structure Filtering (ISF) Methods for Untargeted Mass Spectrometry Imaging of Bacterial Metabolites. *J. Am. Soc. Mass Spectrom.* **2024**, *35* (8), 1743–1755. https://doi.org/10.1021/jasms.4c00129.

(33) Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Van Den Berg, R.; Titov, I.; Welling, M. Modeling Relational Data with Graph Convolutional Networks. In *The Semantic Web*; Gangemi, A., Navigli, R., Vidal, M.-E., Hitzler, P., Troncy, R., Hollink, L., Tordai, A., Alam, M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2018; Vol. 10843, pp 593–607. https://doi.org/10.1007/978-3-319-93417-4_38.

(34) Fey, M.; Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. **2019**. https://doi.org/10.48550/ARXIV.1903.02428.

(35) Andrić, S.; Meyer, T.; Rigolet, A.; Prigent-Combaret, C.; Höfte, M.; Balleux, G.; Steels, S.; Hoff, G.; De Mot, R.; McCann, A.; De Pauw, E.; Argüelles Arias, A.; Ongena, M. Lipopeptide Interplay Mediates Molecular Interactions between Soil Bacilli and Pseudomonads. *Microbiol. Spectr.* **2021**, *9* (3), e02038-21. https://doi.org/10.1128/spectrum.02038-21.

CHAPTER SIX

***

Interpreting the Spatial Dimension of Mass Spectrometry Images

Chapter 6: Contents

6.1 Utilizing spatial dimensions in mass spectrometry imaging 209

6.2 Application of a Deep Learning Model for Image Fusion between Microscope Image and Mass Spectrometry Imaging 213

6.2.1 Abstract 213

6.2.2 Introduction 214

6.2.3 Materials and Methods 219

6.2.3.1 Cell cultures 219

6.2.3.2 MSI acquisition 220

6.2.3.3 Microscope image acquisition 221

6.2.3.4 Image registration 223

6.2.3.5 Manual annotation of cells in the microscope images 224

6.2.3.6 U-Net model adapted for image fusion 225

6.2.3.6.1 Model input and training 225

6.2.3.6.2 Mass signal normalization 226

6.2.3.6.3 U-Net architecture 228

6.2.3.6.4 Tensor cropping 230

6.2.3.6.5 Tensor to mass signal 232

6.2.3.6.6 Generating Fused Images with CAM 233

6.2.4 Results and Discussion 235

6.2.4.1 Data preparation 235

6.2.4.2 MSI pre-processing 236

6.2.4.2.1 Microscope image pre-processing 238

6.2.4.2.2 Ion delocalization on model performances 239

6.2.4.3 Ion signal prediction 243

6.2.4.4 Comparing predicted ion location at the microscope spatial resolution 246

6.2.4.5 Microscopy image segmentation based on fused images 251

6.2.5 Conclusion and perspectives 256

6.1 Utilizing spatial dimensions in mass spectrometry imaging

The distinctive value of MSI is its ability to map molecular signals across a sample. For each *m/z* feature, integrating the measured intensities over all pixels yields an ion image (heat map) that depicts the relative abundance pattern across the surface. Because signal depends on matrix effects, ion suppression, and pixel-to-pixel variability, such images are generally interpreted semi-quantitatively unless calibrated with internal standards and appropriate normalization. MSI datasets are high-dimensional, commonly thousands to tens of thousands of *m/z* features across large pixel grids, so practical interpretation relies on data-reduction workflows: noise and peak filtering (e.g., S/N and mass-accuracy thresholds, isotope/adduct handling), *m/z* binning and peak picking, dimensionality reduction (e.g., PCA or NMF), and spatially aware clustering/segmentation to group pixels or features with shared spatial patterns. These steps reduce complexity while preserving spatial structure, enabling downstream biological interpretation.

Database matching is a widely used first-pass filter for MSI data (see Chapter 3). In practice, accurate-mass *m/z* features are matched to candidate formulas/adducts from curated libraries within a specified mass tolerance, typically with isotope/adduct handling and, where available, FDR control. This step reduces the feature set to a manageable number of ion images, often on the order of tens to a few hundred, depending on tolerance, adducts, and data quality, prioritizing molecules of interest for visualization and downstream analysis. In Chapter 4, we described Kendrick mass defect (KMD) analysis, which projects peaks onto the Kendrick mass scale (e.g., using CH_2 as the base unit) to group homologous series with shared chemistry. KMD thus both reduces the number of images and attaches interpretable chemical attributes to each group without the need for a database. For lipid-rich datasets, CH_2-based KMD consolidates hundreds of exact-mass peaks into series consistent with varying acyl-chain length, and when combined with double-bond-equivalent calculations or class constraints, can highlight series that differ by degree of unsaturation. Note that these assignments are putative without MS/MS or reference standards, and KMD performs best with high mass accuracy and consistent adduct assignment.

When discovery objectives include out-of-library molecules, annotation-driven filters (e.g., database matching or KMD grouping) become less effective. In such cases, it is advantageous to prioritize spatial distribution rather than prior chemical identity.

To automate spatial filtering and interpretation, at either the ion-image or pixel level, methods span from simple statistics to learned models trained on annotated data. Examples include:

(i) **Single-ion spatial saliency:** quantify how structured vs random an ion image is using predefined filters and spatial statistics (e.g., measures developed in Chapter 2), enabling rapid prioritization of informative images without training.

(ii) **Ion-image grouping and compression:** cluster ion images by co-localization or apply matrix factorization (e.g., PCA/NMF) to condense many correlated images into a smaller set of representative patterns, reducing the manual review burden (see Chapters 2 and 4).

(iii) **Pixel-level image segmentation:** partition the tissue into discrete spatial regions. Unsupervised segmentation maximizes differences in spectral composition without labels, whereas supervised segmentation utilizes annotated regions to learn class boundaries and can identify ions that characterize each region via loadings or feature-importance scores. In Chapter 5, we introduced a supervised classifier that represents each spectrum by a graph of mass differences, enabling segmentation based on compositional relationships rather than individual peaks.

This spatial-first workflow preserves the capacity to discover unknowns while still yielding interpretable regional markers for downstream biological analysis.

Limits of regional interpretation at routine MSI resolution. While methods (i–iii) automate the analysis of ion spatial distributions, their interpretation typically remains at the level of large, well-contrasted regions rather than individual cells. This reflects the effective spatial resolution and point-spread characteristics of most MSI modalities (commonly ~10–100 µm for MALDI; SIMS can reach sub-micrometer scales but with different chemical coverage), which introduce partial-volume averaging and mixed-pixel effects. In Chapter 4, we illustrated this on our bacterial colony and mouse brain datasets: the colonies dataset contained three clearly separated compartments (each colony plus the inter-colony zone), and the brain section was partitioned into seven major anatomical regions (Fig. S5, Kune et al. [1]). Such macro-scale patterns are robust; however, because the biological reference unit is the cell, which is generally smaller than routine MSI pixels, attributing each detected ion to individual cells is challenging without single-cell MSI or complementary modalities.

Image fusion refers to combining ion signals from low spatial resolution mass spectrometry imaging (MSI) with higher spatial resolution images, such as optical images, through data processing. In this chapter, we introduce a new algorithm for image fusion designed explicitly for mass spectrometry imaging using a convolutional neural network. This method enables us to predict ion detection levels based solely on microscope images, determine ion signals at a higher spatial resolution, link spatial patterns observed in microscope images to mass signals, and segment microscope images. By fully leveraging image fusion, we can extrapolate the spatial distribution of ions beyond the spatial resolution of MSI, providing deeper insights into the biological samples examined in MSI. This work is from *"Application of a Deep Learning Model for Image Fusion between Microscope Image and Mass Spectrometry Imaging". Written by* ***Raphaël La Rocca****, Virginie Bertrand, Gauthier Eppe, Edwin De Pauw, and Loïc Quinton. Unpublished work.*

6.2 Application of a Deep Learning Model for Image Fusion between Microscope Image and Mass Spectrometry Imaging

6.2.1 Abstract

Mass spectrometry imaging (MSI) generates molecular-rich information from biological tissues. However, interpretation of such data often necessitates the acquisition of microscope images of the same sample. The two image modalities are highly complementary since MSI molecular signals can be highly specific to biological phenotypes, and microscope imaging allows us to obtain a higher spatial resolution than MSI. However, the interpretation of the biological data from the two image modalities is not straightforward. To achieve this goal, image fusion has been proposed; those methods map spatial patterns from the microscope image to ion intensities. By doing so, it is possible to estimate the ion spatial distribution at the resolution of the microscope image. In this work, we proposed a new algorithm for image fusion by adapting a deep neural network architecture called U-Net to image fusion by utilizing Class activation map. U-Net models have been highly efficient in learning spatial patterns from various types of microscope images, eliminating the need for complex image processing. We train and evaluate our method on a coculture of two types of human cells. The training is performed on a coculture where the two cell types are mixed and cannot be distinguished at the MSI spatial resolution. Then, we evaluate the method's capacity to predict each cell type's MS signal. This is performed by assessing the model prediction on monoculture of each cell type by comparing the predicted phospholipid signal to their MS detected signal at MSI spatial resolution (100 µm) and the microscope spatial resolution (1.7 µm). Finally, the predicted phospholipid signal at the microscope resolution is used for segmenting the cell type.

6.2.2 Introduction

Mass spectrometry imaging (MSI) is a powerful imaging technique that enables the detection of diverse compounds from various sample surfaces[2]. The detected signal can then segment the sample spatially into different regions of interest. For instance, MSI has been used in clinical applications to identify healthy and non-healthy parts of tissues by detecting a change in molecular composition[3], such as in cancer biopsies[4]. In addition, MSI has been exploited in biological studies, including single-cell imaging[5,6], amyloid plaque detection[7], and bacterial colony identification[8].

In contrast, bright-field microscopy is a fast-imaging method that provides high spatial resolution, allowing for the visualization of spatial patterns at the cellular level. Due to those advantages, the method has a place of choice in clinical applications. Indeed, microscopy is routinely performed with H&E staining (hematoxylin and eosin) to highlight spatial patterns that increase the specificity of the image toward specific cell phenotypes. This is helpful for pathologists in classifying histological tissue regions[9,10]. However, analyzing this type of data is often a manual-intensive process. Automating the segmentation of microscopy images is crucial for expediting, simplifying, and accelerating microscopy data interpretation by focusing the analysis on a smaller tissue region.

Recent developments in deep learning with Convolutional Neural Networks (CNN) have made it possible to segment microscope images[11,12]. These models can learn from images annotated by histopathologists and build a model that recognizes tissue regions, automating a part of the work done by pathologists. These models allow the recognition of complex image features (or patterns) by themselves (eliminating the need for feature engineering) and predict annotated regions on new microscope images[13]. Spatial patterns, or spatial features, are structured arrangements or recurring configurations of visual elements (such as edges, textures, shapes, or colors) within an image. These patterns can be detected, quantified, and analyzed algorithmically, providing critical insights into the image's content, structure, or context. Complex

spatial pattern can be defined as a combination of multiple simpler spatial pattern elements (e.g., edges, textures, shapes, or colors) into a larger, more intricate, or structured arrangement. CNN models excel at learning complex spatial patterns by hierarchically combining simpler elements, allowing the recognition of cells of tissue regions.

However, training such algorithms requires a significant number of annotated microscope images. This involves manually delineating tissue regions or cell borders, which is a time-consuming and knowledge-intensive task. The U-Net encoder-decoder network architecture is one of the most popular CNN models in biomedical image segmentation, as it requires fewer images than other powerful segmentation models for training. This model has been used to automatically segment cells, track cells, and segment tissues based on microscope images and has already achieved top ranks in biomedical data analysis benchmarks[11,14].

MSI provides valuable information about the molecular composition of spatial regions within tissues, which can assist in understanding biological mechanisms more deeply than microscope images. Moreover, automating the detection of regions of interest by segmentation methods within samples is more straightforward in MSI compared to microscopy images. Unlike microscope images, where a single pixel contains only RGB values, MSI pixels correspond to MS spectra, which consist of the measurement of multiple ions that often have a higher specificity to the underlying materials. Due to this increased specificity, segmentation methods can classify an individual pixel by solely relying on its mass spectrum information. This classification can be performed in a supervised[15–17] or an unsupervised manner for exploratory analysis[18] by using the relative intensity of the pixel's ions. In contrast to RGB images, it is necessary to incorporate the pixel information and the adjacent pixels to classify the pixel, since each pixel's RGB value has poor predictive power.

While MSI can be used for image segmentation of large spatial regions without using complex spatial models, estimating mass signals at the cell level is often necessary for accurate biological interpretation. However, this level of understanding is hampered by the common low spatial resolution of MSI. Indeed, a spatial resolution of 1-10 µm is necessary to distinguish molecules originating from adjacent cells. This convolution issue in MSI can be addressed by increasing MSI spatial resolution. Several in-house methods adapted to customized instrument configurations have been developed to increase MSI spatial resolution. It should be noted that increasing spatial resolution comes at the cost of reduced sensitivity (as less material is being ionized), increased acquisition time (which may not be suitable for large experiments), and increased data storage needs, since MS and high spectral MS resolution generate heavy amounts of information[19].

To fill this need, in addition to the MS image, a microscope image of the same sample is often acquired to assist in interpreting MS ion distribution[20,21]. High-precision image alignment is necessary to compare small elements, such as cells, between multiple images of the same sample. Image registration is the method that aligns images of different imaging modalities, such as MALDI (Matrix-Assisted Laser Desorption/Ionization) imaging and optical microscopy or histological sections. This alignment helps to correlate molecular data with anatomical structures, enabling a more comprehensive analysis and interpretation of the biological sample. Image registration is performed by determining the position of visual clues between the different images and then applying a geometric transformation to each image so that the visual clues end up with the same pixel coordinate system across all images. Finding visual clues and applying the geometric transformation can be manual or semi-manual[9] . Microscope image registration has also been used in different applications such as microdissection based on microscope images[22], acquisition of single cell measurements[5], for classification of epithelial tissue from stroma in esophageal dysplasia by combining MS and microscopy images. Even if microscope image

registration helps in MSI interpretation of well-distinct regions of tissues, MSI spatial resolution still limits the interpretation of biological structures at the MSI resolution.

Image fusion is a credible solution to this limitation. Image fusion is a technique that combines the properties of two images of the same sample by finding relationships between the signals of the images. Ideally, the image pair comes from imaging modalities with complementary information, such as high spatial and spectral resolution. For example, image fusion can combine information from a microscope image and an MSI and form a new image with the spatial resolution of the microscope image and the spectral information of the MSI[23,24]. This increase in resolution can then be used to interpret mass signals of objects not resolved at the MSI spatial resolution.

Image fusion was initially proposed for MALDI imaging in 2015 in Raf Van de Plas–et al [25] work. They proposed a fusion method that combines MALDI MSI and microscope images of stained organ tissue samples to interpret ion images at the spatial resolution of the bright-field microscope image. The technique first expands each pixel's RGB value to a more extensive set of variables by applying predefined image filters. Then, image registration allows finding the MSI pixel corresponding to each microscope pixel. Finally, a multivariate linear model (partial least-squares (PLS)) regression fits the microscope pixel variables to the MSI pixel's *m/z* intensities.
The method was successfully applied to different samples of organ tissues analyzed by MALDI MSI and then stained with various methods. They demonstrate different applications of their method, such as predicting ions at the resolution of a microscope image and predicting ions in tissue areas not measured by MSI.

Our work differs from Van de Plas et al. by proposing an 1/ alternative fusion method based on deep learning CNN algorithms and 2/ evaluating the algorithm at the cell level on a more elementary dataset consisting of a mix of two cell types. CNN methods have the advantage of generating a combination of image filters that best fit the learning task. Those filters should replace image processing and predefined filter applications,

such as in the work of Van de Plas et al. These methods should be more robust to different microscope images without adapting the image filters. In the current study, we propose to use a particular CNN architecture called a U-Net. Our inspiration for using U-Net for image fusion comes from a paper on weakly supervised image cropland segmentation in remote sensing imagery[26]. We have reimplemented their proposed method for image fusion on MS and microscope images. Image fusion methods are often demonstrated on biological tissue images. This type of data makes the interpretation of the fusion method challenging and limits its evaluation to large sections of tissue and not at the cell level. To address this, we propose to generate a simplified biological tissue and evaluate our fusion method on that dataset. This dataset is composed of a mix of only two cell types, a cell culture consisting of MCF-7 cells and WI-38 cells.

We initially trained our model on a coculture of MCF-7/WI-38, where cells are spatially mixed, and evaluated it on a monoculture of each cell type. By separating each cell type, MS signals specific to each type (ground truth) could be determined. Then, the model is evaluated on images of the monoculture by comparing whether the predicted signals conserve the specificity highlighted previously and whether the predicted ion location at the microscope resolution correlates with the cell location from the microscope image.

The evaluation shows that the predicted fused images reach a spatial resolution of 1.7 µm from an MSI acquired at 100 µm spatial resolution. At this spatial resolution level, the model correctly predicts lipid signals specific to each cell type by associating the cells' spatial patterns with the lipid’s expression.

Finally, we investigate the use of fused images for microscope image segmentation. As explained above, molecular information is often richer than the RGB pixels of a microscope image, which is often considered by the segmentation algorithm to classify a pixel solely from the spectrum information. By fusing an MSI image with a microscope

image, we can utilize the richer molecular information as an input for the segmentation algorithm. We therefore applied classification algorithms to a fused image of an MCF-7/WI-38 coculture to segment cell types at the spatial resolution of the microscope image.

6.2.3 Materials and Methods

6.2.3.1 Cell cultures

Two human cell lines were used in this study: WI-38 (human lung fibroblasts) and MCF-7 (human breast adenocarcinoma cells). Both cell lines were cultured using the same protocol and the same culture medium. The culture medium consisted of high-glucose DMEM supplemented with GlutaMAX and sodium pyruvate, complemented with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin.

Cells were thawed directly into this medium and expanded in T25 culture flasks. A single initial T25 flask was split into four T25 flasks, all maintained in the same culture conditions. The flasks were incubated at 37 °C in a humidified atmosphere with 5% CO_2 for 72 h, allowing the cells to reach approximately 80–90% confluence (i.e., 80–90% of the growth surface covered by adherent cells).

To prepare the samples for MALDI-MSI and microscopy, cells were harvested from a T25 flask by enzymatic detachment with trypsin. The detached cells were then transferred onto indium tin oxide (ITO)-coated glass slides. Cell deposition onto the slides was performed using a centrifugation-based concentration device (Cellspin I, Tharmac) equipped with a funnel system that sediments the cells onto the slide surface. Importantly, the cells were not further cultured on the slide; they were physically deposited and immobilized on the ITO surface immediately prior to downstream acquisition.

Three types of slides were prepared a monoculture of WI-38 cells, a monoculture of MCF-7 cells, and a mixed WI-38/MCF-7 coculture.

For the monoculture slides, cells from a single line were deposited onto individual ITO slides. For the coculture slide, WI-38 and MCF-7 cells were pooled before deposition to generate a spatially intermixed population. The coculture was prepared to avoid large segregated regions dominated by only one cell type, with the goal of creating local mixtures of both phenotypes that would not be resolved at the MSI spatial resolution.

These slides were then used for (1) brightfield and phase-contrast microscopy imaging, and (2) matrix-assisted laser desorption/ionization (MALDI) mass spectrometry imaging (MSI), as described in Sections 6.2.3.2 and 6.2.3.3.

6.2.3.2 MSI acquisition

After deposition, each ITO slide was first imaged by optical microscopy (see Section 6.2.3.3). The same slides were then prepared for MALDI mass spectrometry imaging (MSI).

Before matrix deposition, the slides were placed in a vacuum desiccator for approximately 15 minutes, or until no visible residual moisture was present on the surface. Dried samples were then coated with matrix using an automated pneumatic sprayer (SunCollect MALDI spotter, SunChrom, Friedrichsdorf, Germany). The matrix solution consisted of 5 $mg \cdot mL^{-1}$ α-cyano-4-hydroxycinnamic acid (α-CHCA) dissolved in methanol and Milli-Q water acidified with trifluoroacetic acid (MeOH:H_2O:TFA, 9:0.99:0.01, v/v).

MALDI-MSI was acquired on a SolariX 9.4T Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometer (Bruker Daltonics, Bremen, Germany). The acquisition was performed in positive ion mode with a raster step size of 100 µm. The laser was operated at 200 Hz with 20 shots per pixel and a laser power of 70%.

Following acquisition, MSI data were exported in the imzML format. The spatial coordinates of each acquired mass spectrum (pixel positions) were extracted from the corresponding .xeo file generated by the Bruker FlexImaging software and used downstream for spatial registration to the microscope images.

6.2.3.3 Microscope image acquisition

Microscopy was performed on the ITO-coated slides prior to MALDI-MSI acquisition ("pre-MALDI imaging") and again on the same slides after MSI acquisition ("post-MALDI imaging").

For pre-MALDI imaging, slides were imaged using an Olympus IX81 inverted microscope equipped with a 10× objective. Phase-contrast illumination was used to enhance morphological differences between the two cell types. Under these conditions, MCF-7 cells typically appeared as round or slightly flattened spheroidal cells of approximately 20 µm in diameter, adhering weakly to the ITO surface. In contrast, WI-38 cells exhibited an elongated fibroblast-like morphology, spreading broadly on the surface and frequently forming filamentous extensions exceeding 200 µm in length (Figure 6.1b).

After MALDI-MSI acquisition, the same ITO slides were re-imaged on the IX81 microscope in bright-field mode using the same objective but without phase contrast. The purpose of this second acquisition was to visualize the MALDI laser ablation craters. These ablation craters, generated during MSI acquisition, appeared as circular regions with an average diameter of approximately 90 µm (Figure 6.1c). These images are referred to as the post-MALDI microscope images and were used both to assess laser coverage and to support spatial registration between MSI pixels and the optical images.

All microscope images (pre- and post-MALDI) were exported in PNG format for downstream processing, registration, and model input.

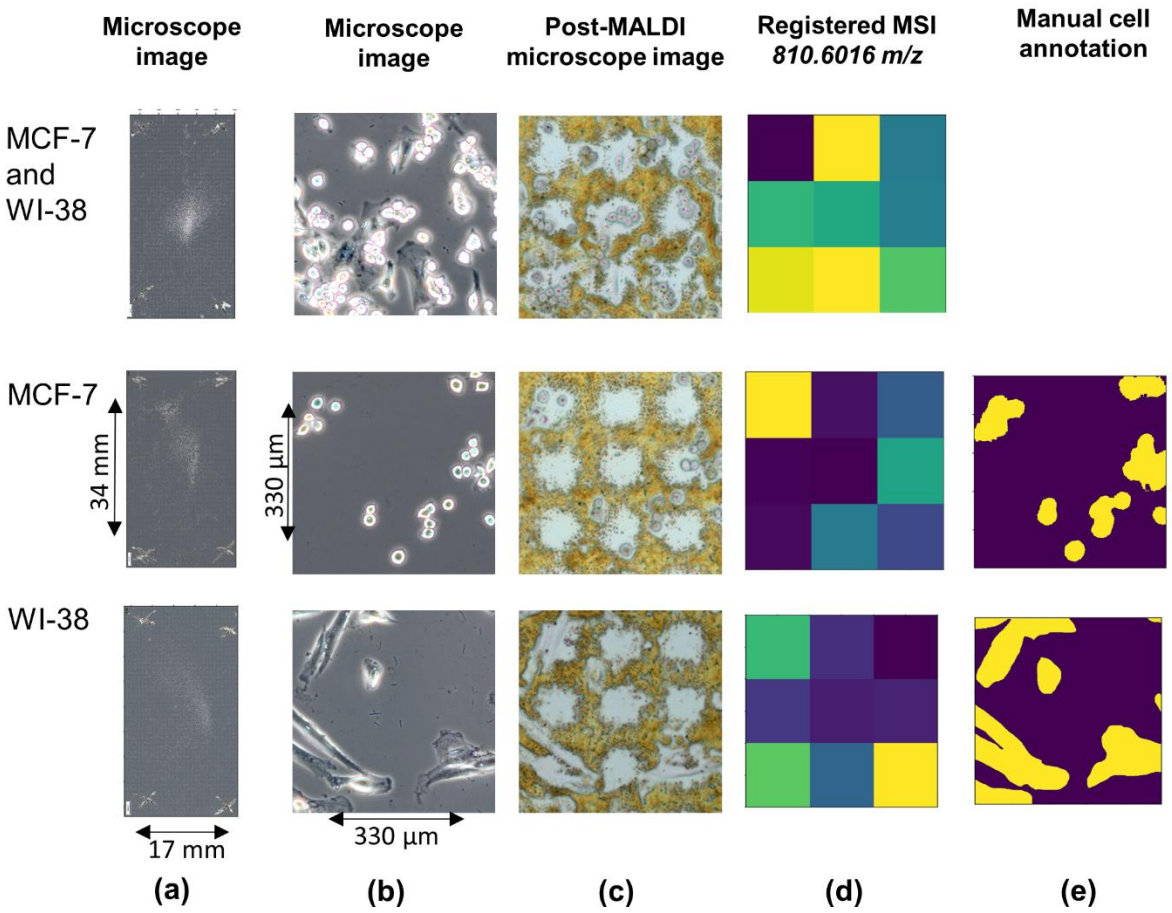


Figure 6.1. *Overview of the imaging modalities and annotations used in this study. Across all panels, the MSI raster step is 100 µm, and the MALDI ablation craters are ~90 µm in diameter, so each crater nearly fills its own raster cell; downstream, the fusion model pairs each MSI spectrum with a microscope patch spanning that crater and its eight nearest neighbors (~3 × 3 raster cells, ≈330 µm; see Figure 6.2a). (a) Full pre-MALDI microscope image of an ITO slide. (b) Higher-magnification phase-contrast views showing cell morphology: MCF-7 cells appear as weakly spread spheroidal cells (~20 µm diameter), whereas WI-38 cells display an extended fibroblast-like morphology with long protrusions (>200 µm). (c) Post-MALDI bright-field image of the same region, highlighting MALDI laser ablation craters (~90 µm diameter) and their spatial correspondence to the pre-MALDI microscope image (b). (d) Registered MSI signal for a representative ion (m/z 810.6016), shown at the MSI raster resolution and aligned to the same field of view. (e) Manual pixel-level annotations of cell identity (WI-38, MCF-7, background) used for evaluation and supervised baselines.*

6.2.3.4 Image registration

To align the MSI data with the corresponding microscope images, we performed a two-step image registration procedure. Image registration is the process of transforming multiple images of the same specimen into a shared coordinate system so that spatial features correspond pixel-to-pixel.

We developed in-house software to perform manual registration using projective homography transformations. A homography is a geometric transformation that maps one image plane onto another while allowing for translation, rotation, scaling, and perspective distortion. This class of transformation is more flexible than a purely affine model and is well-suited for aligning images acquired with different modalities and optical geometries.

In the first step, we registered the MSI pixel grid to the post-MALDI microscope image. We defined 12 pairs of corresponding landmarks between these two modalities. For the post-MALDI microscope image, the landmarks were defined as the centers of the MALDI laser-ablation craters. For the MSI acquisition, the corresponding landmarks were the pixel coordinates extracted from the .xeo file generated by the FlexImaging software. Fitting a homography to these landmark pairs provided a projective transform that maps MSI pixel coordinates into the coordinate frame of the post-MALDI microscope image.

In the second step, we registered the post-MALDI microscope image to the pre-MALDI microscope image of the same slide. Again, we defined 12 pairs of corresponding landmarks, this time using morphological features (cell bodies or cell outlines) that remained partially visible after laser ablation and that could also be identified in the pre-MALDI phase-contrast image. A second homography was fitted to these landmarks to align the post-MALDI and pre-MALDI images.

Finally, the two homography transformations were applied in series. This allowed us to map each MSI pixel (as defined by its coordinates in the .xeo file) into the coordinate

system of the original pre-MALDI microscope image. This registered coordinate system is then used throughout the study to (i) extract the correct microscope image patch corresponding to each acquired MSI spectrum and (ii) overlay MSI ion intensities on microscopy for visualization and model evaluation.

We selected homography-based registration rather than a simpler affine transformation because it provided better alignment of local cellular structures and crater positions across modalities. The outcome of this registration workflow is illustrated in Figure 6.1b–d.

6.2.3.5 Manual annotation of cells in the microscope images

Manual annotations were generated on selected subregions of the monoculture microscope images. In these regions, each pixel was assigned to one of three classes: WI-38 cell, MCF-7 cell, or background. The resulting annotation masks were deliberately drawn slightly larger than the visible cell body to include the bright halo surrounding each cell in phase-contrast microscopy. This halo is an optical artifact caused by phase-contrast imaging and is strongly correlated with cell presence (Figure 6.1e). Including the halo in the annotation reduces the risk of under-labeling cell boundaries.

No manual annotation was performed on the coculture slide. In the coculture, WI-38 and MCF-7 cells are intermixed at a spatial scale below the MSI raster size, and their morphology partially overlaps in crowded regions. This makes it difficult to confidently assign a pixel to a specific cell type based on morphology alone (Figure 6.1b).

6.2.3.6 U-Net model adapted for image fusion

6.2.3.6.1 Model input and training

For each acquired MSI pixel, we generated a paired input-output sample consisting of (i) an optical image patch and (ii) the corresponding mass spectral signal. The optical patch was obtained from the pre-MALDI microscope image using the registration procedure described above (Section 6.2.3.4).

The geometry of this pairing is dictated by the acquisition. The MSI raster step was 100 µm, and the laser ablation craters measured on the post-MALDI images had an average diameter of ~90 µm (Figure 6.1c), so each crater nearly fills its own 100 µm raster cell. For every ablation position, we extracted a square region of approximately 330 × 330 µm centered on the crater, that is, a window spanning the central raster cell and its eight nearest neighbors (~3 × 3 raster cells or MSI pixels; Figure 6). It is important to stress that the mass spectrum we are trying to predict belongs only to the central crater; the eight neighboring craters and the inter-crater regions are included not as prediction targets but as morphological context. This larger field of view ensures that cells that contribute, at least in part, to the central spectrum (including any laterally delocalized material around the crater) fall within the input patch, while also providing the convolutional encoder with sufficient surrounding structure (full cell morphology) to interpret the central region reliably.

The extracted region was saved as an RGB image and resized to 192 × 192 pixels with three color channels (192 × 192 × 3) for use as input to the network (Figure 6.2a); pixel intensities were scaled to the [0,1] range. At this scale, one input pixel corresponds to ≈1.7 µm, and the central crater occupies roughly the central 27% of the patch width (90 µm / 330 µm), a fraction we return to when defining the crop (Section 6.2.3.6.4).

In parallel, the MSI data for that same ablation position were reduced to a vector of selected ion intensities. From the full mass spectrum, we retained only a subset of ions of interest (e.g., lipid species and matrix-related ions) and associated their intensities

with the corresponding microscope patch. This produces, for each patch, an *M*-dimensional target vector, where *M* is the number of selected *m/z* values. Conceptually, the spatial dimension (192 × 192) reflects the microscope view of a single MSI pixel, while the *M*-dimensional vector represents the molecular profile measured at that (central) MSI pixel.

We describe the proposed fusion model as weakly supervised. By this we mean that the only training target attached to each microscope patch is its associated MSI spectrum, a single M-dimensional vector of ion intensities measured at the corresponding acquisition pixel, and that the model is never given manual pixel-level annotations of cell identity or per-pixel ion labels. The supervision is weak in two senses: it is spatially coarse, since one spectrum supervises an entire ~330 µm patch (one MSI pixel) rather than each microscope pixel; and it is indirect, since the target is a molecular profile rather than the cellular structures the network must learn to associate with it. From these patch-level molecular targets alone, the model learns to localize ions at microscope resolution (~1.7 µm) through CAM. This contrasts with fully supervised microscopy segmentation, which relies on dense, manually delineated pixel-level labels; here, such manual annotations (Section 6.2.3.5) are used only for baselines and evaluation, not to train the fusion model.

6.2.3.6.2 Mass signal normalization

The measured MSI ion intensities span several orders of magnitude across different *m/z* values, which would otherwise cause high-abundance ions to dominate the training loss, the scalar objective, minimized during optimization, that quantifies the discrepancy between the network's predicted ion intensities and the measured target values (here the mean squared error, defined in Section 6.2.3.6.5). To balance the contribution of each ion during optimization, we applied the same normalization

procedure to every selected ion (each column in the ion intensity table) before training the network.

For each ion, the following steps were applied:

1. **Log transform with pseudocount**

   The raw ion intensity for that ion was first shifted by adding a constant value of 1 to avoid taking the logarithm of zero, and then transformed using the natural logarithm.

2. **Standardization (z-scoring)**

   The log-transformed values were standardized by subtracting their mean and dividing by their standard deviation, computed for that ion on the training dataset. This produces, for each ion, approximately zero mean and unit variance.

3. **Clipping of extreme values**

   The standardized values were then clipped to the interval $[-2,2]$. Values below -2 were set to -2, and values above 2 were set to 2. This step limits the influence of extreme or noisy measurements (for example, unusually bright spots or local delocalization artifacts) on the loss.

4. **Rescaling to $[0, 1]$**

   After clipping, the values were linearly rescaled from $[-2,2]$ to $[0,1]$ using

$$x_{\text{scaled}} = \frac{x_{\text{clipped}} + 2}{4}$$

After this step, every ion intensity used as a training target lies between 0 and 1. The neural network is therefore trained to predict, for each microscope patch, an M-dimensional vector of normalized ion intensities with values in $[0,1]$, where M is the number of selected *m/z* values.

Because the normalization parameters (mean and standard deviation per ion) are known, it is in principle possible to approximately invert this transformation after prediction: the predicted output in $[0,1]$ can be mapped back to the clipped z-score

range $[-2,2]$, then un-standardized using the stored mean and standard deviation for that ion, and finally exponentiated and shifted back (inverse of the $\log(\text{intensity} + 1)$step). Note that clipping at $[-2,2]$ makes the inversion approximate for saturated values, since information beyond that range is intentionally discarded.

For evaluation, including separate slides acquired in different conditions, we reuse the per-ion mean and standard deviation computed once on the training set; they are not recomputed from the evaluation slides. Re-fitting them per slide would let each dataset's own distribution set its scale, making normalized intensities incomparable across slides and leaking evaluation statistics into preprocessing. Reusing the fixed training parameters keeps ion intensities on a consistent normalized scale across training and testing.

6.2.3.6.3 U-Net architecture

To process the microscope image patches and learn spatial features that explain the mass spectrometry signal, we adapted a U-Net–style encoder–decoder architecture inspired by Ronneberger et al.[14] The network takes as input a $192 \times 192 \times 3$ tensor corresponding to the registered microscope patch (see Section 6.2.3.6.1, “Model input”). The first part of the network (encoder) consists of a sequence of convolutional blocks. Each encoder block applies two 3×3 convolutions with ReLU activation (optionally followed by dropout), followed by 2×2 max pooling to downsample the spatial resolution. After each block, the number of feature channels is doubled (e.g. 32 → 64 → 128 → 256 → 512), and intermediate feature maps are stored to be used as skip connections.

At the bottleneck, a central block of convolutions with dropout is applied at the lowest spatial resolution. The second part of the network (decoder) mirrors the encoder. Each decoder block begins with a 2×2 transposed convolution (stride 2) to upsample the feature map, concatenates the corresponding skip connection from the encoder, and

then applies two additional 3×3 convolutions with ReLU. During decoding, the number of feature channels is halved at each step (e.g. 512 → 256 → 128 → 64 → 32), and spatial resolution is progressively restored to the original 192×192 size.

After the final decoder block, the network outputs a high-resolution feature tensor of shape $192 \times 192 \times F$, where $F = 32$ in our experiments (Figure 6.2). This tensor represents spatial features learned from the microscope image that are predictive of the local MSI signal. We selected $F = 32$ as the base number of feature channels (the initial number of filters in the first encoder block) and $n_{\text{blocks}} = 4$ encoder/decoder stages. These hyperparameters were chosen empirically by monitoring prediction performance on monoculture data and by inspecting the spatial localization in the resulting class activation map (CAMs). The full implementation is provided in our code repository (https://github.com/LaRoccaRaphael/Mass2micro).

This 32-channel feature tensor is then used in two ways downstream: (i) it is spatially pooled and passed through a final layer to predict the ion intensity vector associated with the MSI pixel, and (ii) it is later reweighted with the final layer's weights to generate per-ion fused images at microscope resolution (see below, "Generating fused images with CAM").

6.2.3.6.4 Tensor cropping

Each MSI pixel is associated with a 192 × 192 microscope patch that spans the central crater and its eight neighbors (Section 6.2.3.6.1), but the spectrum we predict was generated only by the central ablation spot. To restrict the network's prediction to the region that actually contributed material to that spectrum, we apply a spatial cropping step. Crucially, this crop is applied to the U-Net output feature map, not to the input image.

The last decoder layer of the U-Net produces a feature tensor of size 192 × 192 × 32. Because the U-Net's skip connections carry high-resolution spatial information from the encoder directly to the decoder, this feature tensor remains spatially registered to the input patch: the center of the feature map corresponds to the center of the microscope patch (the central crater), and its periphery corresponds to the neighboring craters. We therefore apply a custom cropping layer (Crop_Tensor) that retains only the central fraction c of the feature map in both spatial dimensions ($0 < c \leq 1$), yielding a tensor of size $(192c) \times (192c) \times 32$ (Figure 6.2c). For example, $c = 0.30$ keeps the inner ~30% of the feature map around the ablation center.

The crop fraction c thus has a direct geometric meaning relative to the acquisition (Figure 6.2 (geometry panel)). The physical crater occupies ~27% of the patch width (90 µm in a ~330 µm patch), so $c \approx 0.27$ corresponds to the crater core; one full raster cell corresponds to $c \approx 0.30$ (100 µm / 330 µm); and the near edges of the neighboring craters lie ~55 µm from the center, which corresponds to $c \approx 0.33$. For c below this last value the crop stays within the central raster cell and excludes the neighboring crater footprints. Consequently, values of c in the ~0.27–0.33 window admit a thin peripheral halo around the central crater, accounting for local ion delocalization (matrix wetting, solvent spreading, plume effects), while still leaving the neighbors out of the pooled prediction. Smaller c emphasizes the crater core (least delocalization); larger c progressively includes more peripheral signal and, beyond ~0.33, the neighboring craters themselves. This is the precise sense in which cropping both captures genuine local delocalization and mitigates contamination from neighboring craters: it is the upper bound on c, rather than the cropping per se, that separates the two regimes.

We use a square window rather than a circular mask matched to the crater outline, both for implementational simplicity, it is the natural output of an axis-aligned crop on the feature tensor and because the choice of window shape has little effect on the prediction: the subsequent global average pooling, together with the network's large

receptive field, makes the pooled feature vector insensitive to the exact boundary of the retained region

On one hand, ion delocalization means that part of the material giving rise to the measured spectrum originates outside the physical crater. For the model to associate the optical morphology with the spectral signal, the retained region must therefore extend somewhat beyond the crater so that this surrounding material is included; otherwise, the image would lack the very structures that produced part of the signal. On the other hand, the patch also contains material that did not contribute to the central spectrum, neighboring craters and more distant background, and retaining too much of this unrelated input would add noise to the model and weaken the learned image-to-spectrum association. The crop fraction c sets this balance: large enough to capture the delocalized signal around the crater, but small enough to exclude material irrelevant to the central spectrum.

A second, more subtle point concerns where the crop is applied. We crop the last layer of the U-Net (its output feature map) rather than the input image. The reason is that, as the patch passes through the encoder–decoder, the successive convolutions progressively aggregate information across the field of view, so that information from the full patch converges toward the central locations of the feature map. A feature vector at the center of this map, therefore, already summarizes morphology drawn from the whole patch, not only from the pixels directly beneath it. As a consequence, we do not need to worry about whether an object is cut by the crop window: even an object that lies partly outside the cropped region has already contributed to the central features through the network's receptive field. Cropping the feature map thus restricts which locations are pooled into the prediction without discarding the contextual information those locations have already absorbed, which would not be the case had we cropped the raw input image, where any object outside the crop would simply be removed. After cropping, the resulting $(192c) \times (192c) \times 32$ tensor is passed to a global average pooling layer, which averages each of the 32 feature channels across the

cropped spatial window to produce a 32-dimensional feature vector. This pooled vector is then used to predict the ion intensity vector for that MSI pixel (see Figure 6.2d).

6.2.3.6.5 Tensor to mass signal

After cropping the U-Net output feature map to the central region (size $(192c) \times (192c) \times 32$; see previous section), we transform this tensor into a prediction of the MSI ion intensities for that pixel.

This transformation is performed in two steps:

1. Global average pooling

   We apply a 2D global average pooling operation to the cropped tensor. This pooling step averages each of the 32 feature channels over the cropped spatial window, yielding a 32-dimensional feature vector (conceptually equivalent to a $1 \times 1 \times 32$tensor). In other words, this step summarizes "how strongly" each learned spatial feature is present within the ablation footprint.

2. Fully connected output layer

   This 32-dimensional vector is then passed through a final fully connected (Dense) layer. The Dense layer maps these 32 values to an M-dimensional output vector, where M is the number of selected ions (*m/z* values). The output of this layer is interpreted as the predicted normalized intensity of each ion for that MSI pixel (Figure 6.2d). In the implementation, this final Dense layer uses a sigmoid activation and a non-negative kernel constraint, ensuring that the predicted ion intensities lie in the range $[0,1]$, which matches the normalized target scale described above.

The network is trained end-to-end by minimizing the mean squared error (MSE) between the predicted M-dimensional ion intensity vector and the corresponding normalized MSI ion intensities measured experimentally for that pixel.

6.2.3.6.6 Generating Fused Images with CAM

After training, we convert the network's internal feature representation into spatially resolved ion maps using Class Activation Mapping (CAM). CAM provides, for each output neuron, a heatmap that highlights which spatial regions in the input image were most important for that prediction.

In our case, the final prediction layer takes a 32-dimensional feature vector (obtained after global average pooling over the cropped tensor) and maps it to an M-dimensional ion intensity vector using a Dense layer with learned weights. To generate fused images, we reuse those learned weights and project them back onto the high-resolution feature tensor produced by the U-Net decoder before cropping (the $192 \times 192 \times 32$ tensor shown in Figure 6.2b). Concretely, for each ion m, we compute a weighted sum of the 32 feature channels across spatial locations, using the corresponding weights from the final Dense layer. This produces a 192×192 activation map for ion m.

Applying this procedure to all M ions yields a fused image stack of size $192 \times 192 \times M$, where each channel is an estimated spatial distribution of a given ion at the microscope resolution (Figure 6.2e). These CAM-derived fused images can then be interpreted as high-resolution ion localization maps, even though the model was trained only with patch-level spectral targets and not with pixel-level ion labels.

Two properties of the architecture make this possible and biologically interpretable: (i) the U-Net preserves fine spatial structure through its skip connections, so the $192 \times 192 \times 32$ feature tensor retains cell-scale morphology; and (ii) the final Dense layer is constrained to have non-negative weights, which ensures that the CAM maps reflect additive evidence for ion presence rather than arbitrary positive/negative cancellations.

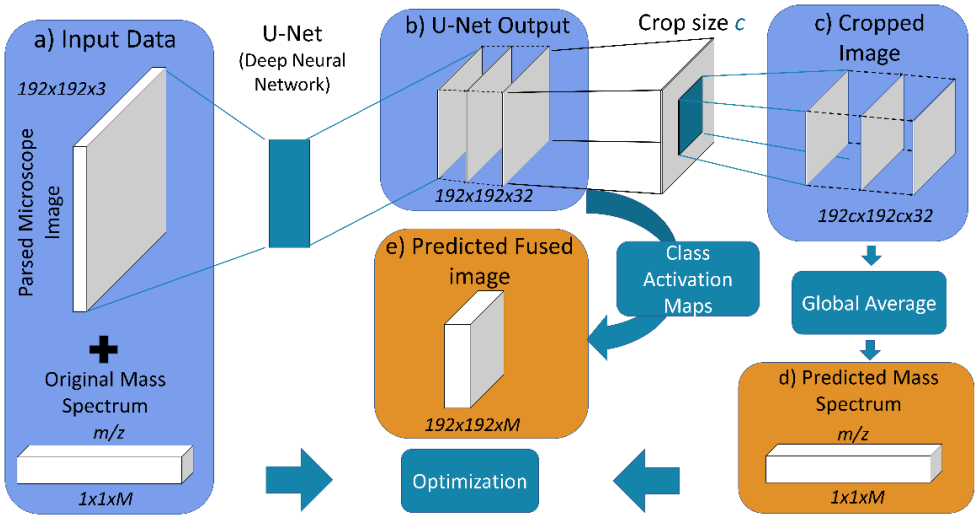


Figure 6.2. *Pipeline of the proposed fusion method combining a U-Net architecture with Class Activation Mapping (CAM). Throughout, the optical patch is centered on a single MSI pixel (the central ablation crater) and also contains its eight neighboring craters as context; the spectrum to be predicted belongs only to the central crater. While CAM is based on the entire patch.*
(a) *For each MSI pixel, we pair the mass spectrum (vector of* M *selected ion intensities; size* $1 \times 1 \times M$*) with the corresponding registered microscope image patch (size* $192 \times 192 \times 3$*).*
(b) *The microscope patch is passed through the U-Net, which produces a high-resolution feature tensor (size* $192 \times 192 \times 32$*).*
(c) *The central region of this tensor is cropped according to a crop parameter* c*, yielding a tensor of size* $(192c) \times (192c) \times 32$*. This crop approximates the laser ablation footprint and possible ion delocalization.*
(d) *The cropped tensor is spatially pooled (global average pooling) and passed through a final fully connected layer to predict the ion intensity vector for that MSI pixel (size* $1 \times 1 \times M$*). The network is trained by minimizing the error between this predicted ion vector and the experimental MSI intensities.*
(e) *After training, CAM is used to project the learned weights of the final layer back onto the full (uncropped) U-Net feature tensor, producing a fused image for each ion. This yields* M *spatial ion maps at the microscope resolution (size* $192 \times 192 \times M$*), effectively estimating ion localization at the cellular scale.*

6.2.4 Results and Discussion

This section first details how we prepared paired inputs from the pre-MALDI microscopy, post-MALDI microscopy, and MSI (registration, patch extraction, and ion selection). We then quantify how the crop fraction c (which approximates the ablation footprint and ion delocalization) affects predictive performance. Finally, we evaluate the model along three axes:

1. **Spectral prediction at MSI resolution**: Compare the model-predicted ion intensity vector for each patch to the experimental MSI intensities (normalized scale).
2. **Spatial localization at microscope resolution**: Use CAM to generate fused ion maps and assess whether predicted ion enrichment aligns with cell morphology and monoculture specificity.
3. **Downstream segmentation**: Test whether fused ion maps improve simple pixel-wise classification of cell type (vs. RGB alone).

Together, these analyses probe accuracy at the spectrum level, fidelity of spatial localization, and practical utility for microscopy segmentation.

6.2.4.1 Data preparation

Before analyzing model performance, we constructed paired inputs from microscopy and MSI for each acquisition pixel. Briefly, MSI data were recalibrated, and a subset of ions (m/z) was selected; pre- and post-MALDI microscope images were co-registered to the MSI grid, and microscopy patches centered on each ablation spot were extracted. Each training example thus consists of (i) a registered RGB microscope patch (≈330 µm field of view, resized to $192 \times 192 \times 3$) and (ii) the corresponding vector of MSI ion intensities for the selected M ions. Ion targets were normalized per ion using the training-set statistics ($\log+1 \rightarrow$ z-score $\rightarrow$ clip to $[-2,2] \rightarrow$ rescale to $[0,1]$), and the same fixed transform was applied to evaluation data. To avoid empty background regions, we retained only positions with sufficient cellular signal (see Section 6.2.4.2.2). The

coculture slide was used to train the model; monoculture slides were reserved for evaluation of spectral prediction and spatial localization.

We also tested the reverse setting, training on monocultures and evaluating on the coculture. This failed consistently. Under weak supervision, the model can learn how the ion signals of the two cell types combine only if it sees mixed patches during training; a monoculture is simply a limiting case of such mixtures (an extreme mixing ratio, e.g. 100% WI-38 / 0% MCF-7), so a model trained on the coculture already encompasses the monoculture compositions and generalizes to them. The reverse setting-training on monocultures and testing on the coculture does not work, because monoculture-only training never observes mixtures, and the network cannot infer composition at test time. We confirm this asymmetry experimentally in Section 6.2.4.4.

6.2.4.2 MSI pre-processing

All MSI datasets (imzML) were first recalibrated using the adaptive pixelwise procedure of La Rocca et al.[27] to minimize mass error across the image. We then queried METASPACE[28] for putative molecular assignments (positive mode) against SwissLipids 2018-02-02 with a 3 ppm tolerance and allowed adducts $[\mathrm{M}+\mathrm{H}]^+$, $[\mathrm{M}+\mathrm{Na}]^+$, and $[\mathrm{M}+\mathrm{K}]^+$. In parallel, we explicitly inspected matrix-related peaks by matching theoretical masses of α-CHCA/CHCA adducts; a dimeric HCCA species $[2\cdot\mathrm{HCCA}-\mathrm{H}+2\mathrm{Na}]^+$ was retained as a matrix/background marker.

Because the fusion model benefits from cell-associated targets rather than matrix/noise, we applied a brief ion-curation pipeline prior to training (Figure S1):

1. **Spatial noise filtering.** Ions with low spatial coherence were removed using the spatial "chaos" score (Chapter 2); we required a score ≥ 0.7 on both monoculture slides.
2. **Cell correlation.** In manually annotated monoculture subregions, we computed, for each MSI pixel, the number of microscope pixels labeled as "cell" (MCF-7 ∪ WI-38) within a central crop (crop fraction $c = 0.3$; Methods). Ions with

Spearman ρ ≥ 0.5 to this cell-area proxy (across both monocultures) were retained.

3. **Annotation quality / relationships.** We prioritized SwissLipids annotations with < 2 ppm error and marked known adduct/isotope relationships by exact mass deltas; remaining series consistent with C_2H_4/H_2 spacing were labeled as propagated lipid homologs (see Supplementary S1 for counts).

From this curated pool, we selected a compact panel of seven ions for the main analyses, chosen for high signal-to-noise and complementary spatial patterns across the two cell types and representing various lipid classes. Six are protonated phospholipids and one is the matrix marker:

1. **PC(38:4)** — *m/z* 810.6016 (1.1 ppm)
2. **PC(36:4)** — *m/z* 782.5698 (0.5 ppm)
3. **SM(36:4)** — *m/z* 725.5600 (1.13 ppm)
4. **PE(24:4)** — *m/z* 572.3343 (0.6 ppm)
5. **PA(40:6)** — *m/z* 749.5124 (1.1 ppm)
6. **CerP(38:5)** — *m/z* 666.4858 (0.2 ppm)
7. **HCCA adduct** — *m/z* 423.0573 (2.3 ppm)

In the remainder of the chapter, we refer to each ion by its molecular family name for brevity. Target intensities were normalized as described in Section 6.2.3.6.2 (“Mass signal normalization”) before model training and evaluation.

From an interpretability standpoint, the model predicts all ions from a single shared 32-dimensional feature basis, with the final non-negative dense layer expressing each ion as an additive combination of these features, effectively a low-rank, NMF-like decomposition. Because MSI ion images are strongly co-localized and thus of low intrinsic dimensionality, the relevant capacity is the number of distinct spatial patterns rather than the number of ions M, so a fixed latent size can support many more ions than the seven-ion panel used here, with diminishing returns once added ions become

redundant. Predicting all ions jointly from this shared basis, rather than one model per ion, also exploits co-localization, shares statistical strength across correlated ions under weak supervision, amortizes the U-Net backbone over all outputs, and yields per-ion CAM maps in a common, comparable feature dictionary.

6.2.4.2.1 Microscope image pre-processing

For each MSI pixel, we paired the spectrum with a registered microscopy patch extracted from the pre-MALDI image (Section 6.2.3.4). Patch geometry was guided by the ablation footprint measured on the post-MALDI images: laser craters had an average diameter of ~90 µm (Figure 6.1c). To capture the full footprint and potential ion delocalization, we extracted a square region covering 3 × 3 raster steps (raster step 100 µm), centered on the ablation position—corresponding to ~330 × 330 µm in the microscope image (Figure 6.1b). Each patch was then resized to 192 × 192 pixels (RGB) and intensity-scaled to [0,1] for input to the network. This larger field of view ensures that cells contributing, at least in part, to the detected spectrum are included in the input patch.

Because a substantial fraction of the ITO surface lacks cells (Figure 6.1a), we restricted analysis to MSI positions with a clear cellular signal. Specifically, using the recalibrated MSI, we retained only pixels where PC(38:4) (hereafter PC (1)) exceeded 14 on the logarithmic scale. PC (1) is abundant in cell membranes of both lines and serves as a robust proxy for local cell density. Applying this filter yielded 11,557 spectra in the coculture, 9,636 in WI-38 monoculture, and 12,042 in MCF-7 monoculture. These positions were used to assemble the paired dataset of (microscopy patch, M-ion intensity vector) for training and evaluation. Because each MSI pixel corresponds to a single acquired mass spectrum at one ablation position, these counts are equivalent to the number of retained spectra, and each retained pixel yields exactly one paired example (microscope patch, M-ion intensity vector) for training and evaluation.

6.2.4.2.2 Ion delocalization on model performances

Ion delocalization denotes the contribution of material located outside the laser ablation footprint to the measured MS signal, e.g., due to matrix wetting/diffusion, solvent spreading, or plume effects during sample prep and acquisition.

In our data, this "material" corresponds to three microscopy-visible categories: MCF-7 cells, WI-38 cells, and background. To account for possible delocalization when linking a microscopy patch to its spectrum, we introduce a crop fraction c applied to the U-Net's final feature map (Section 6.2.3.6.4, *Tensor cropping*). The crop keeps only the central c fraction of the field of view before global averaging, thereby approximating the effective ablation footprint plus any local delocalization. Smaller c emphasizes the crater core; larger c includes more peripheral signal. In the following, we systematically vary c and evaluate how this choice impacts prediction error.To determine an appropriate crop fraction c, we trained separate models across a grid of c values and evaluated them on the monoculture slides. For each c, the model was trained on coculture data only and evaluated on manually annotated monoculture regions by computing the mean squared error (MSE) between the normalized predicted ion intensities and the corresponding experimentally measured, normalized MSI intensities for each patch–spectrum pair (Figure 6.3).In parallel, we fit a simple linear baseline using nonnegative least squares (NNLS)[29] to predict the normalized intensity of each selected *m/z*. Unlike the U-Net (which uses only images and no labels), this baseline uses manual annotations: for each patch and crop fraction c, we compute the counts of microscopy pixels labeled WI-38, MCF-7, and background within the cropped window, yielding the regressor vector

$$[A_{\mathrm{WI38}}, A_{\mathrm{MCF7}}, A_{\mathrm{bg}}].$$

For each ion m, we solve the NNLS problem (no intercept)

$$\hat{y}_m \approx w_{m,1} A_{\mathrm{WI38}} + w_{m,2} A_{\mathrm{MCF7}} + w_{m,3} A_{\mathrm{bg}}, w_{m,k} \geq 0,$$

and we train/evaluate this baseline across the same grid of crop sizes c as the U-Net. Using counts reflects the assumption that a larger labeled area can only increase the collected signal; because we fit a separate model for each c, the scale change with c is absorbed by the fitted weights (weights are not comparable across c, but MSE is). This baseline yields an interpretable "per-class ion yield" and serves as an optimistic upper bound for our weakly supervised model, since it exploits pixel-level class labels that the U-Net never sees. All targets are normalized as in Section 6.2.3.6 to ensure a fair comparison.

Figure 6.3 summarizes the MSE as a function of crop fraction c for both the U-Net and the NNLS baseline. For the NNLS baseline in particular, the c at which the MSE is minimal serves as a proxy for the effective extent of ion delocalization: a larger optimal c implies that accurate prediction requires including more peripheral pixels (greater delocalization), whereas a minimum near the physical crater fraction (~0.27) indicates limited delocalization.

The NNLS baseline (Figure 6.3; green = WI-38, red = MCF-7) shows that WI-38 signals are consistently harder to predict than MCF-7: MSE is higher for WI-38 across crop sizes. The optimal crop fraction c also varies by ion and cell type. For example, HCCA is best predicted at $c \approx 25\%$ for both cell types close to the physical crater fraction, consistent with a homogeneous matrix layer. Among cell-derived ions, PC (1) (PC(38:4)) exhibits the smallest effective delocalization with $c \approx 25\%$ for both lines, whereas PC (2) (PC(36:4)) favors $c \approx 28\%$ in MCF-7 and $c \approx 35\%$ in WI-38. For PE, SM, CerP, and PA, WI-38 generally requires larger crops (often $c > 38\%$) than MCF-7. Overall, the WI-38 dataset tends to prefer larger c, consistent with its more spread morphology on ITO (a greater exposed area corresponds to more opportunities for solvent-assisted diffusion and local delocalization). In contrast, the more spherical, weakly adherent MCF-7 cells likely limit contact with the matrix and show smaller effective delocalization. Ion-to-ion differences within a cell type are less pronounced but plausibly reflect matrix/adduct chemistry, local extraction/solubility, and ion-suppression effects.

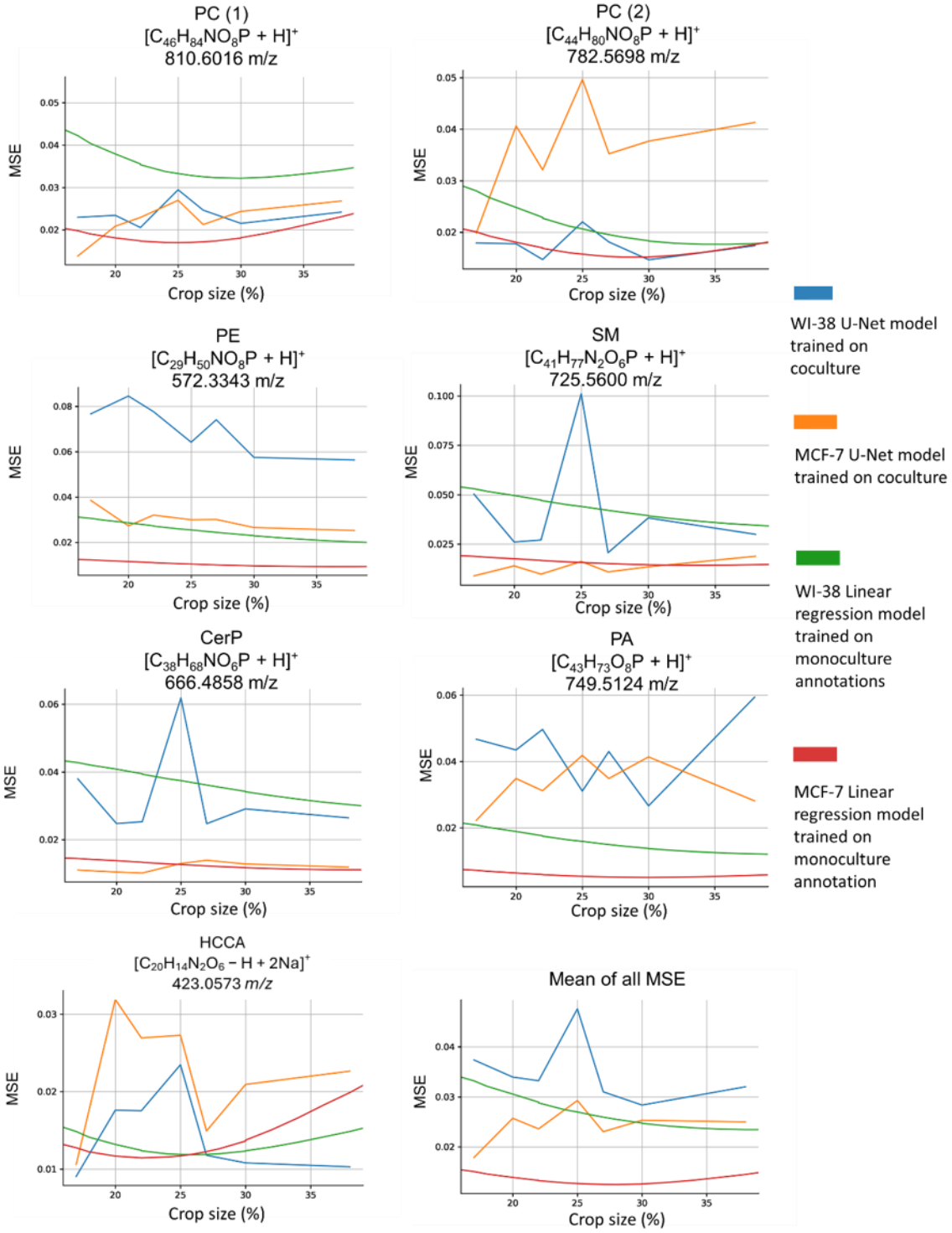

PC (1)
[C46H84NO8P + H]+
810.6016 m/z
PC (2)
[C44H80NO8P + H]+
782.5698 m/z
PE
[C29H50NO8P + H]+
572.3343 m/z
SM
[C41H77N2O6P + H]+
725.5600 m/z
CerP
[C38H68NO6P + H]+
666.4858 m/z
PA
[C43H73O8P + H]+
749.5124 m/z
HCCA
[C20H14N2O6 − H + 2Na]+
423.0573 m/z
Mean of all MSE
MSE
Crop size (%)
WI-38 U-Net model trained on coculture
MCF-7 U-Net model trained on coculture
WI-38 Linear regression model trained on monoculture annotations
MCF-7 Linear regression model trained on monoculture annotation

Figure 6.3. *Effect of crop fraction c on prediction error. Mean-squared error (MSE) between predicted and experimentally measured, normalized MSI intensities as a function of the crop fraction c (percentage of the microscopy patch retained before global averaging). Models are trained on the coculture and evaluated on annotated monoculture regions. The NNLS baseline is shown as green = WI-38 and red = MCF-7; the U-Net is shown blue = WI-38 and orange = MCF-7. Per-ion panels display MSE vs. c; the summary panel reports the mean MSE across ions. The vertical dashed line marks the physical crater fraction (~27%). Note that WI-38 typically favors larger c (greater effective delocalization), whereas MCF-7 often favors smaller c.*

Despite the visible domain shift between coculture and monoculture morphologies (Figure 6.1), the U-Net trained only on coculture generalizes well to monocultures. In Figure 6.3, for WI-38 the U-Net (blue) achieves lower MSE than the NNLS baseline (green) for all ions except PE and PA; for MCF-7, the U-Net (orange) outperforms NNLS (red) for CerP and SM. Our aim is not to "beat" the linear model. NNLS is an optimistic upper bound because it exploits manual pixel labels and is trained/evaluated directly on the monoculture set, yet these results indicate that learned spatial features (U-Net) can carry more predictive information than coarse class-area counts (NNLS), even under domain shift. Detailed spectral prediction accuracy is analyzed in the next section. *(For completeness, when both models are trained and evaluated on monocultures, the U-Net surpasses NNLS across all ions; results not shown.)*

The U-Net's MSE as a function of crop fraction c varies more than the NNLS baseline and exhibits no sharp minima (Figure 6.3; orange = MCF-7, blue = WI-38), which makes precise model selection by c less clear-cut. Nevertheless, two trends emerge from the mean MSE curves: for MCF-7, smaller crops perform slightly better (best around $c \approx 17\%$), whereas for WI-38, larger crops are favored (best around $c \approx 30\%$), with $c = 25\%$ being a local worst case for both. A plausible interpretation is morphological: compact, weakly adherent MCF-7 cells concentrate informative features near the crater core (smaller c), while the more spread WI-38 cells, and their greater effective exposure to solvent/matrix, benefit from including a wider neighborhood (larger c). For all subsequent analyses, we set $c = 30\%$ as a single operational choice: it lies near the physical crater fraction (~27%), is close to the U-Net's optimum for WI-38, and

remains competitive for MCF-7, providing a robust compromise across ions and cell types.

In summary, the U-Net trends suggest an optimal crop of ~17% for MCF-7 and ~30% for WI-38, with ~25% consistently underperforming, likely reflecting a mismatch between the physical crater and the spatial scale of informative morphology under mono- vs coculture conditions. For all subsequent analyses we adopt a single crop fraction of 30%. Using $c = 0.30$ ensures the model "sees" enough of the entire cell morphology, especially for spread WI-38 cells, while global average pooling prevents undue dominance of peripheral pixels. This single setting also avoids post-hoc tuning and keeps comparisons consistent across ions and slides.

6.2.4.3 Ion signal prediction

We evaluated per-patch spectrum prediction at MSI resolution (Figure 6.2d). A U-Net trained on the coculture with crop fraction $c = 0.30$ was applied to the two monoculture slides to predict, for each microscopy patch, the normalized intensities of the M selected ions. Targets were normalized as in Section 6.2.3.6 using the coculture (training) mean and standard deviation, and the same transform was applied to monoculture MSI for comparison. Figure 6.4 shows boxplots that compare the experimentally measured, normalized MSI intensities with the model predictions for each ion and each cell line.

We first examine the experimental MSI distributions (blue/orange in Figure 6.4). The HCCA peak shows comparable levels in MCF-7 and WI-38, which is expected since the mean cellular coverage is similar: matrix ions are typically suppressed over cells by competition for charge, so large differences in HCCA would have indicated unequal cell density. This inference is consistent with the manual annotations on the microscopy images, where the total number of cell-labeled pixels is similar across the two monocultures. Turning to lipids, we observe higher intensities in WI-38 for PC(36:4) (*PC (2)*), SM, and CerP, while PC(38:4) (*PC (1)*) is comparable between lines, and PA exhibits

a higher median in MCF-7. Because targets are normalized (Section 6.2.3.6) and overall cell coverage is comparable, these shifts reflect cell-type-associated biochemical differences rather than trivial differences in the amount of cellular material sampled.

Comparing U-Net predictions to the normalized MSI (Figure 6.4), the predicted distributions are pulled toward the center (≈0.5 on the $[0,1]$ scale), with underestimation of extremes, notably for PE and PA in WI-38. This dynamic-range compression is expected given (i) weak supervision from a homogeneous coculture (limited variation in cell-type proportions across patches), (ii) the use of MSE, which penalizes large errors and encourages mean predictions under uncertainty, and (iii) our normalization (z-score + clipping to $[-2,2]$ and then rescale to $[0,1]$), which makes 0.5 the mapped mean and attenuates tails. A modest domain shift between coculture and monoculture morphologies may further contribute to the shrinkage of high/low values.

Comparing predictions to MSI by cell line (Figure 6.4), the relative ordering of ion intensities between MCF-7 and WI-38 is largely preserved: when an ion is higher in one line in MSI, it is typically higher in the predictions as well. The main exception is PC(38:4) (PC (1)), for which the predicted mean is higher in WI-38 than MCF-7, unlike in the MSI. This mismatch is consistent with the limits of weak supervision in a homogeneous coculture: to deconvolve mixed patches, the model needs (i) variation in mixing ratios across patches and (ii) ions with line-specific enrichment. PC (1) shows little cell-type specificity in MSI, and the coculture provides limited variation in MCF-7/WI-38 proportions within patches; together, these factors leave the model with insufficient signal to assign PC (1) differentially. A modest domain shift in microscopy appearance may further contribute but is unlikely to fully explain this single-ion inversion.

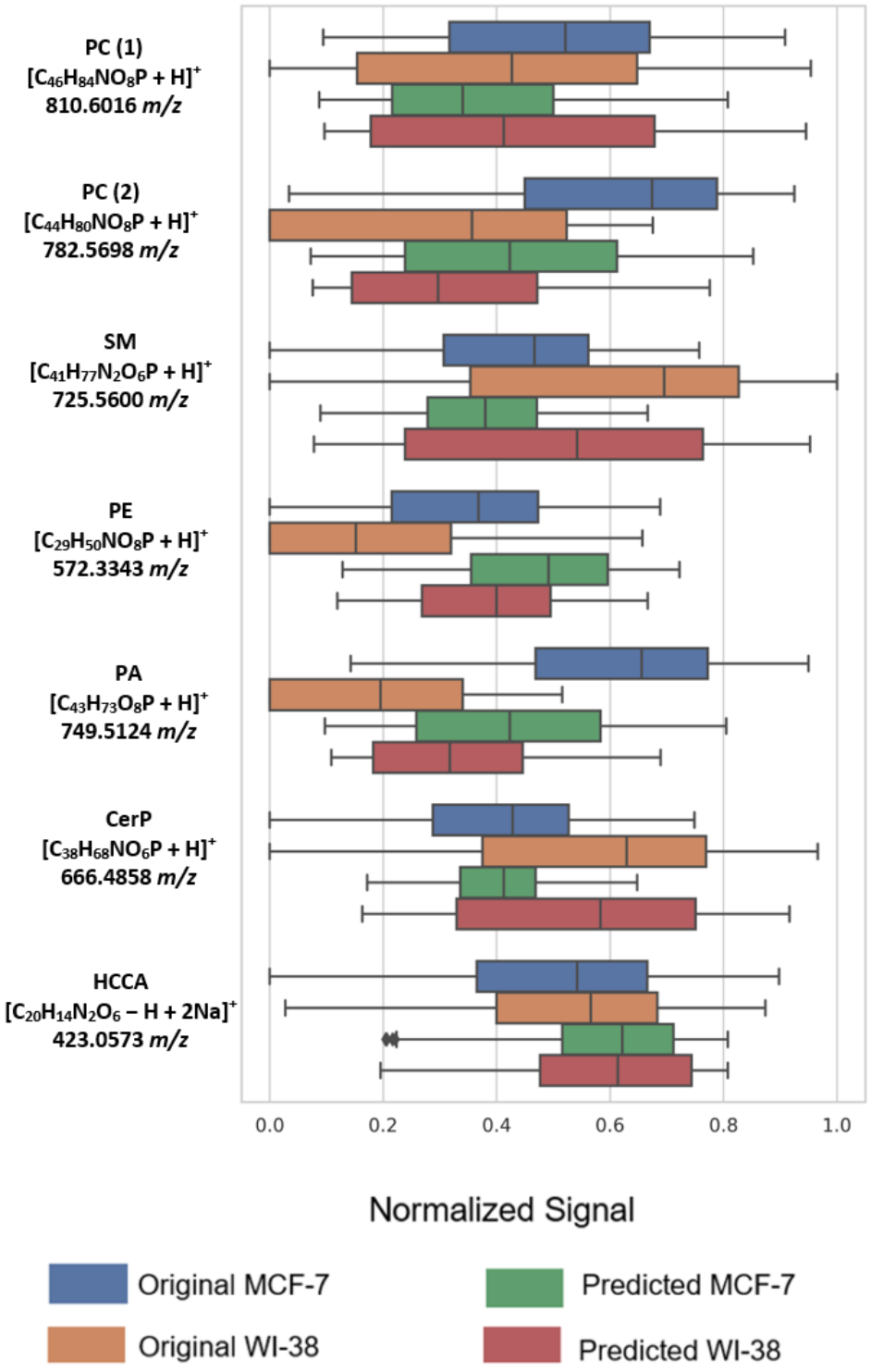

PC (1)
[C46H84NO8P + H]+
810.6016 m/z
PC (2)
[C44H80NO8P + H]+
782.5698 m/z
SM
[C41H77N2O6P + H]+
725.5600 m/z
PE
[C29H50NO8P + H]+
572.3343 m/z
PA
[C43H73O8P + H]+
749.5124 m/z
CerP
[C38H68NO6P + H]+
666.4858 m/z
HCCA
[C20H14N2O6 – H + 2Na]+
423.0573 m/z
0.0
0.2
0.4
0.6
0.8
1.0
Normalized Signal
Original MCF-7
Predicted MCF-7
Original WI-38
Predicted WI-38

Figure 6.4. *Predicted vs. measured ion intensities in monocultures. Box plots compare experimental MSI ("Original") and U-Net predictions ("Predicted") for each ion in MCF-7 and WI-38 monocultures. The model was trained on the coculture with crop fraction $c = 0.30$ and evaluated on monocultures. Boxes show the interquartile range (IQR) with the median line; whiskers extend to 1.5×IQR; points denote outliers.*

Overall, the model recovers the relative signal ratio between MCF-7 and WI-38 for most ions ,i.e., cell-type specificity is largely conserved, but it compresses dynamic range and shows a few ion-specific mismatches (e.g., PC(38:4)). Small microscopy acquisition differences (contrast/illumination) likely bias predictions, and weak supervision from a single coculture slide limits how well the network learns extremes. We already perform cross-slide evaluation (train on coculture A, test on monocultures B), which is appropriate; to strengthen robustness claims, we suggest (i) replicating across additional slides/batches (multiple cocultures and monocultures) to capture biological/technical variability, (ii) training multiple random seeds and reporting dispersion of MAE/MSE, (iii) exploring losses less sensitive to outliers and shrinkage (e.g., Huber/MAE, or per-ion loss reweighting) and (iv) applying color/style normalization or light augmentation to reduce microscope domain shift.

6.2.4.4 Comparing predicted ion location at the microscope spatial resolution

We evaluated whether the trained model can localize ions at microscope resolution on the coculture and both monocultures. For each input microscopy patch ($192 \times 192 \times 3$; field of view ≈ 330×330 µm, i.e. ~1.7 µm/px), class activation map (CAM) (Methods) produce an M-channel fused image with the same spatial size (192×192), yielding one map per ion.

To obtain whole-slide predictions, we tiled the pre-MALDI image into overlapping patches with a 33 µm (~19 px) overlap, computed CAMs for each tile, and formed a seamless mosaic by averaging overlapping pixels for each ion. Because the CAM is read from the full, uncropped U-Net feature tensor (Section 6.2.3.6.6), each tile yields enrichment over its whole extent, including the periphery outside the central crop the

prediction relied on. This off-center signal is trustworthy because the U-Net feature maps vary smoothly, with no systematic difference between the crop region and its surroundings, so it is a smooth extrapolation of the same learned relationship. The overlap is retained only to avoid seams at tile borders: each shared pixel takes the mean of the tiles covering it, a mild smoothing that suppresses edge artifacts without changing the relative-enrichment interpretation.

The CAM for a given ion indicates which regions of the input patch contributed most to the model's predicted intensity for that ion. It should therefore be read as a relative enrichment pattern, a spatial description of the morphology that co-varies with the ion signal, not as a localization probability or a per-pixel reconstruction of ion abundance. In short, it shows which spatial pattern correlates with the ion signal, but it is not an ion image.

Because the network is trained against per-ion normalized intensities (Section 6.2.3.6.2), its predictions, and hence the CAM enrichment, lie on a common normalized scale that can be compared across patches, regions, and slides. This cross-slide comparability also assumes that the new images resemble those seen during training: when a slide differs too strongly in appearance (e.g. contrast, illumination, or morphology), the network operates outside its learned distribution and its CAM enrichment is no longer reliably comparable.

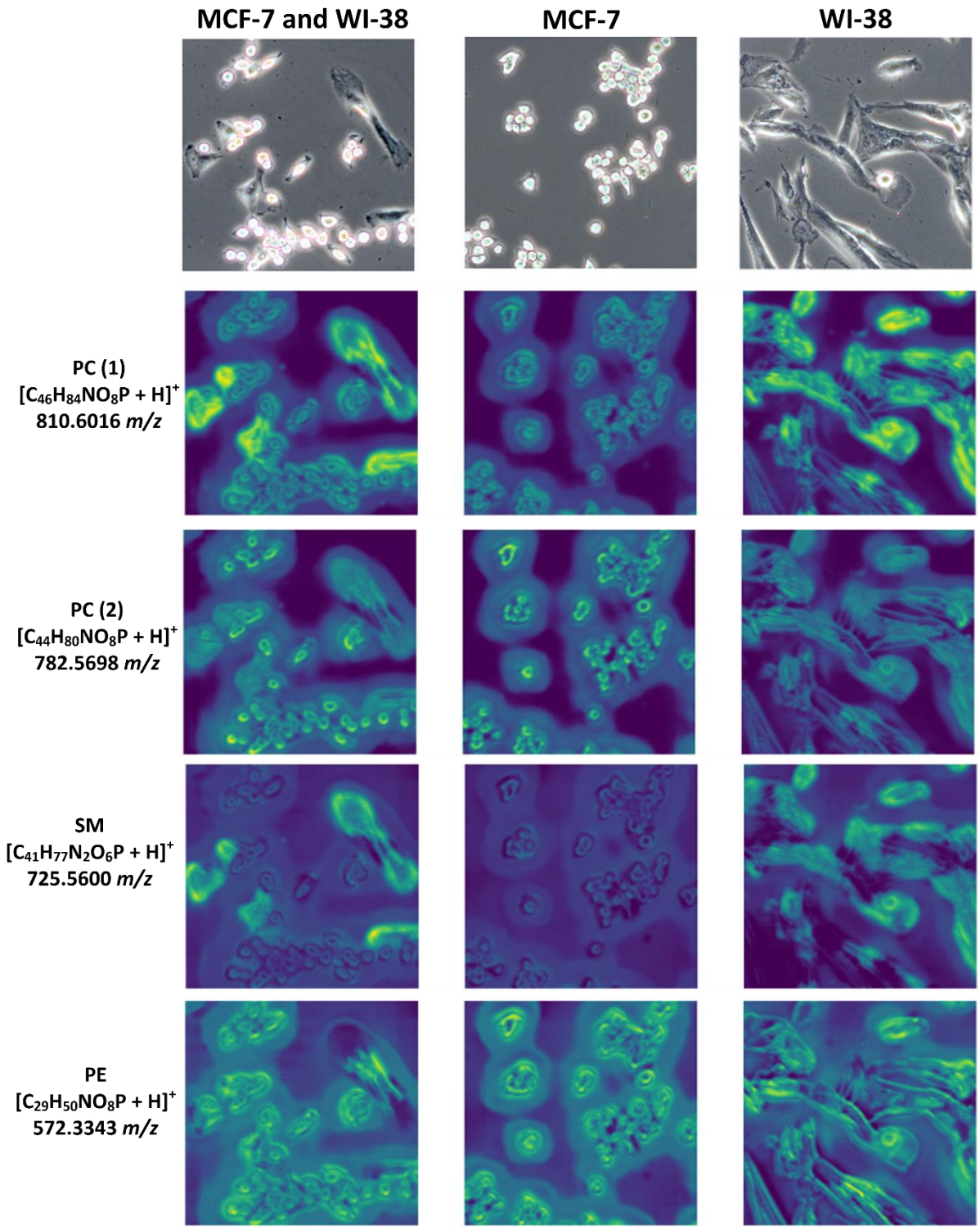
CAM values can also be produced in regions not sampled by MSI, which is useful for exploration but should be interpreted cautiously. A practical advantage of this design is computational: because the network outputs full-resolution maps per patch, whole-slide inference is fast, in contrast to CNN regressors that emit a single scalar per patch and must be slid pixel-by-pixel across the image.

This tiling and averaging procedure yields whole-slide fused ion maps at microscope resolution, including areas not sampled by MSI (extrapolated from morphology). Figure 6.5 presents representative fields from the coculture and both monocultures, showing the predicted per-ion localization patterns.

The CAM maps largely mirror the cell-type trends seen in Figure 6.4: SM and CerP are enriched over WI-38 morphology, while PA shows stronger enrichment over MCF-7. For PC (1), CAM also shows higher signal over WI-38, which is consistent with the model's spectral prediction but not with the MSI ground truth (where PC (1) is similar across lines). Thus, for most ions the spatial localization agrees with spectrum-level specificity; PC (1) remains the main exception. As expected, HCCA is depleted in cells and elevated in the background. A modest residual HCCA signal within WI-38 cells is often observed, consistent with their more spread morphology on glass (greater matrix-contact area; Figure 6.1).

The pre-MALDI images were acquired in phase contrast, which produces a characteristic bright/dark halo around high-refractive–index structures (cell edges). This halo is a known optical artifact: phase differences are converted to intensity and fringes are amplified at boundaries. Because the halo is highly correlated with cell presence, the network's CAM sometimes attributes part of an ion's localization to the halo ring rather than to the cell body itself (Figure 6.5). In practice the halo signal is secondary to morphology but still visible. This artifact should be kept in mind when interpreting fine, ring-like attributions; simple mitigations include using light contrast/illumination augmentation to reduce halo reliance.

Apparent over-detection/specificity of SM or CerP in WI-38 should not be taken as direct evidence of overexpression. WI-38's more spread morphology increases contact with matrix/solvent and can alter extraction/ionization efficiency, shifting detected profiles independent of intracellular abundance. Without orthogonal controls (e.g., LC/MS confirmation,), we cannot fully disentangle biology from ionization geometry. Accordingly, CAM maps are best interpreted as relative enrichment conditioned on imaging context, not absolute per-pixel concentration.


MCF-7 and WI-38
MCF-7
WI-38
PC (1)
$[C_{46}H_{84}NO_8P + H]^+$
810.6016 *m/z*
PC (2)
$[C_{44}H_{80}NO_8P + H]^+$
782.5698 *m/z*
SM
$[C_{41}H_{77}N_2O_6P + H]^+$
725.5600 *m/z*
PE
$[C_{29}H_{50}NO_8P + H]^+$
572.3343 *m/z*

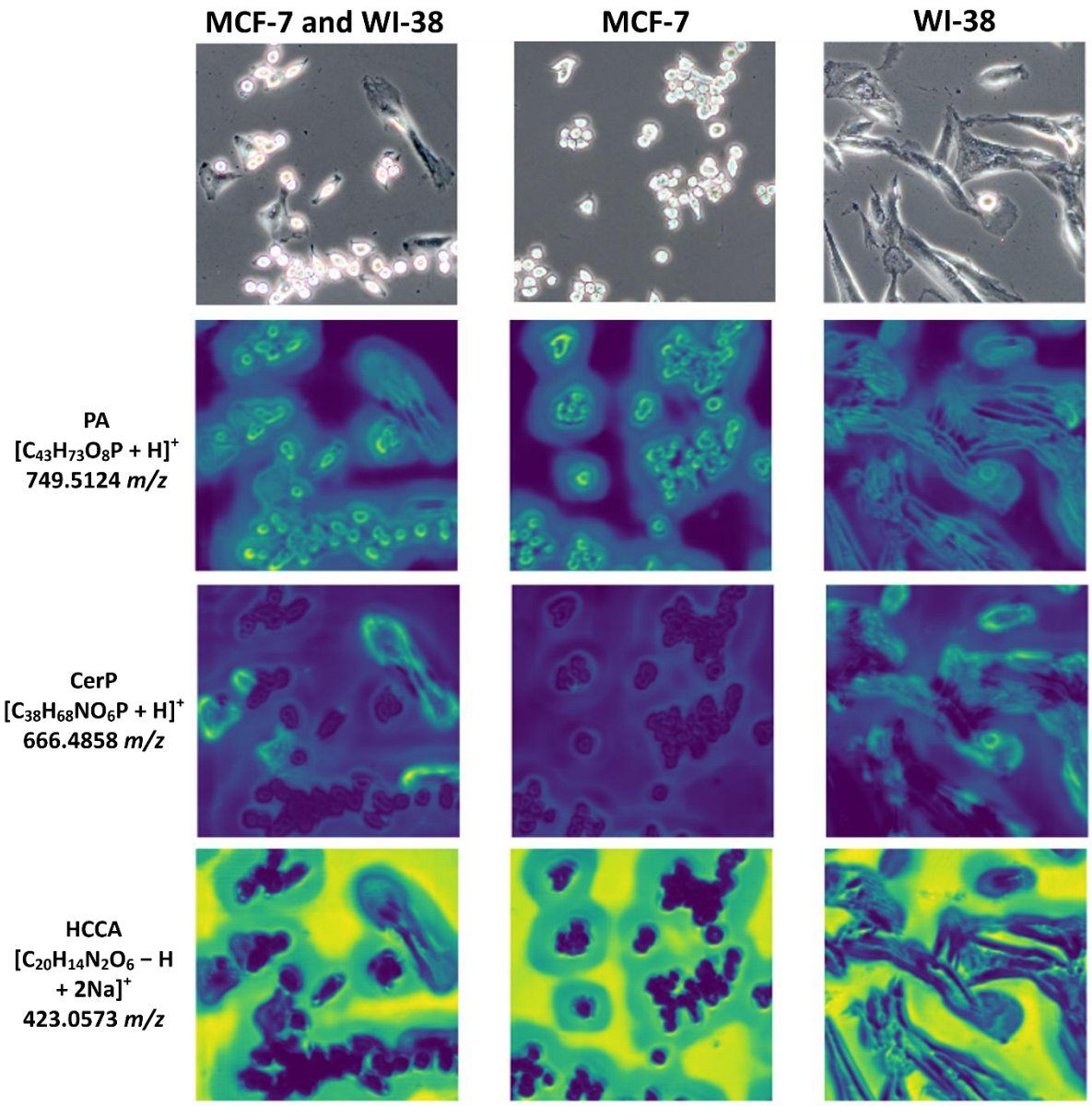


Figure 6.5. *Fused ion maps (CAM) at microscope resolution. Representative fields from MCF-7, WI-38, and coculture slides. Each field spans ~3×3 MSI pixels (raster 100 µm; field of view ≈ 330×330 µm, ~1.7 µm/px). First row: microscope image. Other rows: CAM-based fused maps for each ion, shown as heat maps (dark blue = low, yellow = high). Maps are displayed as relative enrichment (per-ion, normalized for visualization). Spatial patterns mirror spectrum-level specificity (e.g., SM/CerP enrichment over WI-38); HCCA is depleted over cells and elevated in background. Overlapping tiles were averaged to form seamless mosaics.*

The model learns robust spatial features from microscopy alone and links them to ion signals without any pixel-level annotations or handcrafted preprocessing (weak supervision). Using CAM, it recovers per-ion localization at microscope resolution (≈1.7 µm/px) from MSI acquired at 100 µm sampling, enabling high-resolution fused maps over entire slides. For most ions, the cell-type enrichment pattern observed in MSI is preserved in the fused images, and the model generalizes across acquisition differences (coculture to monocultures). CAM values should be interpreted as relative enrichment rather than absolute per-pixel concentration, but they provide consistent, biologically meaningful spatial cues suitable for downstream visualization and segmentation.

6.2.4.5 Microscopy image segmentation based on fused images

We evaluated whether fused ion maps can improve microscopy segmentation. MSI provides rich per-pixel spectral features, but its native spatial resolution is too coarse for cell-level labels. Our fusion model projects MSI information onto the microscope grid (Section 6.2.3.6), yielding per-pixel ion features at ~1.7 µm. We then ask: given only these fused features, can a simple linear classifier segment MCF-7, WI-38, and background in the coculture image more accurately than using RGB alone?

For pixel classification, we used multinomial logistic regression[30], a linear classifier that maps a feature vector to class probabilities via a softmax over a linear score (with L2 regularization). Concretely, each pixel's input is the vector of normalized fused ion intensities; the model outputs probabilities for background, MCF-7, and WI-38. We use a standard configuration (multinomial, L2, solver *lbfgs*, class-weighting to mitigate imbalance). Logistic regression does not model spatial context, which makes it a weak choice for raw RGB segmentation. Here, however, the fused spectral features provide strong per-pixel biochemical contrast, so a simple linear decision boundary is sufficient to test whether MSI-derived information improves segmentation over RGB alone.

We derive pixel labels from the manual masks described in Section 6.2.3.5, assigning each microscopy pixel of the monoculture to one of three classes: background, MCF-7, or WI-38 (with no cell class overlap) . For every labeled pixel, the input feature vector is the M-dimensional fused ion intensity produced by the coculture-trained U-Net at that location. The logistic regression model is trained on the two monoculture slides using these (feature, label) pairs, learning to map fused ion vectors to the three classes. Quantitative performance is computed on held-out monoculture pixels; the trained model is then applied to the coculture for qualitative assessment.

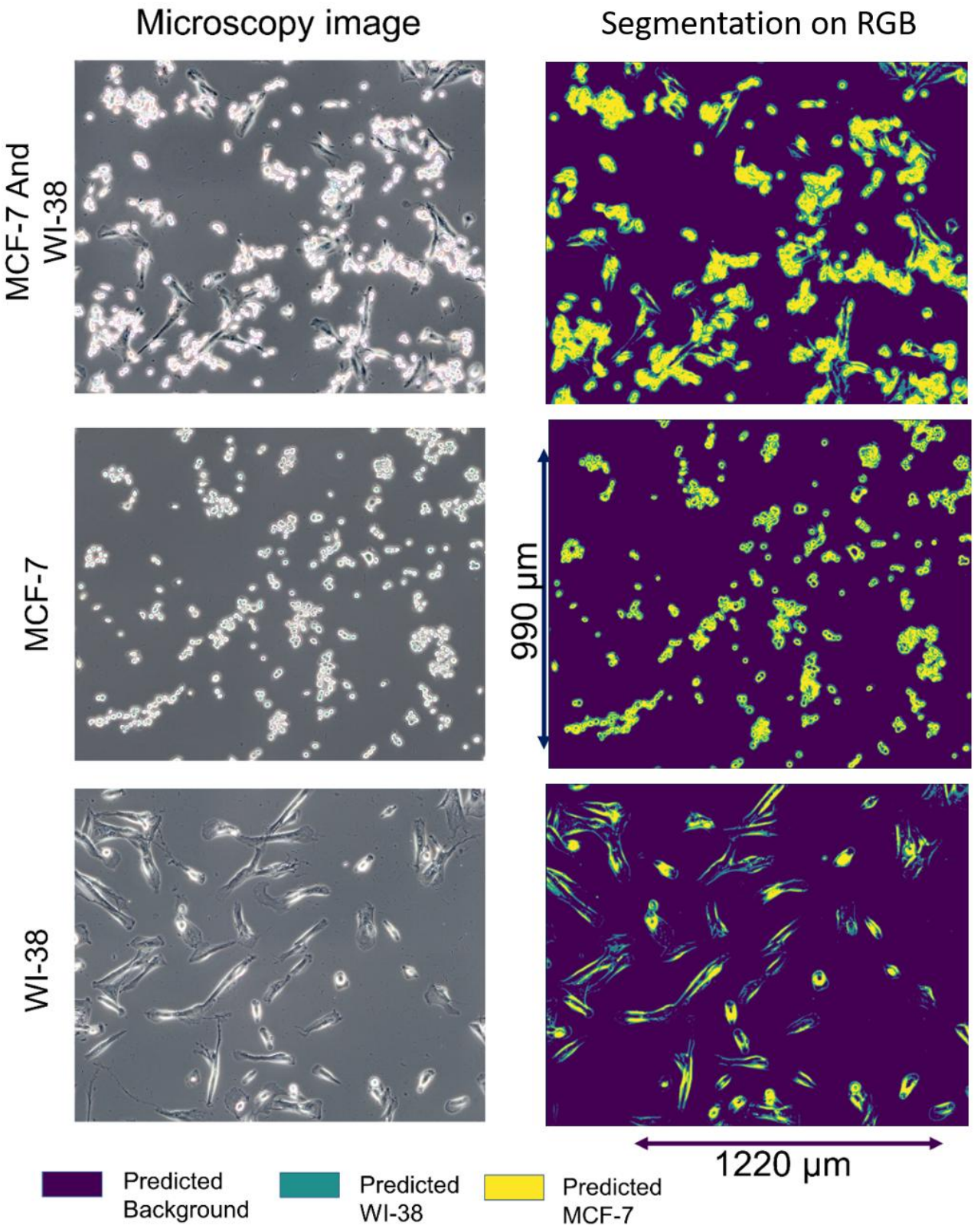
Manual annotations were drawn only on the monoculture slides, where each pixel belongs to a single class (background and one cell type), and no mixing between the two cell types is present. Quantitative evaluation of the segmentation is therefore restricted to monocultures. The coculture does, however, contain regions of potential overlap between the two cell types, precisely the situation the segmentation classifier never encountered during training, since it was trained on monoculture pixels only. In such regions, a single-class assignment is not necessarily meaningful, and it may be more appropriate to treat the pixels as mixed or ambiguous rather than force them into one cell type, for example, by examining the predicted class probabilities (a low or evenly split maximum probability indicating an uncertain, likely mixed pixel).

Each pixel is represented by its M-dimensional vector of normalized fused ion intensities (from the coculture-trained U-Net), and the classifier outputs one of the three classes. To build the dataset, we used a block-wise split on each monoculture slide to limit spatial autocorrelation: the image was partitioned into non-overlapping tiles, a ~1% random sample of pixels from the training tiles formed the training set, and all pixels from the held-out tiles formed the test set (fixed random seed for reproducibility). After fitting a multinomial logistic regression model, we predicted test pixels and evaluated performance using balanced accuracy, which gives equal weight to background, MCF-7, and WI-38 regardless of class imbalance.

Balanced Accuracy is a performance metric used in machine learning classification tasks, particularly when the dataset is imbalanced, where one class has significantly more instances than another. It is designed to give equal weight to the performance of each class, regardless of their proportion in the data (see Chapter 5). To quantify the benefit of fused ion features over raw brightfield microscope imagery, we trained a matched logistic regression baseline on RGB only: each pixel's feature vector is its 3-channel intensity from the pre-MALDI microscope image (per-channel standardized on the training tiles), with the same train/test tiles, class weighting, regularization, and solver as the fused-feature model. This ensures a direct comparison by isolating the impact of MSI-derived fused features on segmentation performance.

The model trained on fused ion features achieved a balanced accuracy of 0.80, while the matched RGB-only logistic regression reached 0.55 on the same data. This gap confirms that the fused maps encode biochemically meaningful cell-type information, enabling even a simple linear classifier to distinguish MCF-7, WI-38, and background without spatial context. A direct MSI baseline is not applicable here because the 100 µm MSI sampling cannot resolve individual cells; only the fused images provide per-pixel features suitable for segmentation.

Figure 6.6 shows predictions from the monoculture-trained logistic regression applied to both monocultures and the coculture. In all three cases, most pixels are correctly assigned to MCF-7, WI-38, or background. Residual errors are localized and interpretable: in the MCF-7 monoculture, mislabels occur where cells are unusually spread on the ITO, mimicking WI-38 morphology; in the WI-38 monoculture, errors arise where cells appear brighter and more compact, resembling MCF-7. In the coculture, MCF-7 predictions tend to be slightly dilated relative to monoculture, likely because the phase-contrast halo is stronger in the coculture acquisition, inflating per-pixel fused features near cell boundaries. Note that segmentation inherits any biases present in the fused maps (e.g., halo influence), so these effects are expected and visible at the margins.


Microscopy image
Segmentation on RGB
MCF-7 And WI-38
MCF-7
WI-38
990 µm
1220 µm
Predicted Background
Predicted WI-38
Predicted MCF-7

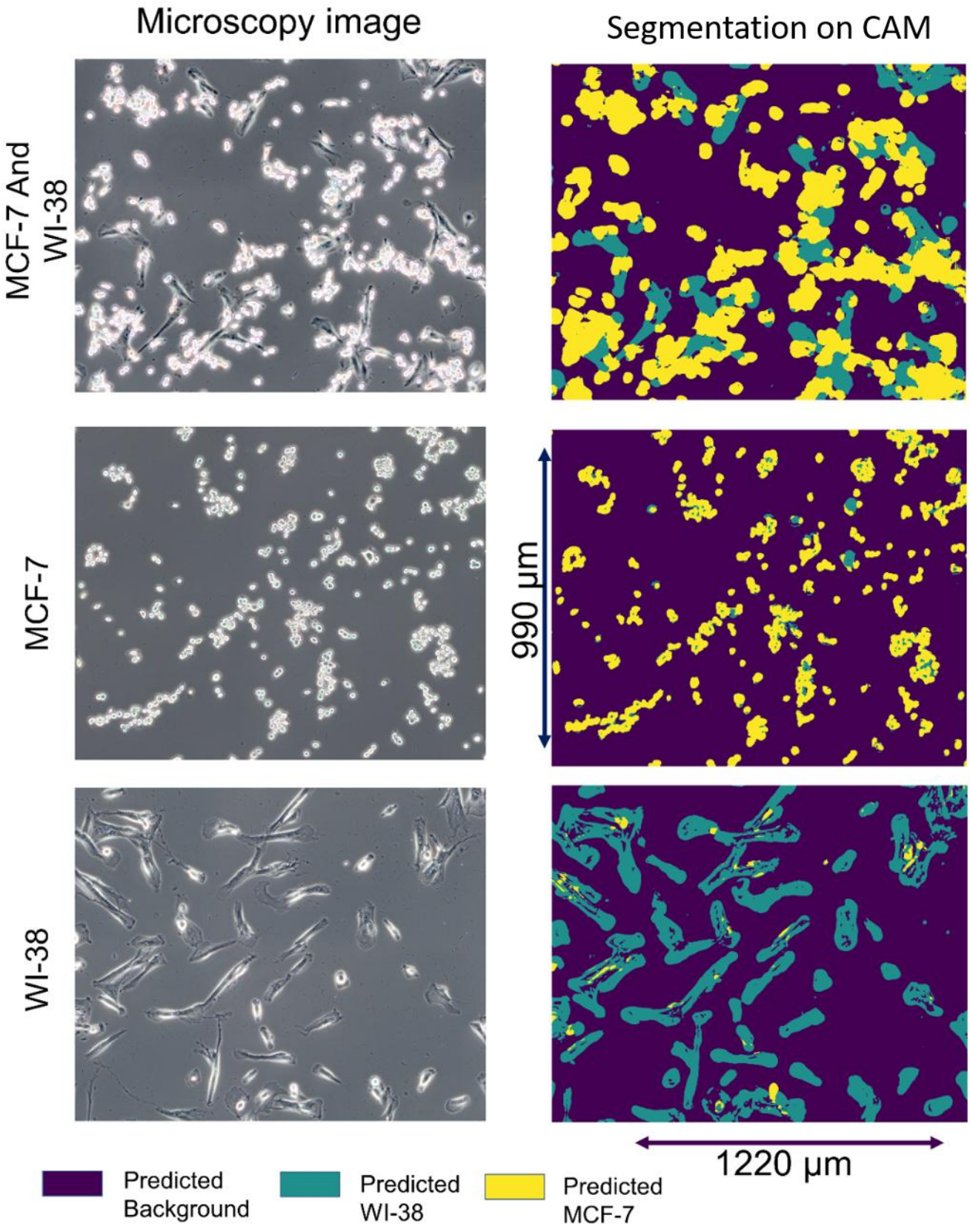

Microscopy image
Segmentation on CAM
MCF-7 And WI-38
MCF-7
WI-38
990 µm
1220 µm
Predicted Background
Predicted WI-38
Predicted MCF-7

Figure 6.6. *Pixel-wise segmentation from RGB values (top) and fused ion features (down). Representative fields from MCF-7 monoculture, WI-38 monoculture, and coculture showing predictions of a multinomial logistic regression trained on fused ion maps (features = per-pixel normalized CAM intensities from the coculture-trained U-Net) or RGB values. Pixels are colored by predicted class (MCF-7, WI-38, background; see legend). Training was done on monoculture tiles; shown monoculture panels use held-out, non-overlapping tiles for testing, and the coculture panel is qualitative (no ground truth labels). Fields are ~330×330 µm (≈ 3×3 MSI pixels at 100 µm raster), ~1.7 µm/px.*

Fused ion maps substantially improve microscopy segmentation: with the same train/test protocol, a simple multinomial logistic regression reaches 0.80 balanced accuracy on fused features vs 0.55 on RGB alone. This demonstrates that projecting MSI signal onto the microscope grid provides biochemically informative per-pixel features that complement morphology. While our classifier is intentionally context-free (no spatial modeling), the results highlight the value of combining spectral (from MSI) and spatial (from microscopy) information; modest spatial post-processing or context-aware models would likely yield additional gains.

6.2.5 Conclusion and Perspectives

In this chapter, we introduced a U-Net + CAM fusion approach that projects MSI information onto the microscope grid, producing microscope-resolution per-ion maps from 100 µm MSI. We evaluated the method on an artificial MCF-7/WI-38 coculture with companion monocultures, assessing (i) patch-level spectrum prediction at MSI resolution and (ii) per-ion spatial localization at ~1.7 µm/px.

The method shows tolerance to ion delocalization: sweeping the crop fraction c identified operating points near the physical crater fraction (~0.27–0.30), with larger c preferred for spread WI-38 cells, consistent with greater effective delocalization, while fused maps preserved the expected cell-type enrichments. The coculture-trained model also generalized across acquisition differences to monocultures, though residual bias from microscopy contrast/illumination and phase-contrast halos remains. A key limitation is dynamic-range compression: extremes are under-predicted, likely arising

from the combination of weak supervision (a single coculture with limited variation in mixing ratios), the target normalization, and the MSE loss. The log transform, per-ion z-scoring, clipping, and rescaling, together with the MSE loss, are deliberate design choices here rather than fundamental requirements. They are motivated by the objective pursued in this work, balanced, outlier-robust training for relative ion localization, with targets matched to the sigmoid output range, and not by precise abundance estimation. These same choices are partly responsible for the dynamic-range compression discussed in Section 6.2.4.3, and they are natural candidates for revision in future work: lighter or no clipping, a loss less sensitive to outliers and shrinkage (e.g., Huber or MAE), per-ion loss reweighting, or a quantile/rank normalization could reduce this compression should quantitative prediction become the goal. Accurate deconvolution further requires sufficient variation in cell-type composition across patches; when these are weak (e.g., PC(38:4)), the model struggles to separate contributions. If the goal is precise intensity prediction, additional work on loss design, normalization, and supervision diversity is needed.

A limitation of the present cropping scheme is that it applies a uniform (top-hat) weight to every retained location before pooling. This does not reflect the physics of the measurement: ionization yield is typically highest near the crater center, and delocalized contributions fall off with distance from it, so peripheral pixels should count less than central ones rather than equally. A more principled alternative is a soft radial weighting, e.g., a Gaussian centered on the crater, applied before pooling. This would down-weight peripheral contributions in line with the smooth radial decay of the effective sampling sensitivity, provide a closer approximation of the MSI sampling/point-spread function, and, being differentiable in its width, allow that width to be learned from the data rather than fixed through a grid search over c.

By fusing MSI with microscopy, the model decomposes mixed MSI spectra into per-pixel ion enrichment maps aligned with cellular morphology (cells vs background). This directly mitigates the spatial-resolution mismatch in MSI (100 µm sampling) by

providing microscope-scale (~1.7 µm/px) fused maps. A practical advantage of our architecture is that U-Net+CAM produces a full 192×192 map per patch in one pass, so whole-slide outputs are obtained efficiently by tiling and overlap-averaging, rather than sliding a scalar regressor.

For downstream analysis, these fused per-pixel features enable simple segmentation: a multinomial logistic regression trained on a small set of manual masks on monocultures achieves 0.80 balanced accuracy, compared with 0.55 using RGB alone (Section 6.2.4.5). Importantly, fusion training itself does not require pixel-level molecular annotations (weak supervision from spectra only). In practice, the fused maps can serve as pseudo-labels/features to pre-train or bootstrap microscopy segmentation with minimal manual annotation, after which lightweight classifiers (or context-aware models) can refine full-slide segmentations.

Our current model uses phase-contrast/brightfield microscopy as the sole visual input; contrast differences and morphology provide enough cues to link image patterns to MSI signals. Richer spatial cues could be obtained with histological stains (e.g., H&E), and evaluating the fusion on stained tissues is a natural next step. However, fair comparison on tissue is harder: pixel/instance annotations are scarce, heterogeneity is higher, and stain variability introduces domain shift. Two practical extensions can mitigate current limits: (i) add the post-MALDI brightfield image as an extra input channel to inform the network about crater geometry, residual material, and non-ionized regions, reducing reliance on a fixed crop fraction c (the model can learn an effective footprint); and (ii) apply style/contrast augmentation and, for H&E, simple stain normalization/deconvolution to reduce acquisition and staining bias. Together, these changes should improve localization fidelity and lessen the need for hand-tuned crop parameters while making the method more robust across imaging setups and sample types.

In practice, the fused maps can also serve as pseudo-labels/features to pre-train microscopy segmentation backbones; however, to make a definitive claim we need systematic comparisons and broader data. Concretely, we will (i) benchmark against representative MSI–microscopy fusion baselines (e.g., PLS/NNLS fusion (Caprioli/Van de Plas), patch-based super-resolution methods, and deep backbones such as ResNet/FPN/UNet variants trained to predict spectra or per-ion maps) under identical preprocessing, registration, normalization, and splits; (ii) evaluate on additional slides/batches (multiple cocultures and monocultures, and, where feasible, stained tissues), with tile-held-out and leave-one-slide-out protocols; (iii) report reproducibility over multiple random seeds and ion-panel selections (mean ± SD of MAE/MSE, rank correlations, calibration slopes), since performance can vary with initialization and target choice.

Because delocalization and matrix deposition affect fusion, we will explicitly compare deposition strategies (e.g., spray vs sublimation) and incorporate the post-MALDI brightfield as an additional input channel. Finally, to test whether fused maps faithfully reflect sub-cellular spatial patterns, we will compare predictions from low-resolution MSI (100 µm) against higher-spatial-resolution MSI acquisitions on the same samples. Together, these steps will determine whether the proposed U-Net-CAM approach, whose full-patch outputs enable fast whole-slide prediction, retains its advantages over existing fusion methods across datasets, seeds, and experimental conditions.

REFERENCES:

(1) Kune, C.; McCann, A.; Raphaël, L. R.; Arias, A. A.; Tiquet, M.; Van Kruining, D.; Martinez, P. M.; Ongena, M.; Eppe, G.; Quinton, L.; Far, J.; De Pauw, E. Rapid Visualization of Chemically Related Compounds Using Kendrick Mass Defect As a Filter in Mass Spectrometry Imaging. *Anal. Chem.* **2019**, *91* (20), 13112–13118. https://doi.org/10.1021/acs.analchem.9b03333.

(2) Buchberger, A. R.; DeLaney, K.; Johnson, J.; Li, L. Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights. *Anal. Chem.* **2018**, *90* (1), 240–265. https://doi.org/10.1021/acs.analchem.7b04733.

(3) Seeley, E. H.; Caprioli, R. M. MALDI Imaging Mass Spectrometry of Human Tissue: Method Challenges and Clinical Perspectives. *Trends Biotechnol.* **2011**, *29* (3), 136–143. https://doi.org/10.1016/j.tibtech.2010.12.002.

(4) Holzlechner, M.; Eugenin, E.; Prideaux, B. Mass Spectrometry Imaging to Detect Lipid Biomarkers and Disease Signatures in Cancer. *CANCER Rep.* **2019**, *2* (6). https://doi.org/10.1002/cnr2.1229.

(5) Neumann, E. K.; Ellis, J. F.; Triplett, A. E.; Rubakhin, S. S.; Sweedler, J. V. Lipid Analysis of 30 000 Individual Rodent Cerebellar Cells Using High-Resolution Mass Spectrometry. *Anal. Chem.* **2019**, *91* (12), 7871–7878. https://doi.org/10.1021/acs.analchem.9b01689.

(6) Rappez, L.; Stadler, M.; Triana, S.; Gathungu, R. M.; Ovchinnikova, K.; Phapale, P.; Heikenwalder, M.; Alexandrov, T. SpaceM Reveals Metabolic States of Single Cells. *Nat. Methods* **2021**, *18* (7), 799–805. https://doi.org/10.1038/s41592-021-01198-0.

(7) Tian, X.; Xie, B.; Zou, Z.; Jiao, Y.; Lin, L.-E.; Chen, C.-L.; Hsu, C.-C.; Peng, J.; Yang, Z. Multimodal Imaging of Amyloid Plaques: Fusion of the Single-Probe Mass Spectrometry Image and Fluorescence Microscopy Image. *Anal. Chem.* **2019**, *91* (20), 12882–12889. https://doi.org/10.1021/acs.analchem.9b02792.

(8) Watrous, J. D.; Dorrestein, P. C. Imaging Mass Spectrometry in Microbiology. *Nat. Rev. Microbiol.* **2011**, *9* (9), 683–694. https://doi.org/10.1038/nrmicro2634.

(9) Balluff, B.; Heeren, R. M. A.; Race, A. M. An Overview of Image Registration for Aligning Mass Spectrometry Imaging with Clinically Relevant Imaging Modalities. *J. Mass Spectrom. Adv. Clin. Lab* **2022**, *23*, 26–38. https://doi.org/10.1016/j.jmsacl.2021.12.006.

(10) Race, A. M.; Sutton, D.; Hamm, G.; Maglennon, G.; Morton, J. P.; Strittmatter, N.; Campbell, A.; Sansom, O. J.; Wang, Y.; Barry, S. T.; Takáts, Z.; Goodwin, R. J. A.; Bunch, J. Deep Learning-Based Annotation Transfer between Molecular Imaging Modalities: An Automated Workflow for Multimodal Data Integration. *Anal. Chem.* **2021**, *93* (6), 3061–3071. https://doi.org/10.1021/acs.analchem.0c02726.

(11) Falk, T.; Mai, D.; Bensch, R.; Çiçek, Ö.; Abdulkadir, A.; Marrakchi, Y.; Böhm, A.; Deubner, J.; Jäckel, Z.; Seiwald, K.; Dovzhenko, A.; Tietz, O.; Dal Bosco, C.; Walsh, S.; Saltukoglu, D.; Tay, T. L.; Prinz, M.; Palme, K.; Simons, M.; Diester, I.; Brox, T.; Ronneberger, O. U-Net: Deep Learning for Cell Counting, Detection, and Morphometry. *Nat. Methods* **2019**, *16* (1), 67–70. https://doi.org/10.1038/s41592-018-0261-2.

(12) Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. **2014**. https://doi.org/10.48550/ARXIV.1411.4038.

(13) Bychkov, D.; Linder, N.; Turkki, R.; Nordling, S.; Kovanen, P. E.; Verrill, C.; Walliander, M.; Lundin, M.; Haglund, C.; Lundin, J. Deep Learning Based Tissue Analysis Predicts Outcome in Colorectal Cancer. *Sci. Rep.* **2018**, *8* (1), 3395. https://doi.org/10.1038/s41598-018-21758-3.

(14) Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*; Navab, N., Hornegger, J., Wells, W. M., Frangi, A. F., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2015; Vol. 9351, pp 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

(15) Tideman, L. E. M.; Migas, L. G.; Djambazova, K. V.; Patterson, N. H.; Caprioli, R. M.; Spraggins, J. M.; Van de Plas, R. Automated Biomarker Candidate Discovery in Imaging Mass Spectrometry Data through Spatially Localized Shapley Additive Explanations. *Anal. Chim. Acta* **2021**, *1177*, 338522. https://doi.org/10.1016/j.aca.2021.338522.

(16) Hanselmann, M.; Köthe, U.; Kirchner, M.; Renard, B. Y.; Amstalden, E. R.; Glunde, K.; Heeren, R. M. A.; Hamprecht, F. A. Toward Digital Staining Using Imaging Mass Spectrometry and Random Forests. *J. Proteome Res.* **2009**, *8* (7), 3558–3567. https://doi.org/10.1021/pr900253y.

(17) Meding, S.; Nitsche, U.; Balluff, B.; Elsner, M.; Rauser, S.; Schöne, C.; Nipp, M.; Maak, M.; Feith, M.; Ebert, M. P.; Friess, H.; Langer, R.; Höfler, H.; Zitzelsberger, H.; Rosenberg, R.; Walch, A. Tumor Classification of Six Common Cancer Types Based on Proteomic Profiling by MALDI Imaging. *J. Proteome Res.* **2012**, *11* (3), 1996–2003. https://doi.org/10.1021/pr200784p.

(18) Verbeeck, N.; Caprioli, R. M.; Van de Plas, R. Unsupervised Machine Learning for Exploratory Data Analysis in Imaging Mass Spectrometry. *Mass Spectrom. Rev.* **2020**, *39* (3), 245–291. https://doi.org/10.1002/mas.21602.

(19) Ščupáková, K.; Balluff, B.; Tressler, C.; Adelaja, T.; Heeren, R. M. A.; Glunde, K.; Ertaylan, G. Cellular Resolution in Clinical MALDI Mass Spectrometry Imaging: The Latest Advancements and Current Challenges. *Clin. Chem. Lab. Med. CCLM* **2020**, *58* (6), 914–929. https://doi.org/10.1515/cclm-2019-0858.

(20) Porta Siegel, T.; Hamm, G.; Bunch, J.; Cappell, J.; Fletcher, J. S.; Schwamborn, K. Mass Spectrometry Imaging and Integration with Other Imaging Modalities for Greater Molecular Understanding of Biological Tissues. *Mol. Imaging Biol.* **2018**, *20* (6), 888–901. https://doi.org/10.1007/s11307-018-1267-y.

(21) Verbeeck, N.; Yang, J.; De Moor, B.; Caprioli, R. M.; Waelkens, E.; Van de Plas, R. Automated Anatomical Interpretation of Ion Distributions in Tissue: Linking Imaging Mass Spectrometry to Curated Atlases. *Anal. Chem.* **2014**, *86* (18), 8974–8982. https://doi.org/10.1021/ac502838t.

(22) Dewez, F.; Martin-Lorenzo, M.; Herfs, M.; Baiwir, D.; Mazzucchelli, G.; De Pauw, E.; Heeren, R. M. A.; Balluff, B. Precise Co-Registration of Mass Spectrometry Imaging, Histology, and Laser Microdissection-Based Omics. *Anal. Bioanal. Chem.* **2019**, *411* (22), 5647–5653. https://doi.org/10.1007/s00216-019-01983-z.

(23) Lin, L.-E.; Chen, C.-L.; Huang, Y.-C.; Chung, H.-H.; Lin, C.-W.; Chen, K.-C.; Peng, Y.-J.; Ding, S.-T.; Wang, M.-Y.; Shen, T.-L.; Hsu, C.-C. Precision Biomarker Discovery Powered by Microscopy Image Fusion-Assisted High Spatial Resolution Ambient Ionization Mass Spectrometry Imaging. *Anal. Chim. Acta* **2020**, *1100*, 75–87. https://doi.org/10.1016/j.aca.2019.11.014.

(24) Ščupáková, K.; Terzopoulos, V.; Jain, S.; Smeets, D.; Heeren, R. M. A. A Patch-Based Super Resolution Algorithm for Improving Image Resolution in Clinical Mass Spectrometry. *Sci. Rep.* **2019**, *9* (1), 2915. https://doi.org/10.1038/s41598-019-38914-y.

(25) Van de Plas, R.; Yang, J.; Spraggins, J.; Caprioli, R. M. Image Fusion of Mass Spectrometry and Microscopy: A Multimodality Paradigm for Molecular Tissue Mapping. *Nat. Methods* **2015**, *12* (4), 366–372. https://doi.org/10.1038/nmeth.3296.
(26) Wang, S.; Chen, W.; Xie, S. M.; Azzari, G.; Lobell, D. B. Weakly Supervised Deep Learning for Segmentation of Remote Sensing Imagery. *Remote Sens.* **2020**, *12* (2), 207. https://doi.org/10.3390/rs12020207.
(27) La Rocca, R.; Kune, C.; Tiquet, M.; Stuart, L.; Eppe, G.; Alexandrov, T.; De Pauw, E.; Quinton, L. Adaptive Pixel Mass Recalibration for Mass Spectrometry Imaging Based on Locally Endogenous Biological Signals. *Anal. Chem.* **2021**, *93* (8), 4066–4074. https://doi.org/10.1021/acs.analchem.0c05071.
(28) Alexandrov, T.; Ovchinnikova, K.; Palmer, A.; Kovalev, V.; Tarasov, A.; Stuart, L.; Nigmetzianov, R.; Fay, D.; Key METASPACE contributors; Gaudin, M.; Lopez, C. G.; Vetter, M.; Swales, J.; Bokhart, M.; Kompauer, M.; McKenzie, J.; Rappez, L.; Velickovic, D.; Lavigne, R.; Zhang, G.; Thinagaran, D.; Ruhland, E.; Sans, M.; Triana, S.; Sammour, D. A.; Aboulmagd, S.; Bagger, C.; Strittmatter, N.; Rigopoulos, A.; Gemperline, E.; Joensen, A. M.; Geier, B.; Quiason, C.; Weaver, E.; Prasad, M.; Balluff, B.; Nagornov, K.; Li, L.; Linscheid, M.; Hopf, C.; Heintz, D.; Liebeke, M.; Spengler, B.; Boughton, B.; Janfelt, C.; Sharma, K.; Pineau, C.; Anderton, C.; Ellis, S.; Becker, M.; Pánczél, J.; Violante, G. D.; Muddiman, D.; Goodwin, R.; Eberlin, L.; Takats, Z.; Shahidi-Latham, S. *METASPACE: A Community-Populated Knowledge Base of Spatial Metabolomes in Health and Disease*; preprint; Systems Biology, 2019. https://doi.org/10.1101/539478.
(29) Lawson, C. L.; Hanson, R. J. *Solving Least Squares Problems*; Society for Industrial and Applied Mathematics, 1995. https://doi.org/10.1137/1.9781611971217.
(30) Schmidt, M.; Le Roux, N.; Bach, F. Minimizing Finite Sums with the Stochastic Average Gradient. *Math. Program.* **2017**, *162* (1–2), 83–112. https://doi.org/10.1007/s10107-016-1030-6.

Conclusion and Perspectives

Contents

1 Conclusion 265

1.1 Overall significance 265

1.2 A practical pipeline to prioritize biology-relevant ions (Chapter 2) 265

1.3 Robust mass accuracy for reliable identification (Chapter 3) 266

1.4 Structuring high-resolution spectra beyond databases (Chapter 4 and Chapter 5) ... 267

1.5 Bridging MSI and microscopy to interpret space (Chapter 6) 267

1.6 How these pieces fit together 268

2 Perspectives 269

2.1 Smarter acquisition 269

2.2 Molecular formula computation 270

2.3 Noise and low-abundance signal 272

2.4 Graph based on molecular formula 274

2.5 Homogeneous family segmentation 277

2.6 Multimodal image fusion 279

2.7 Scalable, standardized evaluation 280

2.8 Closing statement 281

1 Conclusion

1.1 Overall significance

Mass spectrometry imaging (MSI) uniquely links molecular composition to spatial context across tissues, microbial systems, and cell cultures. Turning these rich, high-resolution signals into biological insight, however, is limited by (i) pixel-to-pixel mass shifts that weaken identification confidence, (ii) limited annotation coverage when databases or scoring fail to capture spectral complexity (adducts, isotopes, in-source fragments), and (iii) a spatial-resolution gap between MSI and microscopy that blunts cellular interpretation.

This thesis addresses those bottlenecks with an integrated informatics workflow. First, it stabilizes spectra through adaptive, pixel-wise recalibration to restore mass accuracy and increase FDR-controlled annotations. Second, it structures spectral information beyond database hits by exploiting high-resolution features, mass defects, and systematic mass differences through Kendrick-based filtering and a graph representation of spectra coupled with relational graph neural networks. Third, it bridges MSI and microscopy by utilizing deep image fusion to predict ion distributions at the microscopic scale, thereby improving interpretability at the (near) cellular scale. Together, these contributions move MSI analysis from manual curation toward robust, explainable, and scalable workflows, enabling more reliable discovery of condition-specific molecules and more faithful mapping of chemistry to phenotype.

1.2 A practical pipeline to prioritize biology-relevant ions (Chapter 2)

Building on the general introduction to MSI, Chapter 2 develops a general preprocessing and analysis workflow for MSI feature prioritization. Rather than focusing primarily on a single biological application, the chapter establishes the analytical framework

required to transform raw MSI data into a structured, analysis-ready representation. It describes the conversion of raw files into standardized formats, the construction of mean spectra, peak detection to define a common m/z feature space, the generation of ion images, and the organization of these data into representations suitable for downstream comparison across samples and conditions.

A central contribution of this chapter is to show that feature prioritization in MSI should not rely only on signal presence or relative intensity, but also on the spatial organization of ion images. Chapter 2 establishes a general strategy for reducing the complexity of MSI datasets and identifying a smaller set of ions that are most relevant to the biological question under study. This workflow was applied to a case study in bacterial natural product discovery, where it enabled the prioritization of ions associated with specific growth conditions in Streptomyces scabies. Beyond this application, Chapter 2 also laid the conceptual and computational foundation of the *deltamsi* package, in which these preprocessing and prioritization steps were generalized into a reusable framework that is extended in the following chapters.

1.3 Robust mass accuracy for reliable identification (Chapter 3)

Pixel-to-pixel mass shifts hinder MSI analysis by dispersing true peaks within tolerance ranges and distorting ion images. Chapter 3 countered this with an adaptive, pixel-wise recalibration that selects locally endogenous calibrants and recalibrates the spectrum individually. This recalibration process enables tighter m/z tolerances after correction and restores image morphology for downstream segmentation across diverse datasets. The quality of recalibration is reflected by an increase in METASPACE annotation. Because METASPACE annotation is FDR-controlled via a target–decoy framework based on ion spatial distribution, these gains reflect a real increase in reliable metabolite assignments rather than thresholding artifacts. Complementing the post-acquisition correction, we also document pre-acquisition controls, notably total-ion-current

management and 1ω/2ω detection in FT-ICR, that suppress drift at the source and sustain resolving power above 10^6 over whole images, further stabilizing the pipeline that feeds identification. Together, these measures re-establish mass accuracy as a dependable axis in MSI analysis and set up Chapter 4 and Chapter 5's mass-defect/graph-based interpretation and later spatial fusion investigation in Chapter 6.

1.4 Structuring high-resolution spectra beyond databases (Chapters 4 and 5)

Chapters 4 and 5 reintroduced mass-defect thinking into HR-MSI via Kendrick mass-defect (KMD) filtering, rapidly grouping homologous series and related ion families to prune background and guide family-level interpretation when database hits were sparse or ambiguous. We then generalized beyond KMD by representing each spectrum as a graph of mass differences (adducts, isotopes, neutral losses, repeating units), enabling simultaneous reasoning over multiple chemical relations, unlike classical KMD analyses that target one base at a time. Based on this representation, we developed a Relational Graph Convolutional Network (R-GCN) that robustly classified spectra under mass-shift and ion-loss perturbations, while remaining interpretable through class-activation maps that highlight coherent ion families across datasets. Collectively, these contributions enabled analysis of unannotated *m/z* values, reduced dependence on database matching, stabilized interpretation under common acquisition variability, and made HR-MS features, mass defects, and exact-mass relations necessary signals for MSI classification and downstream spatial analysis.

1.5 Bridging MSI and microscopy to interpret space (Chapter 6)

MSI's spatial resolution is limited to below the cell-level interpretation. Chapter 6 addressed this by fusing MSI with bright-field microscopy using a U-Net–based model that learns to project ion maps at microscope resolution. The approach linked morphological patterns to specific ion signals, improved interpretability beyond native

MSI resolution, and enabled downstream cell-type segmentation from the fused maps rather than from raw RGB alone. We observed practical robustness to ion delocalization and acquisition differences, while noting dependencies on sufficient training diversity. In effect, this chapter narrowed the resolution gap, transforming microscopy into a proxy for high-resolution molecular mapping and laying the groundwork for single-cell analyses and clinical tissue workflows that integrate histology with spatial metabolomics.

1.6 How these pieces fit together

Taken together, the contributions of this thesis form a coherent workflow for extracting biologically meaningful information from raw MSI data. Chapter 2 provides the foundation by organizing MSI data into analysis-ready representations and prioritizing ions of interest through complementary signal-based and spatially aware criteria. In doing so, it reduces the complexity of MSI datasets and defines the subset of molecular features that can be carried forward for more detailed interpretation.

Chapter 3 then stabilizes the m/z axis via adaptive, pixel-wise recalibration so downstream identification (e.g., METASPACE FDR calls) is richer and more reliable, while restoring ion-image morphology for segmentation and interpretation. Building on accurate spectra, Chapter 4 structures high-resolution signals with KMD filtering and a graph of exact-mass differences; Chapter 5 continues with an R-GCN on this representation, enabling robust, interpretable classification and family-level reasoning beyond database hits, enriching annotation and interpretation by leveraging mass-defect structure and exact-mass relations. Finally, Chapter 6 bridges MSI with microscopy, learning fused ion maps at microscope resolution so molecules selected and interpreted earlier can be localized at (near) cellular scales.

Thus, the different parts of the thesis are sequentially and conceptually linked: preprocessing and feature prioritization define the relevant ions; recalibration

improves their annotation; graph- and mass-defect-based analyses organize them into chemically meaningful structures; and image fusion places them back into a higher-resolution spatial context. Together, these stages advance MSI from ad hoc exploration toward a calibrated, interpretable, and scalable framework for spatial molecular analysis.

2 Perspectives

In the following section, we discuss seven possible perspectives based on the methods and results acquired during this thesis.

2.1 Smarter acquisition

High-resolution MSI (FT-ICR/Orbitrap) provides the mass accuracy required for confident formula assignment but remains costly in terms of acquisition time and data; a single high-resolution section can take ~1 day, limiting the number of replicates and statistical power. Accelerated acquisition with reconstruction has been proposed within MSI only, e.g., dynamic sparse sampling, which actively selects informative pixels and reconstructs images from a fraction of measurements[1], and joint compressed sensing and subspace modeling for FT-ICR, which shortens transient sampling and reconstructs HR signals[2]. Other studies use microscopy/histology to define regions of interest (ROIs) and then image those ROIs at high resolution[3–5]. Finally, fusion methods can predict HR-like ion maps from microscopy[6] (see also Chapter 6). These strategies can be combined in a single, two-stage, ROI-driven workflow. The first step is to get a rapid overview to delineate ROIs from morphology or coarse molecular heterogeneity. This can be done by microscopy and/or fast reduced spectral-resolution MSI (e.g., TOF). The second step rests on the acquisition of a (very) high-resolution image for those selected ROIs.

Because acquisition is dedicated to specific ROIs, parameters can be tuned per region to improve mass accuracy and resolving power. This is done by adapting FT-ICR settings to control TIC and laser dosing and, where appropriate, using short-transient 2ω detection followed by adaptive pixel-wise recalibration before annotation (Chapter 3). Moreover, since both overview and HR tiles are acquired on the same or adjacent sections, the Chapter 6 fusion model can be trained on HR/LR pairs within measured ROIs and then used to estimate HR-like ion maps in unacquired areas, yielding a near-full HR image from sparse HR sampling.

Once HR pixels per ROI are obtained, the data are fully compatible with our analysis stack: ISF/DSF for selecting condition-specific ions (Chapter 2), and mass-defect/graph-of-mass-differences for family-level interpretation (Chapter 4). Expected benefits are fewer HR pixels, shorter runs, and smaller data volumes while maintaining identification yield at fixed FDR and comparable spatial interpretability.

2.2 Molecular formula computation

In this thesis, small-molecule annotation relied on database matching (METASPACE) and organism-specific candidate lists. With high-resolution analyzers, a complementary approach is to estimate elemental formulas directly from accurate m/z values and isotopic patterns, which is especially valuable for unknowns absent from libraries.

However, ambiguity rises rapidly with mass: even at <0.1 ppm, a single accurate mass admits many candidate formulas at higher m/z, and isotopic-pattern agreement is needed to remove the large majority of them[7]. To constrain this space, Kind and Fiehn's "Seven Golden Rules" combine element-count limits, valence consistency (Lewis/Senior), isotopic-pattern fit, and admissible H/C and heteroatom (N, O, P, S) ratios, reducing all plausible CHNOPS formulas up to 2,000 Da from ~8 billion to ~623 million[8]. Of these, the isotope, valence, and element-ratio rules transfer directly to our underivatized HR-MSI setting (the trimethylsilylation rule does not), which makes

isotopic-envelope quality, and hence the mass-accuracy and S/N work of Chapter 3, the decisive factor in narrowing candidates. Database-driven matching encodes a useful prior: molecules seen in similar samples are more likely to recur, but it biases against novelty and is often unspecific to the sample/condition (e.g., broad resources such as HMDB, LIPID MAPS, or organismal collections like PAMDB). In natural-product discovery, genome mining provides a more targeted prior on what a strain can make: antiSMASH and PRISM identify biosynthetic gene clusters and predict monomers, tailoring, and candidate scaffolds/structures, yielding plausible compositions that constrain search space before MS interrogation[9,10].

Our Chapter 2/4 tools, ISF (*deltamsi*) to isolate condition-specific ions and mass-difference graphs to group co-varying families, fit naturally with these genome-derived expectations by (i) reducing the set of m/z to those most informative and (ii) pre-clustering m/z into chemically homogeneous families prior to identification.

Fragment information can push beyond sum formulas. ROI-targeted MS/MS can be matched to in-silico spectra (e.g., CFM-ID 4.0[11]). When on-tissue MS/MS is scarce, in-source fragments combined with ion-image co-localization enable pseudo-MS/MS (e.g., PICA[12]) that can also be scored against predicted fragments. Ion mobility adds an orthogonal, reproducible CCS constraint where available[13].

To improve formula estimation and coverage, we should investigate:

(i) Constrained formula estimation enforcing the Seven Golden Rules and explicit isotope-pattern agreement, supported by ROI pooling and targeted acquisition to boost envelope S/N ;

(ii) Strain/species-specific candidate sets from genome mining (antiSMASH/PRISM) used as priors to narrow plausible formulas for the studied BGCs ;

(iii) ROI-targeted MS/MS with in-silico matching (CFM-ID) to resolve isomers/isobars; and

(iv) Leveraging in-source fragmentation by deconvolving parent–fragment relationships from image similarity (pseudo-MS/MS) and scoring them against in-silico spectra, with ion-mobility (CCS) constraints where available.

Across these options, mass accuracy control from Chapter 3 remains foundational, lower ppm error directly shrinks candidate sets and improves both database- and constraint-based identification, while Chapters 2 and 4 help focus the search by prioritizing the m/z of interest and grouping them into chemically coherent families.

2.3 Noise and low-abundance signal

Low-abundance signals are frequent but underexplored in MSI. Low signal quality, driven by ion suppression, limited ion statistics per pixel (yielding poor spatial images), and incomplete isotopic envelopes, depresses annotation yield when identification relies on MS1-level evidence. The METASPACE engine explicitly scores isotope fit and spatial “chaos” under FDR control, and even with recent learning-based improvements, overall coverage remains limited, especially at stringent FDR[14]. Ion-suppression and preparation/topography effects are well-documented contributors to weak or variable signal in MALDI-MSI[15,16].

Chapter 2’s *deltamsi* package and ISD algorithm already leverage spatial structure to surface informative low-S/N ions: ISF quantifies spatial chaos to distinguish structured ions from random, noisy images and selects ions uniquely expressed under specific conditions, even when some are weak, thus acting as a conservative, image-organization–based denoiser. Mass-structure priors can also help rescue faint but relevant ions. As shown in Chapter 4, R-GCN + CAM uses mass defects and systematic mass-difference edges to highlight class-linked ions, even when some family members are weak or spatially diffuse, complementing intensity-centric filters.

In other omics fields, recovering noisy signal for interpretation is routine. In single-cell transcriptomics, the MSI pixel×m/z matrix parallels the cell×gene matrix: variance-

stabilizing transforms (sctransform), probabilistic denoising (scVI[17]), and neighborhood smoothing (MAGIC[18]) recover weak but structured expression. Brought to MSI, the same mindset aims to (i) restore spatial coherence in ion images, (ii) improve intensity estimates for faint peaks, and (iii) reconstruct isotopic patterns, changes that should translate into more confident METASPACE annotations at fixed FDR[14]. We have, however, to keep in mind that aggressive imputation can introduce false positives, so evaluation must be stringent.

Functionally, this can be assessed as in Chapter 3: compare pre/post processing by the gain in accepted annotations at fixed ppm/FDR and by improvements in isotope-fit, since METASPACE's framework is FDR-controlled and incorporates isotope and spatial terms.

However, weak signals behave differently across analyzers, notably in mass accuracy. In MALDI-FT-ICR, pixel-to-pixel mass shifts are strongly driven by ion-current fluctuations (space charge); mitigation via acquisition control and post-hoc recalibration (Chapter 3). In MALDI-TOF, sample topography/matrix heterogeneity dominates intensity/shift patterns, motivating topography-aware setups. In Orbitrap instruments, achieved mass accuracy degrades with falling S/N, supporting intensity-adaptive tolerances for weak peaks[19]. These differences argue for analyzer-aware settings and for checking that denoising does not bias low-S/N features.

Looking ahead, it will be important to investigate transcriptomics-style strategies for MSI and evaluate them on annotation-centric endpoints: apply denoisers that stabilize variance and leverage local neighborhoods/ROIs to reconstruct isotopic envelopes and refine intensities (e.g., deep autoencoder denoising such as DCA, by analogy), then quantify gains in METASPACE identifications at fixed FDR/ppm; make tolerances intensity-adaptive and analyzer-aware to avoid over-penalizing weak peaks; combine graph-based feature selection (R-GCN + CAM) with denoising to prioritize chemically coherent low-S/N ions; and establish a low-abundance benchmark (curated weak

signals and/or spike-in titrations) so MSI denoisers can be compared on standardized tasks, using isotope-fit recovery, spatial coherence, and annotation lift as primary metrics.

2.4 Graph based on molecular formula

A central limitation of Chapter 4's graph-of-mass-differences approach is selecting the set of allowed Δm: enlarging it captures more chemistries but inflates spurious edges and increases model complexity, even with R-GCN supervision. Mass-difference networks are well established in untargeted metabolomics for tracing biochemical transformations, and pairing them with Kendrick/RKMD analyses helps organize homologous series (e.g., CH_2 ladders, unsaturation series) before graph learning [20–22].

Chemically informed edge control is therefore key (e.g., adduct/isotope rules, matrix-aware exclusions, m/z-windowed Δm). A complementary route is a learned edge filter**:** for a given MSI dataset, derive weak-supervision labels from ion-image co-localization (e.g., cosine similarity), then train a lightweight classifier (e.g., XGBoost) on mass-only features of a pair of *m/z*, such as *m/z*, mass defects, mass difference, and mass defect of the mass difference, to predict keep/drop for candidate edges. This leverages the empirical enrichment of certain Δm and defect patterns among co-localized ions without hard-coding a fixed Δm list. A caveat is that family members need not perfectly co-localize; lipid unsaturation/isomerism often yields divergent spatial patterns, so these labels should be treated as weak supervision**.**

A second direction is formula-aware edges that do not require co-localization. Weight each edge by the minimal elemental-difference that explains Δm within tolerance under restricted elements/ranges, selecting the solution with minimal (fewest atoms changed or minimal mass) to favor parsimonious transformations, with explicit handling of isotopes/adducts. This mirrors the feasibility logic used by SIRIUS to prefer chemically consistent candidates and curbs implausible jumps in large graphs[23]. After

edge filtering, fuse this chemistry layer with a spatial co-localization layer (Chapter 4) into a multi-layer graph, then cluster with scalable community detection (Leiden/Louvain) on weighted edges; Leiden yields better-connected, higher-quality partitions in large networks[24]. An approximate version of this strategy is already implemented in the *deltamsi* package. In the current framework, graph edges are built by matching observed mass differences to a predefined catalog of candidate Δm values (and, optionally, associated molecular formula), each associated with a score that determines edge strength. When several candidate transformations explain the same observed Δm within tolerance, the candidate with the best score is retained, with mass error used as a tie-breaker. Leiden clustering is also implemented for community detection on the resulting weighted graph. However, a systematic evaluation is still needed to determine whether it improves molecular-family recovery, robustness, and interpretability across MSI datasets.

To evaluate such an approach**,** we need MSI datasets where many ions carry family/compound annotations (to compute purity/NMI and retrieval of homologous series) or where orthogonal confirmation (e.g., ROI-targeted LC–MS/MS) can supply family labels. This will provide task-appropriate metrics beyond MS1 FDR counts and quantify robustness to Δm-set size and ppm tolerances.

A promising perspective for MS1-level representation is to move from explicit graph construction toward transformer architectures, which appear well-suited to incorporating high-resolution mass features and the relationships derived from them. The R-GCN of Chapter 5 builds these relationships explicitly: peaks are connected through a small, curated set of chemically meaningful, typed mass differences, and the model learns in a supervised fashion from the comparatively few labeled MS1 images available in a given study. A transformer offers the opposite trade-off; it does not fix the relevant mass differences in advance but discovers them by comparing all *m/z* values against one another, learning these relationships implicitly from large amounts of unlabeled data.

DreaMS illustrates the potential of this approach on MS/MS data[25]. Each peak's *m/z* is encoded with high-resolution Fourier features, which efficiently represent both exact mass and mass defect for the deep-learning model, and all pairs of peaks are then compared through self-attention. Crucially, DreaMS injects the pairwise mass differences between peaks directly into its attention layers in a Graphormer-style[26] mechanism so that mass-difference and neutral-loss relationships are learned implicitly rather than enumerated by hand. Because self-attention is a graph-like operation in which every peak attends to every other, DreaMS captures inside its attention layers the same pairwise mass-difference relationships that the R-GCN encodes explicitly, expressing them as continuous, learned attention biases rather than as a curated set of discrete, typed edges. This expressiveness has costs: attention compares every pair of peaks, so it scales quadratically with the number of peaks, whereas R-GCN message passing scales linearly with the number of edges; and the far larger parameter count makes large-scale self-supervised pre-training a practical necessity. DreaMS is pre-trained on hundreds of millions of unlabeled spectra with two objectives: reconstructing masked m/z peaks and predicting the chromatographic retention order of two spectra from the same LC–MS/MS run. After this, the backbone is adapted to supervised tasks by attaching and fine-tuning a new task-specific head.

The same design principles, Fourier-encoded high-resolution *m/z*, attention over pairwise mass differences, and self-supervised pre-training on large unlabeled corpora, could be transferred to MS1 imaging, where peaks correspond to distinct ions and their mass differences encode transformations between molecular families. The masked-peak objective would carry over naturally; the retention-order objective would not, as imaging lacks a chromatographic dimension, so an analogous spatial pretext task (e.g., predicting ion co-localization or neighboring-pixel structure) would need to be devised. Such a model could offer a self-supervised, MS1-native alternative to the explicit mass-difference graph of Chapters 4 and 5, learning chemically meaningful mass

relationships directly from the large volumes of unlabeled imaging data that current supervised graph approaches cannot exploit.

2.5 Homogeneous family segmentation

Most current MSI segmentation methods operate on intensity patterns of individual ions, which can split regions that are chemically similar at the *family* level (e.g., GPC 34:x vs. GPC 35:x) while merging regions that are compositionally distinct (e.g., cholesterol-rich vs. phosphatidylcholine-rich regions). To increase spatial segmentation quality, we propose *family-aware* segmentation: regions should be homogeneous with respect to molecular families rather than single peaks. Practically, this means collapsing adduct/isotope variants and leveraging mass-structure relations (mass defects, exact Δm ladders) before or during segmentation, so that pixels are compared in a representation where members of the same family contribute coherently. This extends our Chapter 4 and 5 results, where a graph of mass differences and an R-GCN captured chemically related ions beyond raw intensity features.

A concrete path is to adapt the Chapter 5 R-GCN from *spectrum classification* to *unsupervised pixel embedding* for segmentation. Each pixel's spectrum is mapped to a graph (nodes = peaks, edges = chemically plausible Δm relations), and an R-GCN (or attention-based variant) produces a low-dimensional embedding per pixel that emphasizes family-level structure while de-emphasizing nuisance variation (adducts/isotopes). Neighboring pixels in the tissue graph are then encouraged to be close in embedding space, while distant pixels are pulled apart, mirroring successful graph-autoencoder approaches used to delineate spatial domains in transcriptomics (e.g., STAGATE; SpaGCN)[27]. Compared with classical MSI segmentation (e.g., spatial shrunken centroids in Cardinal[28]), a family-aware embedding should reduce over-segmentation driven by small chain-length/unsaturation shifts and improve boundary fidelity for compositionally distinct domains.

Two implementation details are important. First, build the pixel representation on top of the *chemistry graph* already validated in Chapter 4, Kendrick/RKMD cues to group homologs, plus Δm edges for isotopes, adduct swaps, and small transformations, so family membership informs the embedding rather than single peaks. Second, regularize with spatial neighbors (k-NN on coordinates) using attention or contrastive losses so that adjacency guides, but does not override, chemical evidence, conceptually similar to attention-based smoothing at domain boundaries in STAGATE.

Evaluation should mirror what is already done in segmentation evaluation: compare resulting segments to histology/ROI labels when available; quantify within-segment family enrichment (e.g., RKMD-organized lipid series enriched per region) and between-segment separation; and report robustness across Δm lists and ppm tolerances. Image co-localization metrics (e.g., cosine, as benchmarked in ColocML[29]) remain useful diagnostics to ensure that family-aware segments are not artifacts of a single dominant ion. Finally, CAM-style readouts from the adapted R-GCN can identify which ions/families drive each segment, preserving interpretability as in Chapter 5.

In sum, recasting segmentation to operate on *family-aware, graph-informed pixel embeddings*, learned with spatial attention and grounded in the Chapter 4 mass-structure graph, should yield regions that better reflect underlying molecular composition, reduce spurious splitting within families, and align more closely with biological domains.

2.6 Multimodal image fusion

MSI is information-rich but expensive and often limited in spatial resolution. Predictive image fusion extends its reach by learning cross-modal mappings from microscopy to ion images, enabling molecular interpretation at (near) optical resolution and on samples where MSI acquisition is impractical. In Chapter 6, we introduced a U-Net+CAM strategy that trains on the paired dataset at hand (no pre-trained cross-modal model required) to predict selected ion images from raw microscope tiles, then "lifts" the network's attention back to high-resolution fused maps suitable for downstream analysis. This differs from regression- or handcrafted-fusion approaches and avoids restricting fusion to modalities for which a pre-trained model exists.

Interesting extensions of our method could be to implement:

(i) Latent targets in place of raw ions. We could apply dimensionality-reduction methods (e.g., PCA or NMF) to the MSI cube to obtain a compact set of components that summarizes the full ion space. The U-Net+CAM model will be trained to predict these latent component maps directly from microscopy, and specific ions can then be reconstructed from the predicted components when needed. This removes the need for a fixed, limited ion list while keeping both training and outputs compact.

(ii) Family-level targets from Chapter 4. The mass-difference graph/embedding from Chapters 4 and 5 could be used to define *m/z* family clusters (e.g., lipid series/adduct groups). Select one representative m/z per cluster (e.g., cluster centroid), and the U-Net+CAM model will be trained to predict these representative ions. This preserves separation between compound families without predicting each ion individually.

Methodologically, our U-Net + CAM strategy can train without large external corpora, but its infancy warrants stronger evaluation: hold-out ions and slides, cross-sample generalization, and calibration of fused intensity (correlation/SSIM against MSI, isotope-pattern recovery where S/N permits). Extending beyond bright-field cell

culture to H&E tissue is a priority; paired, open MSI–histology datasets remain limited, which has historically constrained benchmarking in MSI fusion.

2.7 Scalable, standardized evaluation

At present, MSI lacks broadly accepted benchmarks, richly annotated datasets, and task-appropriate evaluation protocols. METASPACE is the main community resource for sharing imzML data and obtaining FDR-controlled MS1 annotations, and we leveraged it here, both to assess recalibration (Chapter 3) and to source datasets for Chapters 4–5. Yet this infrastructure is insufficient for many downstream tasks. In particular, MS1-only labels are too sparse to evaluate (i) ion clustering/family discovery (e.g., Chapters 4 and 5), (ii) graph-based classification, and (iii) spatial segmentation and fusion at cell scale.

We see two concrete gaps and how to close them:

(i) Chemically grounded clustering benchmarks. To judge whether a method groups ions into correct molecular families, we need datasets where a large fraction of m/z values have confirmed identities or at least family labels. Building such sets requires orthogonal chemistry on the *same sections*: ROI-targeted LC-MS(/MS) (laser microdissection or surface microextraction) to confirm exemplar ions per family; spike-in standards and homologous ladders to create known series; and family taxonomies (e.g., lipid classes) to provide label structure. For microbes, genome mining (antiSMASH/PRISM) can supply strain-specific expected families/formulas so clustering can be scored at the family level even when full structures are not known. These ingredients let us report family purity/NMI, retrieval/enrichment of homologous series, and robustness across Δm lists and ppm tolerances, metrics that FDR counts alone cannot provide.

(ii) Spatial segmentation (including cell level) and fusion. Organ-level ROIs exist for coarse segmentation, but cell-level ground truth is scarce, precisely what is needed to evaluate Chapter-6 image fusion. Benchmarks should therefore include co-registered multiplex immunofluorescence/RNAscope to provide cell-type masks and boundaries

on the same samples. Fusion and segmentation can then be scored with Dice/ARI, boundary F1, and calibration of predicted ion intensities against MSI (and isotope-pattern recovery where S/N permits). This closes the loop between morphological detail and molecular validity.

Putting this together, future releases should bundle: (i) data (imzML + co-acquired LC-MS/MS and, where relevant, multiplex IF and omics tables); (ii) labels (confirmed ions/families, ROIs/cell masks, genome-derived priors); (iii) tasks (recalibration, ion family clustering, graph-based classification, spatial segmentation, and fusion); and (iv) metrics matched to each task (FDR-controlled identification, family coherence/retrieval, segmentation accuracy/boundary metrics, and interpretability readouts such as CAM coverage of family graphs).

2.8 Closing statement

This thesis presents a comprehensive toolkit that stabilizes spectral data, structures chemical analysis, and enhances spatial interpretation. It emphasizes high-resolution spectral features such as mass defects and exact-mass relations as primary elements. Additionally, it integrates MSI with microscopy through deep learning, transitioning MSI analysis from manual, ad-hoc methods to reliable, explainable, and scalable workflows. As instrumentation (e.g., FT-ICR) and bioinformatics (e.g., graph learning, FDR-aware annotation) continue to mature, MSI is poised to become a routine engine for spatial biochemistry, from microbial ecology to pathology and precision medicine.

REFERENCES:

(1) Hu, H.; Helminiak, D.; Yang, M.; Unsihuay, D.; Hilger, R. T.; Ye, D. H.; Laskin, J. High-Throughput Mass Spectrometry Imaging with Dynamic Sparse Sampling. *ACS Meas. Sci. Au* **2022**, *2* (5), 466–474. https://doi.org/10.1021/acsmeasuresciau.2c00031.

(2) Xie, Y. R.; Castro, D. C.; Rubakhin, S. S.; Sweedler, J. V.; Lam, F. Enhancing the Throughput of FT Mass Spectrometry Imaging Using Joint Compressed Sensing and Subspace Modeling. *Anal. Chem.* **2022**, *94* (13), 5335–5343. https://doi.org/10.1021/acs.analchem.1c05279.

(3) Joignant, A. N.; Ritter, M. M.; Knizner, K. T.; Garrard, K. P.; Kullman, S. W.; Muddiman, D. C. Maximized Spatial Information and Minimized Acquisition Time of Top-Hat IR-MALDESI-MSI of Zebrafish Using Nested Regions of Interest (nROIs). *J. Am. Soc. Mass Spectrom.* **2023**, *34* (9), 2043–2050. https://doi.org/10.1021/jasms.3c00210.

(4) Heijs, B.; Abdelmoula, W. M.; Lou, S.; Briaire-de Bruijn, I. H.; Dijkstra, J.; Bovée, J. V. M. G.; McDonnell, L. A. Histology-Guided High-Resolution Matrix-Assisted Laser Desorption Ionization Mass Spectrometry Imaging. *Anal. Chem.* **2015**, *87* (24), 11978–11983. https://doi.org/10.1021/acs.analchem.5b03610.

(5) Patterson, N. H.; Tuck, M.; Lewis, A.; Kaushansky, A.; Norris, J. L.; Van De Plas, R.; Caprioli, R. M. Next Generation Histology-Directed Imaging Mass Spectrometry Driven by Autofluorescence Microscopy. *Anal. Chem.* **2018**, *90* (21), 12404–12413. https://doi.org/10.1021/acs.analchem.8b02885.

(6) Van De Plas, R.; Yang, J.; Spraggins, J.; Caprioli, R. M. Image Fusion of Mass Spectrometry and Microscopy: A Multimodality Paradigm for Molecular Tissue Mapping. *Nat Methods* **2015**, *12* (4), 366–372. https://doi.org/10.1038/nmeth.3296.

(7) Kind, T.; Fiehn, O. Metabolomic Database Annotations via Query of Elemental Compositions: Mass Accuracy Is Insufficient Even at Less than 1 Ppm. *BMC Bioinformatics* **2006**, *7* (1), 234. https://doi.org/10.1186/1471-2105-7-234.

(8) Kind, T.; Fiehn, O. Seven Golden Rules for Heuristic Filtering of Molecular Formulas Obtained by Accurate Mass Spectrometry. *BMC Bioinformatics* **2007**, *8* (1), 105. https://doi.org/10.1186/1471-2105-8-105.

(9) Blin, K.; Shaw, S.; Augustijn, H. E.; Reitz, Z. L.; Biermann, F.; Alanjary, M.; Fetter, A.; Terlouw, B. R.; Metcalf, W. W.; Helfrich, E. J. N.; van Wezel, G. P.; Medema, M. H.; Weber, T. antiSMASH 7.0: New and Improved Predictions for Detection, Regulation, Chemical Structures and Visualisation. *Nucleic Acids Research* **2023**, *51* (W1), W46–W50. https://doi.org/10.1093/nar/gkad344.

(10) Skinnider, M. A.; Johnston, C. W.; Gunabalasingam, M.; Merwin, N. J.; Kieliszek, A. M.; MacLellan, R. J.; Li, H.; Ranieri, M. R. M.; Webster, A. L. H.; Cao, M. P. T.; Pfeifle, A.; Spencer, N.; To, Q. H.; Wallace, D. P.; Dejong, C. A.; Magarvey, N. A. Comprehensive Prediction of Secondary Metabolite Structure and Biological Activity from Microbial Genome Sequences. *Nat Commun* **2020**, *11* (1), 6058. https://doi.org/10.1038/s41467-020-19986-1.

(11) Wang, F.; Liigand, J.; Tian, S.; Arndt, D.; Greiner, R.; Wishart, D. S. CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. *Anal. Chem.* **2021**, *93* (34), 11692–11700. https://doi.org/10.1021/acs.analchem.1c01465.

(12) Dong, Y.; Shachaf, N.; Feldberg, L.; Rogachev, I.; Heinig, U.; Aharoni, A. PICA: Pixel Intensity Correlation Analysis for Deconvolution and Metabolite Identification in Mass Spectrometry Imaging. *Anal. Chem.* **2023**, acs.analchem.2c04778. https://doi.org/10.1021/acs.analchem.2c04778.

(13) Zhou, Z.; Luo, M.; Chen, X.; Yin, Y.; Xiong, X.; Wang, R.; Zhu, Z.-J. Ion Mobility Collision Cross-Section Atlas for Known and Unknown Metabolite Annotation in Untargeted Metabolomics. *Nat Commun* **2020**, *11* (1), 4334. https://doi.org/10.1038/s41467-020-18171-8.
(14) Palmer, A.; Phapale, P.; Chernyavsky, I.; Lavigne, R.; Fay, D.; Tarasov, A.; Kovalev, V.; Fuchser, J.; Nikolenko, S.; Pineau, C.; Becker, M.; Alexandrov, T. FDR-Controlled Metabolite Annotation for High-Resolution Imaging Mass Spectrometry. *Nature Methods* **2017**, *14* (1), 57–60. https://doi.org/10.1038/nmeth.4072.
(15) Balluff, B.; Hopf, C.; Porta Siegel, T.; Grabsch, H. I.; Heeren, R. M. A. Batch Effects in MALDI Mass Spectrometry Imaging. *J. Am. Soc. Mass Spectrom.* **2021**, *32* (3), 628–635. https://doi.org/10.1021/jasms.0c00393.
(16) Taylor, A. J.; Dexter, A.; Bunch, J. Exploring Ion Suppression in Mass Spectrometry Imaging of a Heterogeneous Tissue. *Anal. Chem.* **2018**, *90* (9), 5637–5645. https://doi.org/10.1021/acs.analchem.7b05005.
(17) Lopez, R.; Regier, J.; Cole, M. B.; Jordan, M. I.; Yosef, N. Deep Generative Modeling for Single-Cell Transcriptomics. *Nat Methods* **2018**, *15* (12), 1053–1058. https://doi.org/10.1038/s41592-018-0229-2.
(18) Van Dijk, D.; Sharma, R.; Nainys, J.; Yim, K.; Kathail, P.; Carr, A. J.; Burdziak, C.; Moon, K. R.; Chaffer, C. L.; Pattabiraman, D.; Bierie, B.; Mazutis, L.; Wolf, G.; Krishnaswamy, S.; Pe'er, D. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **2018**, *174* (3), 716-729.e27. https://doi.org/10.1016/j.cell.2018.05.061.
(19) Olsen, J. V.; De Godoy, L. M. F.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M. Parts per Million Mass Accuracy on an Orbitrap Mass Spectrometer via Lock Mass Injection into a C-Trap. *Molecular & Cellular Proteomics* **2005**, *4* (12), 2010–2021. https://doi.org/10.1074/mcp.T500030-MCP200.
(20) Amara, A.; Frainay, C.; Jourdan, F.; Naake, T.; Neumann, S.; Novoa-del-Toro, E. M.; Salek, R. M.; Salzer, L.; Scharfenberg, S.; Witting, M. Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation. *Front. Mol. Biosci.* **2022**, *9*, 841373. https://doi.org/10.3389/fmolb.2022.841373.
(21) Traquete, F.; Luz, J.; Cordeiro, C.; Sousa Silva, M.; Ferreira, A. E. N. Graph Properties of Mass-Difference Networks for Profiling and Discrimination in Untargeted Metabolomics. *Front. Mol. Biosci.* **2022**, *9*, 917911. https://doi.org/10.3389/fmolb.2022.917911.
(22) Richardson, L. T.; Neumann, E. K.; Caprioli, R. M.; Spraggins, J. M.; Solouki, T. Referenced Kendrick Mass Defect Annotation and Class-Based Filtering of Imaging MS Lipidomics Experiments. *Anal. Chem.* **2022**, *94* (14), 5504–5513. https://doi.org/10.1021/acs.analchem.1c03715.
(23) Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information. *Nat Methods* **2019**, *16* (4), 299–302. https://doi.org/10.1038/s41592-019-0344-8.
(24) Traag, V. A.; Waltman, L.; Van Eck, N. J. From Louvain to Leiden: Guaranteeing Well-Connected Communities. *Sci Rep* **2019**, *9* (1), 5233. https://doi.org/10.1038/s41598-019-41695-z.
(25) Bushuiev, R.; Bushuiev, A.; Samusevich, R.; Brungs, C.; Sivic, J.; Pluskal, T. Self-Supervised Learning of Molecular Representations from Millions of Tandem Mass Spectra Using DreaMS. *Nat Biotechnol* **2026**, *44* (4), 630–640. https://doi.org/10.1038/s41587-025-02663-3.
(26) Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; Liu, T.-Y. Do Transformers Really Perform Badly for Graph Representation? In *Advances in Neural Information Processing Systems*; 2021; Vol. 34, pp 28877–28888.

(27) Dong, K.; Zhang, S. Deciphering Spatial Domains from Spatially Resolved Transcriptomics with an Adaptive Graph Attention Auto-Encoder. *Nat Commun* **2022**, *13* (1), 1739. https://doi.org/10.1038/s41467-022-29439-6.
(28) Bemis, K. D.; Harry, A.; Eberlin, L. S.; Ferreira, C.; van de Ven, S. M.; Mallick, P.; Stolowitz, M.; Vitek, O. *Cardinal* : An R Package for Statistical Analysis of Mass Spectrometry-Based Imaging Experiments. *Bioinformatics* **2015**, *31* (14), 2418–2420. https://doi.org/10.1093/bioinformatics/btv146.
(29) Ovchinnikova, K.; Stuart, L.; Rakhlin, A.; Nikolenko, S.; Alexandrov, T. ColocML: Machine Learning Quantifies Co-Localization between Mass Spectrometry Images. *Bioinformatics* **2020**, *36* (10), 3215–3224. https://doi.org/10.1093/bioinformatics/btaa085.

Supporting information: Table of Contents

Supporting information of Chapter 2

S1 : Scoring ion image structure

S2 : deltamsi functions descriptions

Supporting information of Chapter 3

S1 : Description of the 31 METASPACE MSI data according to the initial uploader

S2 : Full detailed method section

S3 : Plots for the 31 METASPACE MSI dataset

Supporting information of Chapter 5

S1: Description and id of the 4 METASPACE MSI data according to the initial uploader and an additional dataset from our laboratory

S2: Background and Foreground annotation for each selected pixel of the 4 METASPACE MSI

S3: Set of mass differences of the most frequent mass patterns in MALDI-MSI experiments from the work of Janda M. et al[1]

S4: Annotated m/z of all dataset and their computed CAM values, Fold change and correlation

S5: Ion images corresponding to the annotated m/z in S4, Pseudomonas.

S6: Material and Sample annotations

S7: R-GCN Implementation Description

S8: Performance comparison of different R-GCN parameters

S9: 1D-CNN, SVM, RF, LDA, XGBoost pre-processing

S10: Evaluation of unseen data

S11: Model evaluation details

Supporting information of Chapter 6

S1: Molecular annotation workflow

Supporting information of Chapter 2

S1: Spatial Chaos Score

S1.1 Spatial Chaos Score

The spatial chaos score quantifies the degree of structure or randomness in an ion image based on the spatial distribution of its intensities. The input is a grayscale ion image, where each pixel contains a non-negative intensity value corresponding to the abundance of a given ion at that spatial location. The goal is to assess whether these intensities form coherent, contiguous structures (e.g., interaction zones) or are scattered randomly across the image. To do so, the ion image is transformed into a series of binary images at different intensity thresholds, and the number and size of connected pixel clusters are analyzed (Figure S1).

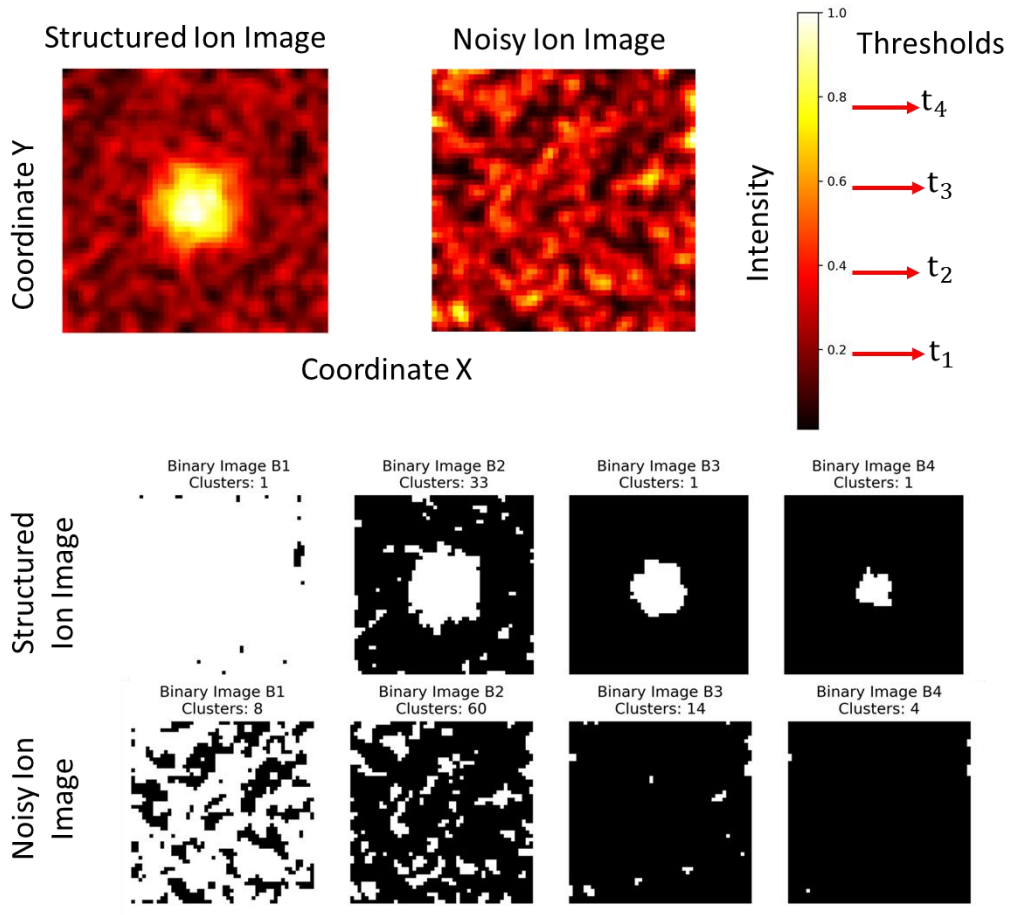


Figure S1. *illustrates the generation of binary images and the clustering of pixels to measure spatial chaos from two ion images: one with a structured distribution (structured ion image) and the other with a random spatial distribution of intensity (noisy ion image). For each ion image, the maximum intensity is normalized to 1 (white), with values ranging from 0 to 1. N thresholds are then selected, evenly spaced within this intensity range (from t1 to t4). For each threshold, a binary image is created, where white pixels correspond to values in the ion image that exceed the threshold. For each binary image, the number of clusters is calculated; clusters are defined as groups of white pixels separated by black pixels. The noisy ion image shows more clusters with fewer pixels than the structured ion image.*

The algorithm used to compute the spatial chaos score can be summarised as follows.

1) Normalization

Let $I(x, y)$ denote the ion image, where (x, y) indexes pixel coordinates and $I(x, y) \geq 0$ is the ion intensity at that pixel. The image is first normalized to the range [0, 1] by dividing by its maximum intensity:

$$I_{\text{norm}}(x, y) = \frac{I(x, y)}{\max I(x, y)}$$

This step ensures that intensity values are on a comparable scale across images, allowing a fixed set of thresholds to be applied consistently.

2) Thresholding into binary images

A set of N thresholds $\{t_k\}_{k=1}^{N}$ is defined, evenly spaced between 0 and 1. For each threshold t_k, a binary image $B_k(x, y)$ is created as

$$\boldsymbol{B_k}(x, y) = \begin{cases} 1 \text{ if } \boldsymbol{I}_{\text{norm}}(x, y) \geq t_k \\ 0 \qquad otherwise \end{cases}$$

This produces a sequence of binary images $\{B_k\}$, each corresponding to a different intensity threshold. Pixels above the threshold are set to 1 (white), and all others to 0 (black).

3) Cluster (connected components) identification

For each binary image B_k, the algorithm identifies clusters of connected pixels with value 1. A cluster (connected component) is a group of adjacent pixels under 4-connectivity (pixels are connected if they share a side). For each threshold t_k, we compute:

The algorithm calculates:

- C_k : the number of clusters (connected components) in B_k.
- P_k: the number of positive (non-zero) pixels in B_k.

Thus, $C_k$counts how many separate “islands” of signal are present at that threshold, and $P_k$measures the total area covered by the signal.

4) Cumulative totals across thresholds

As the algorithm iterates over all thresholds t_k, it accumulates the following totals:

$$C_{tot} = \sum_{k=1}^{N} C_k \, , P_{tot} = \sum_{k=1}^{N} P_k$$

Here, C_{tot} is the total number of clusters summed across all binary images, and P_{tot} is the total number of positive pixels summed across all binary images. These cumulative quantities summaries how "fragmented" the signal is across intensity levels.

5) Overall spatial chaos score

The spatial chaos score S is finally defined as:

After processing all the binary images, the spatial chaos score is calculated as:

$$S = 1 - \frac{C_{tot}}{P_{tot}}$$

By construction, $0 \leq \frac{C_{tot}}{P_{tot}} \leq 1$, so S also lies between 0 and 1. The score provides a measure of spatial organization:

- A score close to 0 indicates a high degree of spatial randomness (high chaos): many clusters with very few pixels each, as would occur if positive pixels were scattered randomly across the image.
- A score close to 1 indicates a more structured and organized spatial distribution: relatively few clusters containing many pixels, as expected for ions forming coherent regions or fronts.

In practice, noisy ion images tend to produce a large ratio $\frac{C_{tot}}{P_{tot}}$ and therefore low S, whereas biologically structured images produce a smaller ratio and thus higher S. Although we refer to S as a "spatial chaos score" for consistency with the original terminology, larger values of S actually correspond to lower spatial chaos (higher spatial structure). This score is then used,

in the following subsection, to compare ion images between control and condition groups via a fold-change analysis of their spatial organisation.

S1.2 Fold Change Calculation

The goal of this step is to compare the spatial organization of ion images between two groups: the control group and the condition (interaction) group. For each m/z value, we first compute a spatial chaos score for every ion image in both groups (Section 2.3.1), then summarize these scores at the group level, and finally calculate a fold change between groups. This fold change plays a role analogous to intensity-based fold change in LC–MS, but here it reflects differences in spatial *structure* rather than in total abundance.

Spatial chaos scores for each ion image

Let $S_{g,s}(m)$ denote the spatial chaos score of the ion image corresponding to m/z value m, for sample s in group g, where $g \in$ {control, interaction}. The score $S_{g,s}(m)$ is computed using the procedure described in Section 2.3.1 (normalization, thresholding, cluster counting, and aggregation across thresholds). By construction, the score lies between 0 and 1:

- $S_{g,s}(m) \approx 0$ for highly chaotic, noisy ion images (many small clusters, few pixels per cluster).
- $S_{g,s}(m) \approx 1$ for highly structured ion images (few clusters, many pixels per cluster).

Although we retain the terminology “spatial chaos score” for consistency, larger values of S in practice correspond to lower spatial chaos (i.e. more spatial structure).

Maximum spatial chaos score per group

For each m/z value m and each group g, we summarise the scores across all MSI datasets in that group by taking the maximum value:

$$S_g^{\max}(m) = \max_{s \in g} S_{g,s}(m).$$

This yields two summary scores for each m/z: one for the control group $S_{\text{control}}^{\max}(m)$ and one for the interaction group $S_{\text{interaction}}^{\max}(m)$. Intuitively, $S_g^{\max}(m)$ represents the most structured ion image observed for that m/z within group g. Using the maximum rather than, for example, the mean score allows us to detect ions that form a clear structure in at least one replicate, even if the pattern is weaker or absent in others.

Fold change between groups

To quantify the change in spatial organization between the two groups, we define a structure-based fold change for each m/z value:

$$\text{FC}_S(m) = \frac{S_{\text{interaction}}^{\max}(m)}{S_{\text{control}}^{\max}(m)}.$$

In practice, if $S_{\text{control}}^{\max}(m)$ is extremely small or zero, a small constant can be added to the denominator to avoid numerical instabilities. Conceptually, however, $\text{FC}_S(m)$ remains the ratio of the most structured ion image in the interaction group to the most structured ion image in the control group.

Interpretation

The fold change $\text{FC}_S(m)$ provides a direct measure of how the spatial structure of an ion image changes between groups:

- $\text{FC}_S(m) > 1$: the ion image is *more structured* in the interaction group than in the control group. Equivalently, the spatial distribution of the ion is more random (chaotic) in the control group, and a coherent spatial pattern emerges under the interaction condition.
- $\text{FC}_S(m) \approx 1$: the spatial organisation of the ion is similar in both groups.
- $\text{FC}_S(m) < 1$: the ion is more structured in the control group than in the interaction group (for example, a structure present at baseline that disappears or becomes more diffuse under the condition).

By analysing $FC_S(m)$ across all detected m/z values, we can identify ions whose spatial patterns change markedly between control and interaction conditions. In particular, ions with high $FC_S(m)$ correspond to features that are spatially disordered in the controls but form clear, localized structures in the interaction group.

S2: deltamsi functions descriptions

Scope and purpose

This document provides the primary function descriptions for the *deltamsi* package, with particular focus on *src/deltamsi/core/msicube.py* (the *MSICube* orchestration layer) and the processing modules that these methods wrap or call, including preprocessing, normalization, peak picking, colocalization, spatial-chaos scoring, recalibration, mass clustering, and Kendrick-space analysis.

The descriptions in this document correspond to the implementation currently available in the *LaRoccaRaphael/deltamsi*. At a high level, *MSICube* implements an end-to-end MSI workflow that (i) scans a project directory for imzML/ibd datasets; (ii) builds per-sample and global mean spectra; (iii) defines study-wide features via peak picking; (iv) extracts a dense pixel × feature matrix into an *AnnData* container; and (v) provides optional steps for intensity normalization/scaling, spatial preprocessing (hotspot capping, quantile thresholding, 2D median filtering), cosine-based ion colocalization, graph-based mass clustering (Leiden), aggregation of features by labels, spatial-chaos scoring and fold-change, robust mass recalibration via KDE + RANSAC, and Kendrick-space projection and annotation. Key external standards and references relevant to these methods include the imzML file format (paired XML + binary)[1], the *AnnData*/Scanpy data model for annotated matrices[2], pyimzML for imzML I/O[3], cosine colocalization approaches in MSI (e.g., ColocML)[4], spatial-chaos metrics used in MSI quality/annotation pipelines[5], robust regression via RANSAC[6], and Leiden community detection for graph clustering [7].

Code referencing note: exact code references are given as file path + symbol name so that each method can be located deterministically by text search within the source tree (for example, by searching *def pick_peaks* in the referenced file).

Repository context and *MSICube* data model

Mass spectrometry imaging data are handled as imzML datasets, which conventionally store metadata in an XML-based *.imzML* file and spectral arrays in a paired binary file (commonly *.ibd*). The imzML format was created to standardize MSI data exchange and tool interoperability[1], and the code here uses the pyimzML stack for reading/writing such datasets[3].

MSICube uses *AnnData* as its central in-memory and on-disk representation. *AnnData* organizes a main matrix *X* with aligned observation metadata (*obs*) and variable metadata (*var*), plus structured slots like *layers*, *obsm*, *varm*, and pairwise variable matrices *varp*, which is consistent with the Scanpy/AnnData ecosystem design[2]. Persistence is implemented using *AnnData.write_h5ad* (HDF5) and *AnnData.write_zarr* (chunked Zarr) with corresponding readers, as documented in the *anndata* API.

MSICube object

Method description (paper-ready): *MSICube* (class) — *Code reference:* *src/deltamsi/core/msicube.py* (class *MSICube*; line ranges unspecified). *MSICube* is the central orchestration layer of the *accessory_function* branch. Rather than representing a single numerical transform, it coordinates the full project lifecycle: discovery of raw imzML/ibd datasets, construction of per-sample and global mean spectra, feature definition by peak picking, pixel-wise extraction of a dense feature matrix, intensity preprocessing and normalization, image-space filtering, pairwise ion analysis, graph-based mass clustering, Kendrick-space projection and annotation, recalibration, and visualization. In that sense, *MSICube* acts as a project controller wrapped around a persistent *AnnData* state rather than as a simple container object.

On initialization, the class validates the provided *data_directory*, scans it for *.imzML* files, verifies that each candidate has a corresponding *.ibd* companion file, and registers valid datasets in *org_imzml_path_dict*. Sample names are inferred from filenames without extension. This discovery step is lightweight and metadata-oriented: it indexes available MSI datasets but does not yet load spectral data into memory. Raw data remain on disk until explicit extraction or recalibration steps are invoked. This behavior is consistent with the paired-file design of the imzML standard and with pyimzML-based workflows for MSI I/O[1,3].

Internally, *MSICube* uses *AnnData* as the canonical in-memory representation. Pixel × feature intensities are stored in *adata.X*; alternative processed matrices can be written to *adata.layers*; pixel metadata such as sample labels are stored in *adata.obs*; feature metadata such as m/z values and cluster annotations are stored in *adata.var*; spatial coordinates are typically stored in *adata.obsm["spatial"]*; feature embeddings such as spatial-chaos matrices are stored in *adata.varm*; pairwise feature relationships such as cosine colocalization or graph connectivities are stored in *adata.varp*; and workflow metadata, options, mean spectra, and summary tables are stored in *adata.uns*. This design follows the annotated-matrix model used in Scanpy/AnnData and makes MSI preprocessing and downstream interpretation share a single synchronized project state[2].

A key aspect of the class is that some methods are thin wrappers over backend numerical routines, whereas others are true orchestration methods that restructure *AnnData* or update project-level file paths. For example, *pick_peaks* can rebuild *adata.var* and resize the full *AnnData* object; *extract_matrix* reconstructs the project-wide pixel × feature matrix and assigns sample/spatial metadata; *recalibrate* physically writes corrected imzML files and then updates *org_imzml_path_dict* and *data_directory* to point to the recalibrated datasets; and persistence helpers (*save*, *load*, *from_file*) make it possible to checkpoint and resume entire MSI projects through *h5ad* or Zarr storage supported by *anndata*[2].

From a workflow perspective, the class therefore serves three complementary roles: (i) a registry of raw MSI datasets, (ii) an *AnnData*-backed state machine for progressive preprocessing and feature engineering, and (iii) a user-facing API that exposes analysis, visualization, and manual-annotation utilities under one object. Complexity is dominated by downstream methods, whereas object construction itself is typically linear in the number of files

present in the directory and does not scale with the total number of pixels or spectral peaks until explicit processing begins.

Additional orchestration, persistence, and utility methods in *MSICube*

Project discovery and persistence

__init__ — *Code reference:* *src/deltamsi/core/msicube.py* (method *MSICube.__init__*; line ranges unspecified). This constructor initializes the project by validating *data_directory*, calling *_scan_imzml_files*, creating the *org_imzml_path_dict* registry, and setting *adata=None*. Its purpose is to create a lightweight project object that knows where the raw MSI files are stored before any heavy numerical processing begins. Complexity is approximately **O(n_files)** in the number of directory entries inspected.

_scan_imzml_files — *Code reference:* *src/deltamsi/core/msicube.py* (method *MSICube._scan_imzml_files*; line ranges unspecified). This internal helper scans a directory for *.imzML*/*.imzml* files, checks for the corresponding *.ibd* companion, and stores only valid dataset pairs in *org_imzml_path_dict*. Files lacking an *.ibd* partner are skipped with a warning rather than raising an exception. The method is therefore responsible for the raw-data registry used by the rest of the *MSICube* workflow. Runtime is linear in the number of filenames examined.

_default_adata_path — *Code reference:* *src/deltamsi/core/msicube.py* (method *MSICube._default_adata_path*; line ranges unspecified). This helper constructs the default persistence target for the project state as *adata.h5ad* or *adata.zarr* inside *data_directory*. Although simple, it is an important part of the class design because it standardizes where checkpointed project states are stored when explicit output paths are not supplied. Complexity is **O(1)**.

save — *Code reference:* *src/deltamsi/core/msicube.py* (method *MSICube.save*; line ranges unspecified). This method serializes the current *AnnData* state to disk in either HDF5-backed *.h5ad* format or Zarr format. In the branch implementation, it also performs branch-specific handling of *adata.uns["intensity_clipping"]* so that lists of dictionaries remain serializable when written to HDF5. The method therefore provides project checkpointing rather than a numerical transform. Runtime and output size scale with the size of the stored *AnnData* object, typically **O(size of serialized adata)**.

load — *Code reference:* *src/deltamsi/core/msicube.py* (method *MSICube.load*; line ranges unspecified). This method restores a previously persisted *AnnData* object from *.h5ad* or Zarr back into *self.adata*. It makes the *MSICube* workflow resumable and allows downstream analyses to be restarted from a saved project state without repeating extraction or preprocessing. Runtime is dominated by deserialization cost and is roughly proportional to the size of the saved object.

from_file — *Code reference:* *src/deltamsi/core/msicube.py* (classmethod *MSICube.from_file*; line ranges unspecified). This factory method combines object

construction and project-state loading into a single call. It is a convenience entrypoint for resuming analyses from an existing saved *AnnData* representation while still rebuilding the surrounding *MSICube* object with its directory context. Complexity is the sum of constructor overhead plus the *load* cost.

Workflow orchestration and cleanup helpers

clear_mean_spectra — *Code reference:* *src/deltamsi/core/msicube.py* (method *MSICube.clear_mean_spectra*; line ranges unspecified). This method removes cached mean-spectrum objects and their provenance entries from *adata.uns*. Its main purpose is to reduce memory footprint and file size once peak picking has been completed and the intermediate mean-spectrum representations are no longer needed. Complexity is effectively **O(1)** with respect to metadata keys removed, excluding Python object deallocation costs.

create_feature_matrix — *Code reference:* *src/deltamsi/core/msicube.py* (method *MSICube.create_feature_matrix*; line ranges unspecified). This high-level convenience method chains *pick_peaks* and *extract_matrix*, thereby constructing the full pixel × feature intensity matrix from the previously computed global mean spectrum. Conceptually, it is the main orchestration shortcut that converts a project with defined spectral references into a data-rich *AnnData* matrix ready for downstream analysis. Complexity is the combined cost of peak picking and matrix extraction.

Visualization and interrogation helpers

plot_ion_images — *Code reference:* *src/deltamsi/core/msicube.py* (method *MSICube.plot_ion_images*; line ranges unspecified). This method visualizes one or more feature images across selected samples, drawing intensities either from *adata.X*/*layers[layer]* or from aggregated representations stored in *adata.obsm*. It is a project-level QC and interpretation endpoint that turns the stored *AnnData* state back into MSI ion images. Runtime is dominated by image assembly and Matplotlib rendering and is typically linear in the number of plotted pixels and requested features.

plot_peak_windows — *Code reference:* *src/deltamsi/core/msicube.py* (method *MSICube.plot_peak_windows*; line ranges unspecified). This method plots zoomed spectral windows around user-specified m/z values using the stored mean spectra. It is mainly intended for quality control of peak alignment, extraction tolerances, and spectral comparability across samples. Computational cost is modest and scales with the number of plotted windows and selected samples, with rendering usually dominating runtime.

plot_recalibration — *Code reference:* *src/deltamsi/core/msicube.py* (method *MSICube.plot_recalibration*; line ranges unspecified). This method generates diagnostic plots for recalibration quality in selected pixels of a given sample. It uses the same database-matching and recalibration parameterization as the recalibration backend but stops at visualization of KDE error profiles and fitted correction models rather than writing corrected imzML files. It is therefore a QC companion to *recalibrate*. Runtime depends on the number of selected pixels and the cost of per-pixel diagnostics.

direct_mass_neighbors — *Code reference: src/deltamsi/core/msicube.py* (method *MSICube.direct_mass_neighbors*; line ranges unspecified). This helper queries the stored mass-difference network and returns the direct neighbors of a chosen feature together with edge annotations and optional cosine-colocalization information. It is a graph-interrogation utility rather than a clustering routine, allowing local inspection of the mass network produced by *cluster_masses*. Complexity is proportional to the degree of the queried node plus table assembly overhead.

Preprocessing and normalization methods in *MSICube*

Intensity clipping/masking and log transform

clip_or_mask_intensities — *Code reference: src/deltamsi/core/msicube.py* (method *MSICube.clip_or_mask_intensities*; line ranges unspecified). Underlying implementation: *src/deltamsi/processing/normalization.py* (function *clip_or_mask_intensities*). This method applies lower and/or upper intensity thresholds to an *AnnData* matrix (*adata.X* or *adata.layers[layer]*) and writes the processed matrix back to *adata.X* or *adata.layers[output_layer]*. The caller specifies *low* and/or *high* thresholds and selects discrete policies for values outside bounds: *low_action ∈ {nan, zero, clip, keep, move}* and *high_action ∈ {clip, nan, keep}*. Operationally, the method performs elementwise threshold tests and either (i) replaces out-of-range values by *NaN* or *0*, (ii) clips them to the boundary, or (iii) in *move* mode subtracts *low* and floors at 0 (background-shift), as implemented in the normalization backend. The method supports dense and sparse matrices; for sparse data, only explicitly stored entries are transformed, while implicit zeros remain zero, which can materially affect behavior when thresholds interact with zeros. The function records a provenance entry in *adata.uns["intensity_clipping"]* (the exact schema is defined in the backend), and optionally returns a copied *AnnData* object if *copy=True*. Time complexity is linear in the number of processed matrix entries: **O(n_obs·n_vars)** for dense matrices and **O(nnz)** for sparse matrices, with memory overhead dominated by whether a copy or new layer is created.

log1p_intensity — *Code reference: src/deltamsi/core/msicube.py* (method *MSICube.log1p_intensity*; line ranges unspecified). Underlying implementation: *src/deltamsi/processing/normalization.py* (function *log1p_intensity*). This method performs variance-stabilizing log transformation of intensities using *log(1+x)* applied to the selected matrix (*adata.X* or *layers[layer]*), optionally converting to a different base via division by *ln(base)* when *base* is provided. The method is designed to preserve sparsity where possible by applying the transform to stored values for sparse matrices (implementation detail in the backend). Output is written to *adata.X* or *layers[output_layer]* (source layer is not modified), and transformation metadata are stored in *adata.uns* (key and schema defined by the backend). Complexity is **O(n_obs·n_vars)** for dense and **O(nnz)** for sparse matrices; memory overhead depends on whether a new layer/copy is created.

MSI-specific image-space preprocessing

msi_cap_hotspots — *Code reference:* `src/deltamsi/core/msicube.py` (method `MSICube.msi_cap_hotspots`; line ranges unspecified). Underlying implementation: `src/deltamsi/processing/preprocessing.py` (function `msi_cap_hotspots`). This method reduces the influence of isolated "hot pixels" by capping each ion image (each feature column) at its per-feature quantile q (default **0.99**). The preprocessing backend first densifies the intensity matrix to a floating array (default dtype **float32**), then processes features in blocks (`chunk_size`, default **256**) to control memory. For each block, it computes `caps_j = quantile(X[:,j], q)` and applies `X[:,j] ← min(X[:,j], caps_j)` in-place. The processed matrix is stored in `adata.X` or `layers[output_layer]`. This is consistent with MSI colocalization workflows that use quantile capping/thresholding to mitigate hotspots prior to similarity computation[4]. Computational cost is dominated by per-column quantile computation and clipping: roughly **O(n_obs·n_vars)** with quantile selection costs depending on NumPy's quantile algorithm; memory is **O(n_obs·n_vars)** because the implementation explicitly densifies and copies the matrix.

msi_threshold_quantile — *Code reference:* `src/deltamsi/core/msicube.py` (method `MSICube.msi_threshold_quantile`; line ranges unspecified). Underlying implementation: `src/deltamsi/processing/preprocessing.py` (function `msi_threshold_quantile`). This method performs per-feature background thresholding by setting all pixels below a chosen quantile q (default **0.5**) to either `0` (`mode="zero"`, default) or `NaN` (`mode="nan"`). The backend densifies the matrix, iterates over features in chunks (`chunk_size=256` by default), computes a per-feature threshold vector `thr_j = quantile(X[:,j], q)`, then masks each block by `X < thr` and assigns either `0.0` or `NaN`. This mirrors MSI preprocessing patterns used in cosine-based ion colocalization pipelines[4]. Complexity follows the same scaling as hotspot capping: approximately **O(n_obs·n_vars)** with densification and quantile computation.

msi_median_filter_2d — *Code reference:* `src/deltamsi/core/msicube.py` (method `MSICube.msi_median_filter_2d`; line ranges unspecified). Underlying implementation: `src/deltamsi/processing/preprocessing.py` (function `msi_median_filter_2d`). This method applies a spatial 2D median filter to each ion image to reduce salt-and-pepper noise while largely preserving edges. The backend reconstructs each ion image on a dense `(H×W)` grid using pixel coordinates from either `adata.obs[[x_key,y_key]]` or `adata.obsm[spatial_key]` (default keys: `x`, `y`, `spatial`), with an adjustable origin (`origin ∈ {"min","zero"}`) and optional explicit `shape=(H,W)`. Non-rectangular sampling regions are supported by inserting observations into a full grid initialized with `fill_value` (default **0.0**). Filtering is performed with `scipy.ndimage.median_filter` using a per-feature window of size `(size×size)` (default **3×3**) and `mode="nearest"`. Median filtering is a classical nonlinear denoising operator; fast 2D median filtering algorithms are well established in the image-processing literature, and median filtering is also explicitly referenced in MSI colocalization workflows[4]. Runtime depends on grid size and filter window; a conservative bound is **O(n_vars·H·W·size²)**, though actual SciPy performance uses optimized internal routines. Memory usage is significant because the method densifies and materializes `(n_block×H×W)` blocks.

Spectrum-level normalization and scaling

normalize_tic — *Code reference:* `src/deltamsi/core/msicube.py` (method `MSICube.normalize_tic`; line ranges unspecified). Underlying implementation: `src/deltamsi/processing/normalization.py` (function `tic_normalize_msicube`). This method performs Total Ion Current (TIC) normalization, scaling each pixel spectrum (each observation row) so that the row sum equals a specified `target_sum` (default **1e6**). TIC normalization is a widely used MSI preprocessing step to reduce multiplicative technical variability (e.g., variable ionization efficiency or matrix/tissue effects), though its appropriateness depends on whether total signal is expected to be stable across space. In the backend implementation, the TIC vector is computed as the per-row sum (dense or sparse), and each row is scaled by `target_sum / TIC_i` when TIC is positive. Optionally, the original TIC is stored in `adata.obs[store_tic_in_obs]` (default key `"tic"`). Output can be written to `adata.X` or `layers[output_layer]`; with `copy=True`, a modified cube is returned. Complexity is **O(n_obs·n_vars)** for dense matrices and **O(nnz)** for sparse matrices, with negligible additional memory beyond the output matrix/layer.

scale_zscore — *Code reference:* `src/deltamsi/core/msicube.py` (method `MSICube.scale_zscore`; line ranges unspecified). Underlying implementation: `src/deltamsi/processing/normalization.py` (function `scale_ion_images_zscore`). This method standardizes each feature's spatial intensity vector using z-score scaling, optionally within grouping strata. The user specifies `mode ∈ {"all","per_sample","per_condition"}` (default `"all"`), which controls whether mean/standard deviation statistics are computed across all pixels, separately per `adata.obs["sample"]`, or separately per `adata.obs["condition"]`. Standardization is controlled by `with_mean` and `with_std` (both default **True**), numerical stability uses `eps` (default **1e-8**), and degrees-of-freedom adjustment is supported via `ddof` (default **0**). Optional clipping to `max_value` bounds standardized values. Results are written to `adata.X` or `layers[output_layer]`, and the method can return group-wise statistics if `return_stats=True`. Complexity is linear in processed entries (dense **O(n_obs·n_vars)**; sparse **O(nnz)**) plus overhead for group splitting.

Kendrick-space analysis and annotation

compute_kendrick — *Code reference:* `src/deltamsi/core/msicube.py` (method `MSICube.compute_kendrick`; line ranges unspecified). Underlying implementation: `src/deltamsi/processing/kendrick.py` (function `compute_kendrick_varm`). This method computes Kendrick coordinates from the feature m/z axis stored in `adata.var[mz_key]` (default `"mz"`) and writes the resulting two-dimensional embedding to `adata.varm`. The transformation is parameterized by a repeating unit `base`, which can be supplied as a chemical formula string (for example `"CH2"`), as an exact mass, or as an `(exact_mass, nominal_mass)` tuple. For each feature, the backend computes the Kendrick mass and a Kendrick mass defect-like coordinate according to `kmd_mode ∈ {"fraction", "defect"}`, then stores the resulting `(n_features × 2)` array under `varm_key` (or an automatically generated key if none is provided). Optionally, when `store_1d_in_var=True`,

the one-dimensional Kendrick mass and Kendrick mass defect vectors are also written to *adata.var* using the provided *var_prefix*. In the *MSICube* workflow, this method provides a feature-space representation tailored to homologous-series analysis and is intended to support downstream visual inspection, manual annotation, and plotting. Runtime is effectively linear in the number of features, **O(n_vars)**, with memory overhead dominated by storing the *(n_vars × 2)* coordinate matrix and any optional one-dimensional columns.

annotate_kendrick — *Code reference:* *src/deltamsi/core/msicube.py* (method *MSICube.annotate_kendrick*; line ranges unspecified). Underlying implementation: *src/deltamsi/plotting/kendrick_manual_label.py* (function *manual_label_vars_from_kendrick*). This method launches an interactive Plotly/IPyWidgets interface for manual labeling of features in Kendrick space. It requires precomputed Kendrick coordinates stored in *adata.varm[varm_key]* and initializes or updates a label column in *adata.var[label_key]* (default *"manual_label"*) with a *default_label* (default *"unlabeled"*). The user-facing interface supports point selection in Kendrick space using a configurable *dragmode* (default *"lasso"*), tooltip display using *mz_key*, and visual tuning of *point_size*, *height*, and a warning threshold for large point clouds (*max_points_warn*). The method returns the interactive UI object together with a mutable state object that mediates label assignment back to the *AnnData* table. In the *MSICube* API, this function is therefore not a numerical transform but an annotation utility for expert-guided curation of homologous series or chemically related feature groups. Computational cost is dominated by preparing and rendering the plotted points, which is typically **O(n_vars)**, while practical performance depends strongly on the number of displayed features and the notebook environment.

plot_kendrick — *Code reference:* *src/deltamsi/core/msicube.py* (method *MSICube.plot_kendrick*; line ranges unspecified). Underlying implementation: *src/deltamsi/plotting/plot_kendrick_cluster_mz.py* (function *plot_kendrick_from_clustering*). This method generates a Kendrick plot for the m/z features stored in *adata.var*, optionally colored by clustering results or annotation labels. The plotting behavior is controlled through a *KendrickPlotOptions* object, with keyword arguments allowed to override individual fields before validation. The method retrieves masses from *adata.var[options.mass_col]*, retrieves cluster labels from *adata.var["mass_cluster"]* unless a label-based single-panel plot is explicitly requested, optionally retrieves family annotations from *options.label_col*, *adata.var["family"]*, or *adata.var["label"]*, and then forwards these arrays to the rendering backend. Depending on *options.two_panels*, the function can generate either a single Kendrick-space panel or a two-panel representation comparing the full dataset with selected clusters/labels. The return value is a *(figure, axes, dataframe)* tuple, where the dataframe contains the processed coordinates and grouping variables used for plotting. Within the *MSICube* workflow, this method is the main static visualization endpoint for Kendrick-space structure after feature extraction and optionally after graph-based mass clustering or manual labeling. Runtime is approximately **O(n_vars)** for coordinate preparation and plotting-data assembly, with practical cost dominated by Matplotlib rendering and annotation options for large feature sets.

Feature definition and matrix construction

Mean spectra computation

_compute_all_mean_spectra *— Code reference: src/deltamsi/core/msicube.py* (method *MSICube._compute_all_mean_spectra*; line ranges unspecified). Underlying implementation: *src/deltamsi/processing/mean_spectrum.py* (function *compute_mean_spectrum*). This internal helper is the sample-level backend used by *compute_mean_spectra(scope="samples")*. It ensures that *self.adata* exists, constructs and validates a *MeanSpectrumOptions* object from the provided keyword arguments, iterates over every discovered imzML sample in *org_imzml_path_dict*, computes one mean spectrum per sample, and stores the resulting *{sample: {"mz", "intensity"}}* mapping in *adata.uns["mean_spectra"]* together with option provenance in *adata.uns["mean_spectra_options"]* and the processed sample list in *adata.uns["mean_spectra_samples"]*. In the *MSICube* workflow, this is the dedicated per-sample spectral summarization step that precedes global combination and peak picking. Runtime is approximately linear in the total number of spectral peaks read across all samples, plus the cost of any centroid smoothing requested by the options.

_compute_global_mean_spectrum *— Code reference: src/deltamsi/core/msicube.py* (method *MSICube._compute_global_mean_spectrum*; line ranges unspecified). Underlying implementation: *src/deltamsi/processing/combine_mean_spectra.py* (function *combine_mean_spectra*). This internal helper is the global-combination backend used by *compute_mean_spectra(scope="global")*. It requires previously stored per-sample spectra in *adata.uns["mean_spectra"]*, builds and validates a *GlobalMeanSpectrumOptions* object, reconstructs the list of *(mz, intensity)* sample spectra, interpolates them onto a common grid, optionally TIC-normalizes each interpolated spectrum, averages them into a single consensus spectrum, and stores the result in *adata.uns["mean_spectrum_global"]* with options saved in *adata.uns["mean_spectrum_global_options"]*. Within the overall workflow, this method creates the reference spectrum used for downstream project-wide peak picking. Runtime is driven by interpolation and averaging across all sample mean spectra and scales roughly with the number of samples times the length of the common grid.

compute_mean_spectra *— Code reference: src/deltamsi/core/msicube.py* (method *MSICube.compute_mean_spectra*; line ranges unspecified). Underlying implementations: *src/deltamsi/processing/mean_spectrum.py* (*compute_mean_spectrum*) and *src/deltamsi/processing/combine_mean_spectra.py* (*combine_mean_spectra*). This method computes project-level mean spectra in two modes controlled by *scope*. With *scope="samples"* (default), the method requires *mode ∈ {"profile","centroid"}* and computes one mean spectrum per imzML sample. The backend iterates over all pixel spectra from each imzML (via pyimzML), applies per-pixel TIC normalization (divide by spectrum sum), bins intensities onto a fixed m/z grid defined by *min_mz* (default **0.0**), *max_mz* (default **2000.0**), and *binning_p* (default **0.001**), and accumulates normalized intensity into a “spike” vector representing the mean spectrum. For centroid data, an additional Gaussian smoothing step is

applied either with ppm-scaled width (*mass_accuracy_ppm*, default **3**) or constant width (*tolerance_da* if provided), reflecting typical differences between profile and centroid acquisition. Results are stored in *adata.uns["mean_spectra"]* with provenance in *adata.uns["mean_spectra_options"]*. With *scope="global"*, the method combines existing per-sample mean spectra into a single global spectrum by interpolating each spectrum onto a common grid (*binning_p* default **0.0001**), optionally restricting to the intersection of m/z ranges (*use_intersection=True*), optionally TIC-normalizing each interpolated spectrum (*tic_normalize=True*), averaging across samples, and optionally compressing bins with zero intensity (*compress_axis=False*). This design leverages imzML for raw data access[1] and pyimzML for parsing and spectrum iteration, and stores results in an *AnnData* structure consistent with Scanpy/AnnData conventions[2]. Time complexity is approximately linear in the total number of peaks read across the dataset plus interpolation costs for the global step; memory is dominated by storing binned spectral arrays and, for global combination, the common grid.

Peak picking and matrix extraction

pick_peaks *— Code reference:* *src/deltamsi/core/msicube.py* (method *MSICube.pick_peaks*; line ranges unspecified). Underlying implementation: *src/deltamsi/processing/peak_picking.py* (function *peak_picking*). This method defines study-wide MSI features by peak picking on the global mean spectrum stored in *adata.uns["mean_spectrum_global"]*. It constructs a *PeakPickingOptions* object from keyword arguments (*topn* default **10000**, *binning_p* default **0.0001**, and optional *distance_da*/*distance_ppm* separation constraints), validates it, and then calls *peak_picking*. The backend first interpolates the (potentially irregular) mean spectrum onto a regular axis with step *binning_p*, detects candidate local maxima using *scipy.signal.find_peaks*, and then applies a greedy non-maximum suppression rule: candidates are considered in descending intensity order and accepted only if they are not within the specified Da and/or ppm distance thresholds of any already selected peak. The final selected peak m/z values are sorted ascending and stored in *adata.var["mz"]*, with an indexed feature identifier format *mz_{idx}_{m:.4f}*. If the number of selected peaks differs from the current *adata.n_vars*, the method reconstructs the *AnnData* object with a zero-filled placeholder *X* of shape *(n_obs × n_new_vars)*, preserving *obs*, *uns*, and *obsm*, and warns that the real intensity matrix must be populated by subsequent extraction. Time complexity is driven by the length of the interpolated axis and candidate peaks (peak detection is roughly linear in axis length, while greedy selection may be up to **O(n_candidates·topn)** in the worst case due to repeated distance checks), and memory is dominated by the interpolated arrays and the (re)constructed *AnnData*.

extract_matrix *— Code reference:* *src/deltamsi/core/msicube.py* (method *MSICube.extract_matrix*; line ranges unspecified). Underlying implementation: *src/deltamsi/processing/peak_picking.py* (function *extract_peak_matrix*). This method constructs the final dense pixel × feature matrix *adata.X* by extracting intensities around each target feature m/z stored in *adata.var["mz"]*. It instantiates *PeakMatrixOptions* from *tol_da* and/or *tol_ppm* arguments and validates them (detailed

validation rules are defined in the options class; behavior when both tolerances are omitted is unspecified in *msicube.py* and should be checked in *PeakMatrixOptions.validate*). For each sample imzML, it calls *extract_peak_matrix(imzml_path, target_mzs, options)*, which iterates all spectra via pyimzML[3] and, for each pixel spectrum, assigns to each target m/z the maximum intensity among experimental peaks within a tolerance window (ppm- or Da-defined) using binary search (*searchsorted*) on the sorted target array. Extracted per-sample matrices are vertically concatenated to form *X_all*, coordinate arrays are concatenated to form *adata.obsm["spatial"]*, and sample provenance is stored as a categorical *adata.obs["sample"]*. A copy of the extracted matrix is saved to *adata.layers["RAW"]* for provenance. This procedure directly operationalizes the imzML → matrix conversion described by the imzML standardization effort[1] and is aligned with common MSI workflows that extract feature-aligned matrices for multivariate analysis. Complexity is approximately **O(Σ_pixels Σ_peaks_per_pixel · log n_targets)** plus the cost of concatenation; memory is **O(n_total_pixels·n_features)** due to constructing a dense matrix.

create_feature_matrix — *Code reference: src/deltamsi/core/msicube.py* (method *MSICube.create_feature_matrix*; line ranges unspecified). This is a high-level convenience wrapper that calls *pick_peaks(**peak_picking_kwargs)* followed by *extract_matrix(**extraction_kwargs)* to produce a fully populated *AnnData* object with *adata.var["mz"]* and *adata.X* aligned to the selected peaks. Inputs are optional dictionaries forwarded to the underlying methods; outputs are stored in-place in *self.adata*. Computational cost is the sum of the peak picking and extraction stages, and the method inherits their assumptions (availability of a global mean spectrum for peak picking; availability and validity of extraction tolerances for matrix extraction).

Spatial similarity, aggregation, matching, and differential analysis

Cosine colocalization and aggregation

compute_cosine_colocalization — *Code reference: src/deltamsi/core/msicube.py* (method *MSICube.compute_cosine_colocalization*; line ranges unspecified). Underlying implementation: *src/deltamsi/processing/colocalization.py* (function *compute_mz_cosine_colocalization*). This method computes a feature–feature spatial similarity matrix using cosine similarity between ion images (columns of the intensity matrix). The backend first computes each feature's L2 norm and normalizes columns to unit length, then computes $S = X^T X$ on the normalized matrix. Two computation modes exist: *"dense"* returns a full square NumPy matrix, while *"topk_sparse"* computes similarities blockwise and retains only the top-*k* neighbors per ion (plus an optional minimum similarity threshold), returning a CSR sparse matrix. This design is intended to scale to large MSI feature spaces where a full **O(n_vars²)** dense similarity matrix is impractical. The result is stored in *adata.varp[params.store_varp_key]* along with a parameter log in *adata.uns*. Cosine-based colocalization scores are explicitly used in MSI colocalization methods such as ColocML, typically after preprocessing steps like hotspot removal, thresholding, and median filtering[4]. Dense mode has time complexity approximately **O(n_obs·n_vars²)** and memory **O(n_vars²)**;

sparse top-k mode reduces memory and can reduce effective runtime, but still requires blockwise matrix multiplies whose cost depends on matrix sparsity and chunking.

aggregate_vars_by_label — *Code reference:* `src/deltamsi/core/msicube.py` (method `MSICube.aggregate_vars_by_label`; line ranges unspecified). Underlying implementation: `src/deltamsi/processing/aggregation.py` (function `aggregate_vars_by_label`). This method collapses multiple m/z features into grouped "component images" based on a categorical variable label stored in `adata.var[label_col]` (e.g., chemical family, cluster assignment). For each unique label, it selects the corresponding feature subset and computes an aggregation statistic per pixel: `mean`, `max`, or `median` (median is explicitly unsupported for sparse matrices in the backend, avoiding implicit densification). The aggregated matrix has shape `(n_pixels, n_labels)` and is stored in `adata.obsm[obsm_key]` either as a NumPy array or a pandas DataFrame (`as_df=True`). Provenance (source label column, layer, aggregation mode) is recorded in `adata.uns`. Complexity is approximately linear in the total data traversed across groups (dense **O(n_obs·n_vars)**; sparse **O(nnz)**), and memory cost is **O(n_obs·n_labels)** for the result matrix.

m/z matching utilities

match_mzs_to_var_simple — *Code reference:* `src/deltamsi/core/msicube.py` (method `MSICube.match_mzs_to_var_simple`; line ranges unspecified). Underlying implementation: `src/deltamsi/processing/mz_matching.py` (function `match_mzs_to_var_simple`). This method matches user-provided query m/z values against the feature m/z axis stored in `adata.var[mz_col]` (default `"mz"`), either by selecting the closest feature (`mode="closest"`) or by returning features within a tolerance window (`mode="tolerance"` with `tol` default **5.0** in units *ppm* or *da*). The backend sorts the feature m/z axis (unless `assume_sorted=True`) and uses binary search (`np.searchsorted`) to identify candidate matches efficiently. In tolerance mode, it can return all matches within tolerance (`return_all_within_tol=True`) or only the best match, reporting errors in both Da and ppm. The method can also annotate matched features by writing to an `annotation_col` in `adata.var` with either overwrite or append semantics. Complexity is **O(n_queries·log n_vars + n_returned_matches)** with low memory overhead beyond the results DataFrame.

Differential ranking of ions

rank_ions_groups_msi — *Code reference:* `src/deltamsi/core/msicube.py` (method `MSICube.rank_ions_groups_msi`; line ranges unspecified). Underlying implementation: `src/deltamsi/processing/discriminant_analysis.py` (function `rank_ions_groups_msi`). This method ranks ions by differential abundance between two groups defined in `adata.obs` (e.g., treated vs. control) using a parameter object `RankIonsMSIParams`. The backend computes effect sizes on pooled pixels (log2 fold change with a pseudocount and delta detection rate above a detection threshold) and adapts statistical testing to the study design. When multiple biological replicates exist for both groups (≥2 samples per group), the method aggregates pixels by sample ("pseudobulk") and applies either Welch's t-test or a Mann–Whitney U test (depending on `method` and replicate counts), followed

by Benjamini–Hochberg FDR correction for multiple testing. For single-sample designs (no replication), the implementation emphasizes effect sizes and can optionally compute spatial block-bootstrap confidence intervals by resampling square spatial tiles, acknowledging spatial autocorrelation in MSI pixels. Optional AUC computation is supported for a limited number of ions using Mann–Whitney U statistics. Results are returned as a ranked DataFrame (top-N) and stored in *adata.uns[params.key_added]* with arrays for downstream visualization. The method's complexity is dominated by computing group means/detection rates (**O(n_obs·n_vars)**) plus any per-feature statistical testing and bootstrap loops (which can be substantial when nonparametric tests or many bootstrap resamples are enabled); memory overhead is primarily the storage of result tables and optional sampled matrices.

Spatial chaos metrics

Spatial chaos metrics quantify the spatial "structuredness" of ion images across thresholds and are used in MSI pipelines to filter uninformative/noisy ions and support annotation scoring[5].

compute_spatial_chaos_scores *— Code reference:* *src/deltamsi/core/msicube.py* (method *MSICube.compute_spatial_chaos_scores*; line ranges unspecified). Underlying implementation: *src/deltamsi/processing/spatial_chaos.py* (*compute_spatial_chaos_matrix*, *spatial_chaos_score*). This method computes a per-ion, per-sample score by reconstructing each ion image on a 2D grid from *adata.obsm[obsm_key]* (default *"spatial"*) and applying a level-set fragmentation metric across *n_thresholds* intensity thresholds (default **30**). The backend normalizes each image to *[0,1]* by its max intensity, then for each threshold constructs a binary mask and computes: (i) *P_k*, the number of positive pixels, and (ii) *C_k*, the number of 4-connected components (clusters). The returned score is *S = 1 − (Σ C_k) / (Σ P_k)*, clamped to *[0,1]*, yielding higher scores for spatially coherent images (few components per positive pixel) and lower scores for fragmented/noisy signals. The resulting matrix of shape *(n_features, n_samples)* is stored in *adata.varm[varm_key]* (default *"spatial_chaos"*) with metadata in *adata.uns["spatial_chaos"]*. The conceptual basis—quantifying level-set complexity/fragmentation as a measure of ion-image "chaoticness" or structure—appears in MSI literature and in annotation-quality pipelines such as METASPACE[5,8]. Runtime is potentially high: in the worst case **O(n_samples·n_vars·n_thresholds·H·W)** due to repeated connected-component labeling over a full grid per ion.

compute_spatial_chaos_fold_change *— Code reference:* *src/deltamsi/core/msicube.py* (method *MSICube.compute_spatial_chaos_fold_change*; line ranges unspecified). Underlying implementation: *src/deltamsi/processing/spatial_chaos.py* (*spatial_chaos_fold_change_from_adata*, *spatial_chaos_fold_change*). This method compares spatial-chaos scores between two experimental groups defined by an *adata.obs[groupby]* column. It requires precomputed per-sample chaos scores stored in *adata.varm[varm_key]* (default *"spatial_chaos"*). The backend maps each sample to a unique group label, extracts the maximum chaos score per ion within each group (*S_control_max*, *S_interaction_max*), and computes a structure-based fold change *FC_S*

= S_interaction_max / max(S_control_max, eps) with *eps* default **1e-6**. The results are stored in *adata.var* under keys derived from *result_key* (default *"spatial_chaos_fold_change"*) and metadata are stored in *adata.uns[result_key]*. This groupwise comparison aligns with MSI use cases where ions can become more or less spatially structured between biological conditions[5]. Runtime is **O(n_vars·n_samples)**, negligible relative to initial chaos-score computation.

Recalibration and mass clustering

Mass recalibration

recalibrate *— Code reference: src/deltamsi/core/msicube.py* (method *MSICube.recalibrate*; line ranges unspecified). Underlying implementations: *src/deltamsi/processing/recalibration_core.py* and *src/deltamsi/processing/recalibration_cli_clean.py*. This method performs pixel-wise m/z recalibration for each imzML dataset, writing new calibrated imzML files and updating *MSICube* to point to the recalibrated paths. The pipeline is: (i) load a sorted unique database of theoretical reference masses from a text/CSV file; (ii) for each pixel spectrum, select the top *n_peaks* most intense peaks, match peaks to database masses within tolerance using binary search, and collect mass errors; (iii) estimate the dominant systematic error as the mode of the error distribution via kernel density estimation (KDE); (iv) restrict matches to a region-of-interest around that mode; (v) fit a robust linear error model across m/z using RANSAC regression; and (vi) subtract the predicted error from each m/z value and write the corrected spectrum to a new imzML. Robust model fitting via RANSAC is classically motivated by the need to fit in the presence of outliers. The implementation includes explicit quality filters (e.g., skipping pixels with too few peaks or too few database hits, and skipping pixels where an error mode or model cannot be fitted), which implies **the recalibrated output can contain fewer spectra than the input** (skipped pixels are not written by the writer). Parallel processing across samples is implemented via *ProcessPoolExecutor* when *n_workers>1*. Complexity depends on the number of pixels, the number of peaks retained, and database size; per pixel, matching is roughly **O(n_peaks·log n_db)**, KDE scales with number of hits, and RANSAC scales with iterations and inlier evaluation. The method is intended to improve cross-sample feature alignment before global peak picking, consistent with standard MSI practice where calibration drift can impede feature matching across acquisitions[1].

Graph-based mass clustering

cluster_masses *— Code reference: src/deltamsi/core/msicube.py* (method *MSICube.cluster_masses*; line ranges unspecified). Underlying implementation: *src/deltamsi/processing/mass_clustering.py* (*cluster_masses_with_candidates*, *cluster_masses_from_colocalization*). This method groups m/z features into “chemical families” using graph community detection. Two graph construction modes are supported via *MassClusteringOptions.method*: *In candidate-difference mode (method="candidates"):* each feature is a node; an undirected edge connects two nodes *(i,j)* if *|mz_i−mz_j|* matches a known candidate mass difference from an input catalog (*candidates_df*) within a specified tolerance (Da or ppm). When

multiple candidate differences match, the implementation selects the candidate with the lowest score, breaking ties by smallest absolute mass error. Edge weights are derived from candidate scores via a configurable transform (default *`1/(1+score)`*), and optional pruning retains only top-k edges per node in either union or mutual k-NN mode. *In colocalization mode (`method="colocalization"`):* edges are built directly from a precomputed similarity matrix (e.g., cosine colocalization), with optional similarity thresholding and k-NN pruning; edge weights are the similarity values. In both cases, Leiden community detection (RBConfigurationVertexPartition) is applied to the resulting weighted graph to infer clusters; singleton clusters are mapped to label *−1*. Leiden is a well-established improvement over earlier modularity-based methods and is widely used for large-scale graph clustering [8]. *`MSICube.cluster_masses`* stores (i) cluster labels in *`adata.var[opts.output_col]`*, (ii) a symmetric weighted adjacency matrix in *`adata.varp["mass_clustering_weights"]`*, (iii) a binary connectivity matrix in *`adata.varp["mass_clustering_connectivities"]`*, (iv) node degree and weighted degree in *`adata.var`*, and (v) a full edge table with feature-name endpoints and summary statistics in *`adata.uns["mass_clustering"]`*. Candidate-based clustering has worst-case time complexity **O(n_vars²·log n_candidates)** due to pairwise mass differences, while colocalization-based clustering is **O(nnz)** in the similarity matrix (or **O(n_vars²)** if dense). Leiden runtime scales approximately with the number of edges.

References

(1) Römpp, A.; Schramm, T.; Hester, A.; Klinkert, I.; Both, J.-P.; Heeren, R. M. A.; Stöckli, M.; Spengler, B. imzML: Imaging Mass Spectrometry Markup Language: A Common Data Format for Mass Spectrometry Imaging. In *Data Mining in Proteomics*; Hamacher, M., Eisenacher, M., Stephan, C., Eds.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2011; Vol. 696, pp 205–224. https://doi.org/10.1007/978-1-60761-987-1_12.

(2) Virshup, I.; Rybakov, S.; Theis, F. J.; Angerer, P.; Wolf, F. A. Anndata: Access and Store Annotated Datamatrices. *J. Open Source Softw.* **2024**, *9* (101), 4371. https://doi.org/10.21105/joss.04371.

(3) Bald, T.; Barth, J.; Niehues, A.; Specht, M.; Hippler, M.; Fufezan, C. pymzML—Python Module for High-Throughput Bioinformatics on Mass Spectrometry Data. *Bioinformatics* **2012**, *28* (7), 1052–1053. https://doi.org/10.1093/bioinformatics/bts066.

(4) Ovchinnikova, K.; Stuart, L.; Rakhlin, A.; Nikolenko, S.; Alexandrov, T. ColocML: Machine Learning Quantifies Co-Localization between Mass Spectrometry Images. *Bioinformatics* **2020**, *36* (10), 3215–3224. https://doi.org/10.1093/bioinformatics/btaa085.

(5) Palmer, A.; Phapale, P.; Chernyavsky, I.; Lavigne, R.; Fay, D.; Tarasov, A.; Kovalev, V.; Fuchser, J.; Nikolenko, S.; Pineau, C.; Becker, M.; Alexandrov, T. FDR-Controlled Metabolite Annotation for High-Resolution Imaging Mass Spectrometry. *Nat. Methods* **2017**, *14* (1), 57–60. https://doi.org/10.1038/nmeth.4072.

(6) Fischler, M. A.; Bolles, R. C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM* **1981**, *24* (6), 381–395. https://doi.org/10.1145/358669.358692.

(7) Traag, V. A.; Waltman, L.; Van Eck, N. J. From Louvain to Leiden: Guaranteeing Well-Connected Communities. *Sci. Rep.* **2019**, *9* (1), 5233. https://doi.org/10.1038/s41598-019-41695-z.

(8) Alexandrov, T.; Ovchinnikova, K.; Palmer, A.; Kovalev, V.; Tarasov, A.; Stuart, L.; Nigmetzianov, R.; Fay, D.; Key METASPACE contributors; Gaudin, M.; Lopez, C. G.; Vetter, M.; Swales, J.; Bokhart, M.; Kompauer, M.; McKenzie, J.; Rappez, L.; Velickovic, D.; Lavigne, R.; Zhang, G.; Thinagaran, D.; Ruhland, E.; Sans, M.; Triana, S.; Sammour, D. A.; Aboulmagd, S.; Bagger, C.; Strittmatter, N.; Rigopoulos, A.; Gemperline, E.; Joensen, A. M.; Geier, B.; Quiason, C.; Weaver, E.; Prasad, M.; Balluff, B.; Nagornov, K.; Li, L.; Linscheid, M.; Hopf, C.; Heintz, D.; Liebeke, M.; Spengler, B.; Boughton, B.; Janfelt, C.; Sharma, K.; Pineau, C.; Anderton, C.; Ellis, S.; Becker, M.; Pánczél, J.; Violante, G. D.; Muddiman, D.; Goodwin, R.; Eberlin, L.; Takats, Z.; Shahidi-Latham, S. METASPACE: A Community-Populated Knowledge Base of Spatial Metabolomes in Health and Disease. Systems Biology February 3, 2019. https://doi.org/10.1101/539478.

Supporting information of Chapter 3

Adaptive Pixel Mass Recalibration for Mass Spectrometry Imaging Based on Locally Endogenous Biological Signals.

Raphaël La Rocca[*,†], Christopher Kune[†], Mathieu Tiquet[†], Lachlan Stuart[‡], Gauthier Eppe[†], Theodore Alexandrov[‡,§], Edwin De Pauw[†] and Loïc Quinton[†]

[†]Mass Spectrometry Laboratory, MolSys Research Unit, Department of Chemistry, University of Liège, Allée du Six Août, 11 - Quartier Agora, 4000 Liège, Belgium

[‡]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany.

[§] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA, USA.

S1 : Description of the 31 METASPACE MSI data according to the initial uploader

S2 : Full detailed method section

S3 : Plots for the 31 METASPACE MSI dataset (1 to 31) for the number of METASPACE annotations of the unprocessed (blue dots) and recalibrated (orange dots), for different mass tolerances (ppm). The number of annotations is shown for FDR ≤ 5%, ≤ 10%, ≤ 20% and ≤ 50%. See original paper https://doi.org/10.1021/acs.analchem.0c05071.

S4: Recalibration evaluation with *deltamsi*

S1: Description of the 31 METASPACE MSI data according to the initial uploader

Dataset	Group	Uploader	Source	Analyzer	Polarity	Resolving Power	Resolving Power at m/z	Organism	Organism Part	Sample Stabilisati	Tissue Modification	MALDI Matrix
1	ICL	James McKenzie	DESI	Orbitrap	Negative	70000	200	Homo sapiens (human)	Colon	Fresh frozen	None	none
2	ICL	James McKenzie	DESI	Orbitrap	Negative	70000	200	Homo sapiens (human)	Colon	Fresh frozen	None	none
3	ICL	James McKenzie	DESI	Orbitrap	Negative	70000	200	Homo sapiens (human)	Colon	Fresh frozen	None	none
4	ICL	James McKenzie	DESI	Orbitrap	Negative	140000	200	Homo sapiens (human)	Lymph node	Fresh frozen	None	none
5	ICL	Zoltan Takats	DESI	Orbitrap	Negative	140000	200	Homo sapiens (human)	Colon	None	None	none
6	JLU Giessen	Dhaka Bhandari	MALDI	Orbitrap	Positive	45000	200	Mus musculus (mouse)	Brain	Fresh frozen	None	DHB
7	JLU Giessen	Mario Kompauer	MALDI	Orbitrap	Positive	60000	200	Mus musculus (mouse)	Brain	Dehydrated	Frozen	DHB
8	M4I	Shane Ellis	MALDI	Orbitrap	Positive	240000	400	Mus musculus (mouse)	Brain	None	None	Norharmane
9	PNNL	Dusan Velickovic	MALDI	FTICR	Negative	120000	400	Homo sapiens (human)	kidney	Fresh frozen	Frozen	Norharmane
10	PNNL	Dusan Velickovic	MALDI	FTICR	Negative	120000	400	Homo sapiens (human)	Kidney	Fresh frozen	None	9AA
11	PNNL	Dusan Velickovic	MALDI	FTICR	Positive	120000	400	Homo sapiens (human)	Kidney	FFPE	N/A	CHCA
12	PNNL	Dusan Velickovic	MALDI	FTICR	Positive	120000	400	Bacillus subtilis	Microbial colony	N/A	N/A	DHB/CHCA
13	PNNL	Dusan Velickovic	MALDI	FTICR	Positive	150000	700	Mus musculus (mouse)	Lung	N/A	N/A	CHCA
14	PNNL	Jessica Lukowski	MALDI	FTICR	Negative	120000	400	Homo sapiens (human)	Kidney	None	None	Norharmane
15	U Rennes 1	LAVIGNE Régis	MALDI	FTICR	Negative	130000	400	Mus musculus (mouse)	Brain	Fresh frozen	Ammonium Acetate wash	DHB
16	U Rennes 1	LAVIGNE Régis	MALDI	FTICR	Positive	130000	400	Mus musculus (mouse)	Brain	Fresh frozen	None	CHCA
17	U Rennes 1	LAVIGNE Régis	MALDI	FTICR	Positive	130000	400	Rattus norvegicus (rat)	Kidney	Fresh frozen	None	DHB
18	U Rennes 1	LAVIGNE Régis	MALDI	FTICR	Positive	130000	400	Mus musculus (mouse)	Brain	Fresh frozen	None	CHCA
19	U Rennes 1	LAVIGNE Régis	MALDI	FTICR	Positive	130000	400	Brassica napus (rapeseed)	Root	Native	None	DHB
20	U Rennes 1	LAVIGNE Régis	MALDI	FTICR	Positive	130000	400	Brassica napus (rapeseed)	Stem	Fresh frozen	None	DHB
21	UTHSA	Annapurna Pamreddy	MALDI	Orbitrap	Negative	120000	200	Mus musculus (mouse)	Kidney	Fresh frozen	None	DAN
22	UTHSA	Annapurna Pamreddy	MALDI	Orbitrap	Negative	120000	200	Homo sapiens (human)	Kidney	Fresh frozen	None	DAN
23	UTHSA	Annapurna Pamreddy	MALDI	Orbitrap	Positive	120000	200	Mus musculus (mouse)	Kidney	Fresh frozen	None	DHB
24	UTHSA	Annapurna Pamreddy	MALDI	Orbitrap	Positive	120000	200	Homo sapiens (human) | Mus musculus (mouse)	Brain	Fresh frozen	None	DHB
25	UTHSA	Guanshi Zhang	DESI	Orbitrap	Negative	60000	200	Mus musculus (mouse)	Kidney	Fresh frozen	None	none
26	Waters	Bindesh Shrestha	DESI	Synapt XS	Positive	20000	200	Mus musculus (mouse)	Brain	None	Frozen	none
27	Waters	Bindesh Shrestha	DESI	Synapt XS	Positive	40000	200	Mus musculus (mouse)	Brain	None	Frozen	none
28	Waters	Bindesh Shrestha	DESI	Synapt XS	Positive	30000	200	Gallus domesticus (chicken)	Liver	None	Frozen	none
29	Waters	Bindesh Shrestha	MALDI	Synapt	Positive	20000	200	Rattus norvegicus (rat)	Brain	None	Frozen	none
30	Waters	Emrys Jones	DESI	qTOF	Negative	30000	800	Mus musculus (mouse)	Kidney	Fresh frozen	None	none
31	Waters	Emrys Jones	DESI	TOF reflector	Negative	30000	585	Sus scrofa domesticus (pig)	Kidney	None	None	none

.

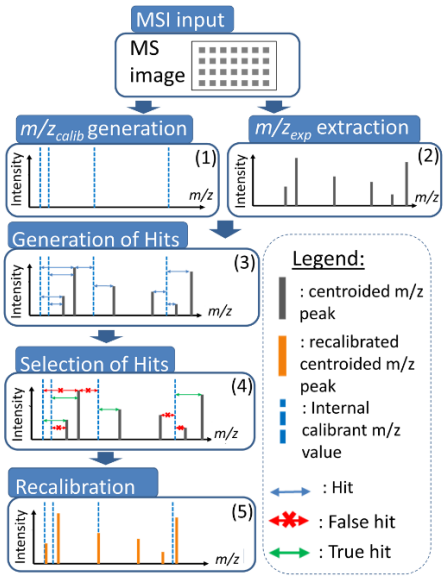
S2: FULL DETAILED METHOD

(1) General approach

In this work, the term “hit” is used to depict a match between experimental signal *m/z* (*m/z*$_{exp}$.) and internal calibrating ion *m/z (m/z*$_{calib.}$*)* within a given mass tolerance from the *m/z*$_{calib.}$ We will call “mass error” the difference in mass between the *m/z*$_{exp}$ and its linked *m/z*$_{calib}$ (Equation 1). A hit is considered as true if the experimental signal corresponds to the calibration ion. Otherwise, an erroneous match between the experimental signal and the internal calibrating ion is a false hit. In MSI, the mass spectra are information-rich and often contain 10e4 peaks per spectrum. Therefore, it is particularly difficult to discriminate true from false hits as multiple signals can be found in the mass range of a given *m/z*$_{calib}$. This difficulty is amplified by the presence of mass shifts since higher mass tolerance is required. However, increasing the mass tolerance inevitably increases the number of false hits, decreasing the recalibration performances. Therefore, the algorithm presented in this work optimizes the selection of true hits for recalibrating each pixel individually. We assume that true hits have mass errors directly correlated to the mass shifts. Then, the calculated mass error is used as a criterion for the selection of hits, considering that hits with the most frequent mass errors are predominantly true. These hits are thus selected to fit a linear model of the mass errors according to *m/z*. Finally, the recalibration is performed by removing the estimated errors in every detected *m/z* values.

This algorithm is divided into 5 steps (Figure 1). (1) The generation of the list of m/z$_{calib.}$ for the whole MSI data according to similar public MSI data sets from METASPACE (i.e. representing the same kind of biological samples) since we assume them to share metabolites with the sample of interest. (2) Centroid MS spectra are extracted from each pixel. From this step, each pixel is processed individually. (3) The calibrant hits are generated by computing the mass errors between the list of potential internal calibrating ions and the spectrum signals for each pixel.

(4) For each pixel individually, the preferential calibrant hits are selected according to their mass errors by discarding hits with infrequent mass errors. (5) A linear model for predicting the mass shifts based on the preferential hits is then constructed to correct every mass of the spectrum.



(2) Generation of the list of internal calibrating ions

Generating the $m/z_{calib.}$ list covering the signals from the sample of interest is a critical step for accurate estimation of mass shifts as it directly influences the collection of true hits. Therefore, for each MSI data subject to recalibration, a list of annotations from similar public METASPACE datasets is generated. Similar METASPACE MSI data are selected based on their metadata which entails the sample type and experimental parameters of the MSI experiment. We search for data with similar acquisition mode, organisms and organs (regardless of the molecular database used in METASPACE for annotation). Once selected, the combined list of annotations is reduced to annotations identified with a false discovery rate [1] (FDR) ≤ 10% and detected in, at least, 10% of considered public datasets. This aims at removing any atypical compound identified only in a few METASPACE MSI datasets, which would have a low chance to be present in the sample of interest. To increase the number of potential calibrating ions we include the two most intense isotopes for each selected calibrating ion. A unique m/z_{calib} list is generated for

recalibrating all the pixels of the MSI data. However, it is not mandatory to use METASPACE to generate the m/z_{calib} as any list of m/z values could work, yet using METASPACE will result in fewer false hits.

(3) Generating the Hits

The m/z_{exp} list is made of the 300 most intense peaks, for each pixel (Figure 1.2). This number is set in advance for all MSI data and is optimized to maximize the true signals over the noisy signals. This selection is not necessary if denoising is already applied but mandatory when no information is known concerning the generation of the centroid MSI (see material). The 300 *m/z* values are then compared to the list of $m/z_{calib.}$, within a mass tolerance called T1 in (Figure 1.3). This mass tolerance window is considered large enough to encompass most of the extreme mass shifts in the data, and small enough to discard the contribution of non-relevant peaks and isotopes. The mass errors in Da, used to calculate the regression and the error distribution, are calculated for each hit using Equation 1 where M_{error}, M_{exact} and $M_{experiment}$ are, respectively, the mass error, the exact mass (from $m/z_{calib.}$ list) and the experimental mass (from the $m/z_{exp.}$ list).

$$M_{error} = M_{experiment} - M_{exact} \qquad \text{Equation 1}$$

M_{error} values are expressed in Da instead of ppm to ensure a linear evolution of M_{error} errors along the m/z axis (See Figure 2A).

(4) Selection of Hits

Since the mass tolerance window for finding true hits T1 (e.g. set at ±0.01 Da for all the MSI data) can be large, a preselection of hits is necessary before fitting any model (Figure 1.4). The calculated mass errors of the hits (with Equation 1) are used for discarding false hits. As discussed previously, the assumption is that true hits display similar mass errors and the most populated errors should therefore contain a maximum of true hits. A kernel density estimation of the distribution of the errors is estimated with the Python library SciPy by a Gaussian Kernel [2] as illustrated in Figure 2.a, the bandwidth is adapted for each pixel by dividing a constant equal to 0.002 by the standard deviation of the computed mass errors. The hits of interest are

finally selected within a certain range from the maximum of this distribution. This range will be called T2 (i.e. set at ±0.002 Da for all the MSI data treated here).

(5) Mass errors estimation and recalibration

Hits selection reduces the probability to select false hits. Unfortunately, the application of simple linear regression function such as OLS (*Ordinary Least Squares*) leads to poor results since false hits are still present due to the complexity of biological signals. Therefore, RANSAC (*Random Sample Consensus*) algorithm from the scikit-learn library (Python) [3] is used for regression as it is more robust to outliers compared to OLS. A minimum of 10 hit (i.e. calibration point) and an upper limit of 300 trials for the random points selection are imposed to increase the chance of performing a correct regression. If this number of hits is met and if RANSAC algorithm is able to perform a linear regression (Figure 2.b), the MS spectrum (MSI pixel) recalibration is performed by removing the mass errors estimated by the linear fit from each detected *m/z* value. This step is made for each pixel of the MSI data. (Figure 1.5)

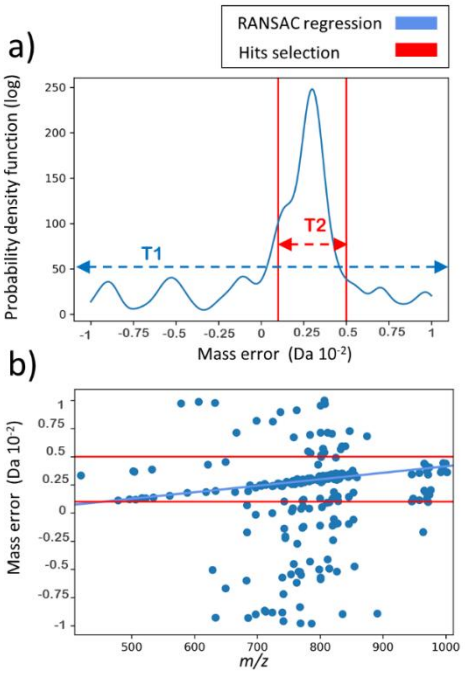


Figure 2. Recalibration algorithm output example from the hits of a single pixel. (a) Selection of hits according to the most frequent errors (b) Mass errors estimation and recalibration.

The Figure 2 highlights the application of our algorithm to a single pixel of the zebrafish MSI data. The selection of hits based on the density estimation of the mass errors is shown in Figure 2.a, the selected hits are the points located within the red lines. The goal of this step is to filter out as many as possible of those hits which have too high error deviations from true hits, to maximize the chance of not including false hits during the model estimation. The RANSAC linear model is then estimated on the selected hits. Since many outliers are still present in the selected hits, the use of a robust linear estimator is necessary for detecting only true hits (Figure 2b).

The reference implementation of the developed method is available at https://github.com/LaRoccaRaphael/MSI_recalibration.

The full method is written in the programming language Python with the following libraries, NumPy, Pandas, metaspace, SciPy, pyimzML and scikit-learn. The processing time is mainly attributable to the recalibration part of the method (step 3 to 5) which depends on the number of pixels in the image. For example, the recalibration of the imzML of the zebrafish, containing 11050 pixels, took 845 seconds on a computer with 16 GB of RAM, 2.6GHz four-core Intel i7 processor with 8MB of cache running Ubuntu.

S4: Recalibration evaluation with *deltamsi*

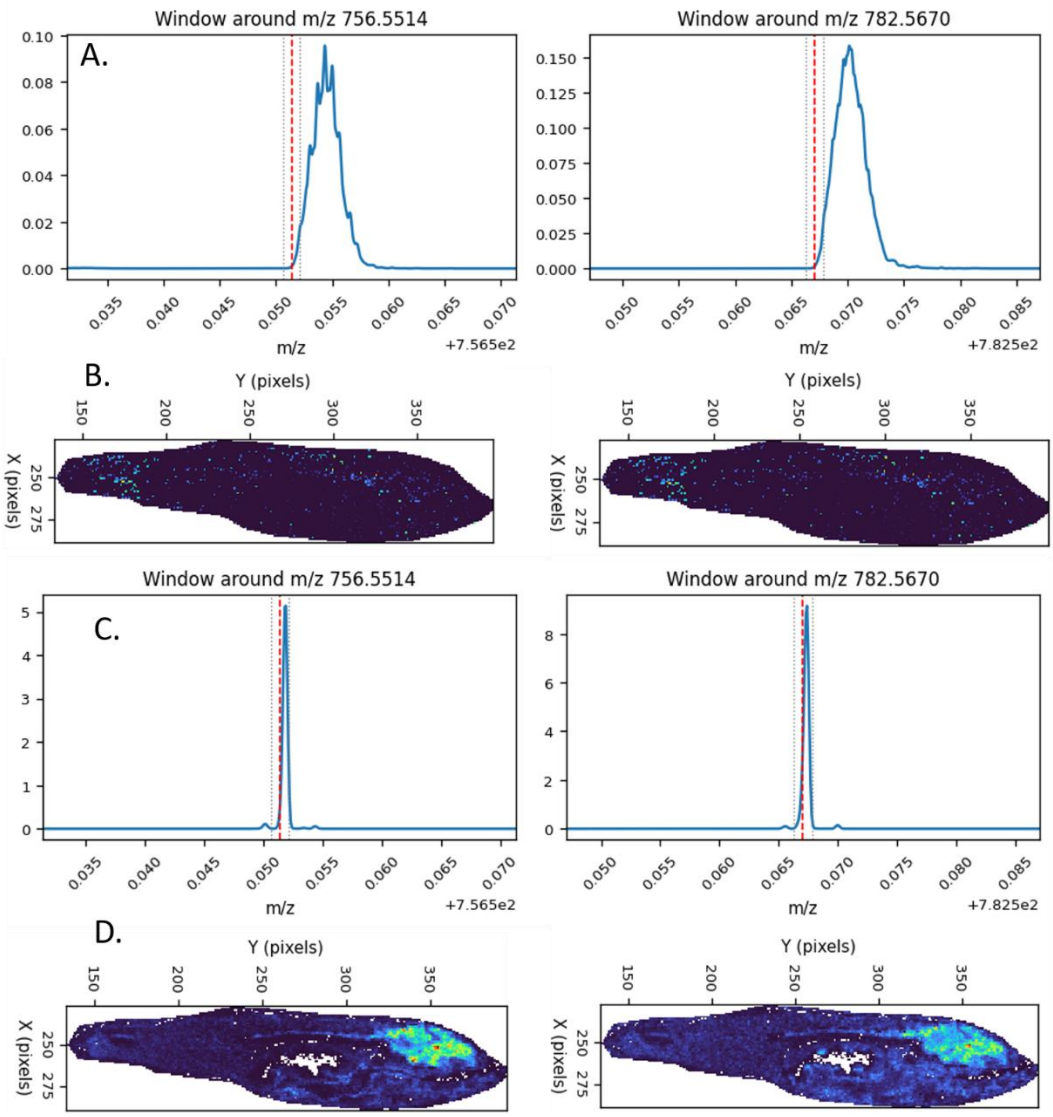


Figure 3. From a zebrafish MSI data. Comparison of the average spectra before and after recalibration (A and C) for two different well-characterized $[C_{40}H_{80}NO_8P+Na]^+$ *and* $[C_{42}H_{82}NO_8P+Na]^+$, with a theoretical *m/z* of 756.5514 *m/z* and 782.5670 *m/z*, respectively (right and left column) (ppm error 0.56 and 0.48, respectively), as well as the consequence on the reconstructed images at ± 1 ppm (B and D). The images are reconstructed from the theoretical masses of the two ions using the *deltamsi* package. White pixels correspond to spectra where not enough hits were found, no calibration was performed, leaving an empty region. The calibration used the same calibrant list as the one used to compute Chapter 3 zebrafish recalibration (Figure 3.4) (excluding all adducts of $C_{40}H_{80}NO_8P$ *and* $C_{42}H_{82}NO_8P$).

Reference

[1] Palmer, A.; Phapale, P.; Chernyavsky, I.; Lavigne, R.; Fay, D.; Tarasov, A.; Kovalev, V.; Fuchser, J.; Nikolenko, S.; Pineau, C.; Becker, M.; Alexandrov, T. FDR-Controlled Metabolite Annotation for High-Resolution Imaging Mass Spectrometry. *Nat Methods* **2017**, *14* (1), 57–60. https://doi.org/10.1038/nmeth.4072.

[2] Ledl, T. Kernel Density Estimation: Theory and Application in Discriminant Analysis. *AJS* **2016**, *33* (3), 267–279. https://doi.org/10.17713/ajs.v33i3.441.

[3] Fischler, M. A.; Bolles, R. C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM* **1981**, *24* (6), 381–395. https://doi.org/10.1145/358669.358692.

Supporting information of Chapter 5

Relational Graph Convolutional Network for Robust Mass Spectrum Classification

Raphaël La Rocca[1], Anthony Cioppa[2], Enrico Ferrarini[3], Monica Höfte[3], Edwin De Pauw[1], Marc Van Droogenbroeck[2], Gauthier Eppe[1] and Loïc Quinton[1*]

[1] Mass Spectrometry Laboratory, MolSys Research Unit, University of Liège, Liège, Belgium

[2] TELIM, Montefiore Institute, University of Liège, Liège, Belgium

[3] Department of Plants and Crops, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium

*E-mail: loic.quinton@uliege.be

S1 : Description and id of the 4 METASPACE MSI data according to the initial uploader and an additional dataset from our laboratory

S2 : Background and Foreground annotation for each selected pixel of the 4 METASPACE MSI

S3 : Set of mass differences of the most frequent mass patterns in MALDI-MSI experiments from the work of Janda M. et al[1]

S4 : Annotated m/z of all dataset and their computed CAM values, Fold change and correlation

S5 : Ion images corresponding to the annotated m/z in S4, Pseudomonas.

S6 : Material and Sample annotations.

S7: R-GCN Implementation Description

S8: Performance comparison of different R-GCN parameters.

S9: 1D-CNN, SVM, RF, LDA, XGBoost pre-processing.

S10: Evaluation of unseen data.

S11: Model evaluation details

S1-S5 See original paper https://doi.org/10.1021/jasms.5c00055.

S6: Material and Sample annotations

S6.1. METASPACE MSI

Four public MSI datasets are selected from METASPACE to showcase different experimental conditions such as different MALDI matrices (2,5-dihydroxybenzoic acid (DHB) and a-Cyano-4-hydroxycinnamic (HCCA)), different HRMS instruments (Orbitrap and FT-ICR), and different organism origins such as human biopsies (lung and kidney) and bacterial colonies. All datasets are downloaded as centroided imzML[2]. We generate a global dataset by randomly sampling 2,500 spectra from each image, which adds up to 10,000 spectra for all METASPACE images. From each spectrum, a mass range from 200-1400 m/z is selected, and 2000 of the most intense centroids are selected.

S6.2. *Pseudomonas* bacterial colonies MSI

Indium tin oxide glass slides (Bruker, Bremen, Germany) are coated, as described by Skinner et al. [3], with 1/5 strength CAA agar medium (1/5 casamino acids (Becton, Dickinson and Company) 1 g/L, MgSO4·7H2O 0.05 g/L, K2HPO4 0.2 g/L, agar (Bacto TM) 1.5 g/L). Slides are deposited in sterile petri dishes on top of a humidified filter paper disc. *Pseudomonas sessilinigenes* CMR12a is cultured overnight on King's B broth. Bacteria are pelleted by centrifugation (4000g, 10 min), resuspended in 10 mM MgSO4, and appropriately diluted. A volume of 10 µL was uniformly distributed on the slide. The inoculated slides are cultured at 18°C for 16-18h, in the dark. Slides are dried in a vacuum chamber for 48 hours at room temperature. The slides are then covered by HCCA matrix and imaged on a SolariX XR 9.4T (Bruker) [4]. A single image is selected and converted into a centroided imzML format, a standard data format for MSI using FlexImaging 5.0. From this image, 3000 pixels are randomly picked. Two mass ranges are selected, generating two different datasets; the first contains 200-2100 m/z, and the other contains 200-1400 m/z. Finally, we select the 2000 most intense centroids in each spectrum.

In the section R-GCN signal interpretability, the MSI is recalibrated using the following method. We recalibrate the masses of each spectrum using the method from La Rocca et al. [5]. The following parameters are used for the recalibration function: -st 0.0005 -tl 0.01 -lm 0.002. The list of m/z values for recalibration can be found in S4 (calibration_mass_pseudo). The dataset can be downloaded from https://metaspace2020.org/api_auth/review?prj=966e90da-5502-11f0-a049-4353c03f6897&token=X5pJU9I26dbd.

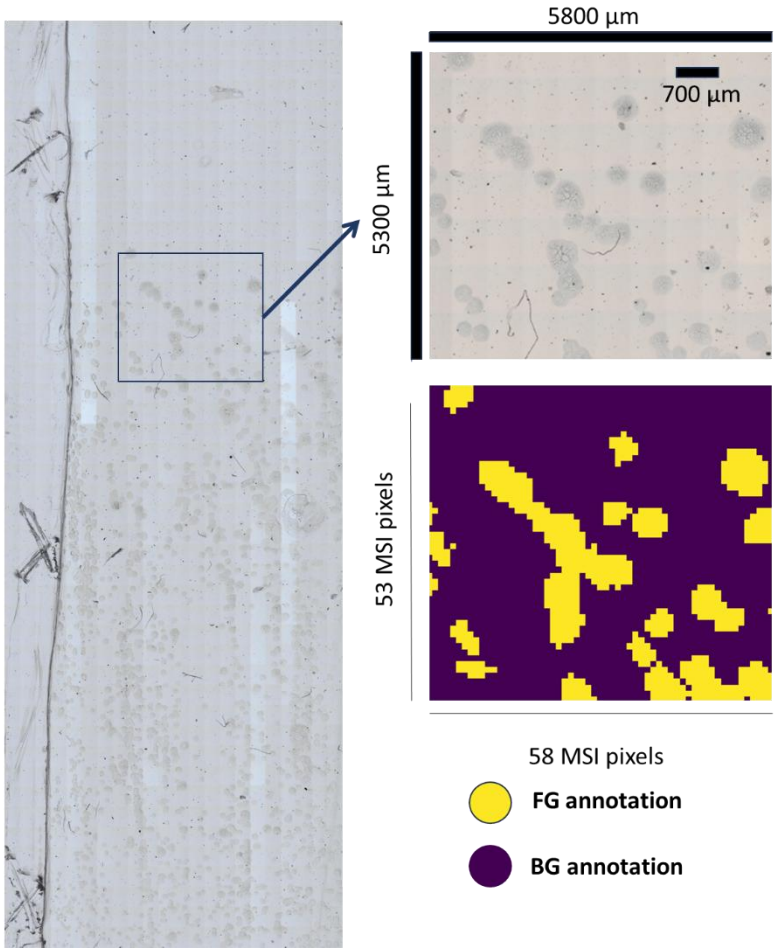


Figure S0: **Left:** Brightfield microscopy image of the full *Pseudomonas* culture on the ITO plate. **Top right:** Close-up of the region analyzed by MSI. **Bottom right:** Corresponding annotation of the foreground (FG) and background (BG) regions used for classification.

S6.3. MSI for background and foreground annotation

All MSI spectra are annotated into one of two classes: The background (BG), corresponding to regions without biological material, and the foreground (FG), corresponding to the targeted sample (see the last column in Figure S1). To speed up the annotation process, a K-means algorithm[6] is first used to cluster all mass spectra of each pre-processed image into 2 clusters. Then, a visual inspection of the spatial distribution of these clusters and their associated signals helped us rapidly determine which cluster corresponds to either the background or the foreground. Visual inspection of the cluster spatial distribution is enough to determine the position of the bacterial colonies in the two bacterial datasets from METASPACE (see Figure S1). Then, for the *Pseudomonas* bacterial colonies grown on agar, we validate the cluster classes by comparing them with a microscope image of the plate where bacterial colonies are visible. Finally, we confirmed the cluster classes of the biopsy samples by analyzing the molecular composition of each class. The molecular composition was determined with the help of their Kendrick plots Figure S2.

Those plots show that the most intense signals in the foreground correspond mostly to lipids, and the main signals in the background correspond to DHB ions without lipid ions. This confirms that the background classes are correctly attributed to the 2 biopsy images.

S6.4. Hepatocytes Dataset

From the human HepaRG hepatocyte cell culture datasets generated by Rappez et al[7], we select as our two classes the datasets containing hepatocytes in the homeostatic state (control) and the datasets modeling non-inflammatory Non-Alcoholic Fatty Liver Disease (NAFLD). We have a total of 8 datasets, 4 for the control and 4 for the non-inflammatory NAFLD) downloaded as centroided imzML. We generate a global dataset by randomly sampling 1,250 spectra from each image, which adds up to 10,000 spectra. From each spectrum, a mass range from 200-1100 m/z is selected, and 2000 of the most intense centroids are chosen.

S6.5. Streptomyces scabiei Dataset

The second case study consists of bacterial culture of *Streptomyces scabiei* for natural product discovery, using two MALDI-MSI datasets from Burguet et al[8]. These datasets were designed to compare culture conditions that either activate or repress the production of specialized metabolites in Streptomyces scabiei.

Dataset 1: Cellobiose-Induced Pathogenicity

Here, we compare spectra from *S. scabiei* cultures exposed to cellobiose (which induces virulence-related metabolite production) against negative controls treated with water or sugar. This dataset contains 3 MSI (centroided imzML), one for the experimental condition and one for each control.

We generate a global dataset by randomly sampling 2000 spectra from each image, which adds up to 6,000 spectra. From each spectrum, a mass range from 200-1600 m/z is selected, and 2000 of the most intense centroids are chosen.

Dataset 2: Iron-Regulated Siderophore Production

In this dataset, *S. scabiei* is grown under iron-depleted versus iron-supplemented conditions to study siderophore biosynthesis. This dataset contains 2 MSI (centroided imzML), one for the experimental condition and one the control.

We generate a global dataset by taking all spectra 3300 (control) and 2400 (FeCl3) spectra from each image, which adds up to 5700 spectra. From each spectrum, a mass range from 200-1600 m/z is selected, and 2000 of the most intense centroids are chosen.

S6.5. Figure S1: Description of the different MALDI MSI datasets

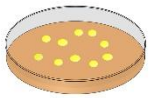
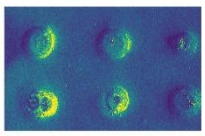
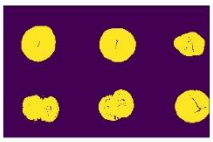
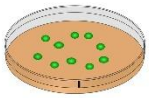
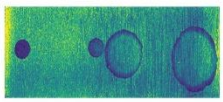
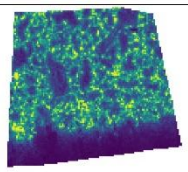
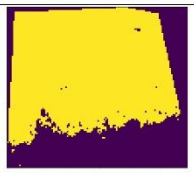
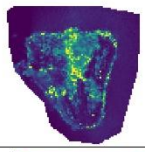
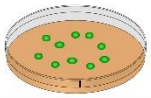
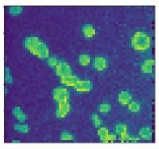
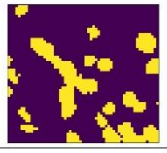
Organism	MALDI Matrix	Mass Analyzer	Mass Signal Example	Annotation **Background** vs **Foreground**
Streptomyces 	DHB HCCA	Orbitrap		
Pseudomonas 	DHB HCCA	Orbitrap		
Lung 	DHB	FT-ICR		
Kidney 	DHB	FT-ICR		
Pseudomonas 	HCCA	FT-ICR		

Figure S1. Description of the different MALDI MSI datasets used in this work (organism of origins, MALDI matrixes, mass analyzers). The signal example corresponds to the intensity distribution of one ion. The last column corresponds to the foreground (FG) versus background (BG) annotations.

S6.6. Figure S2: background/foreground Kendrick analysis

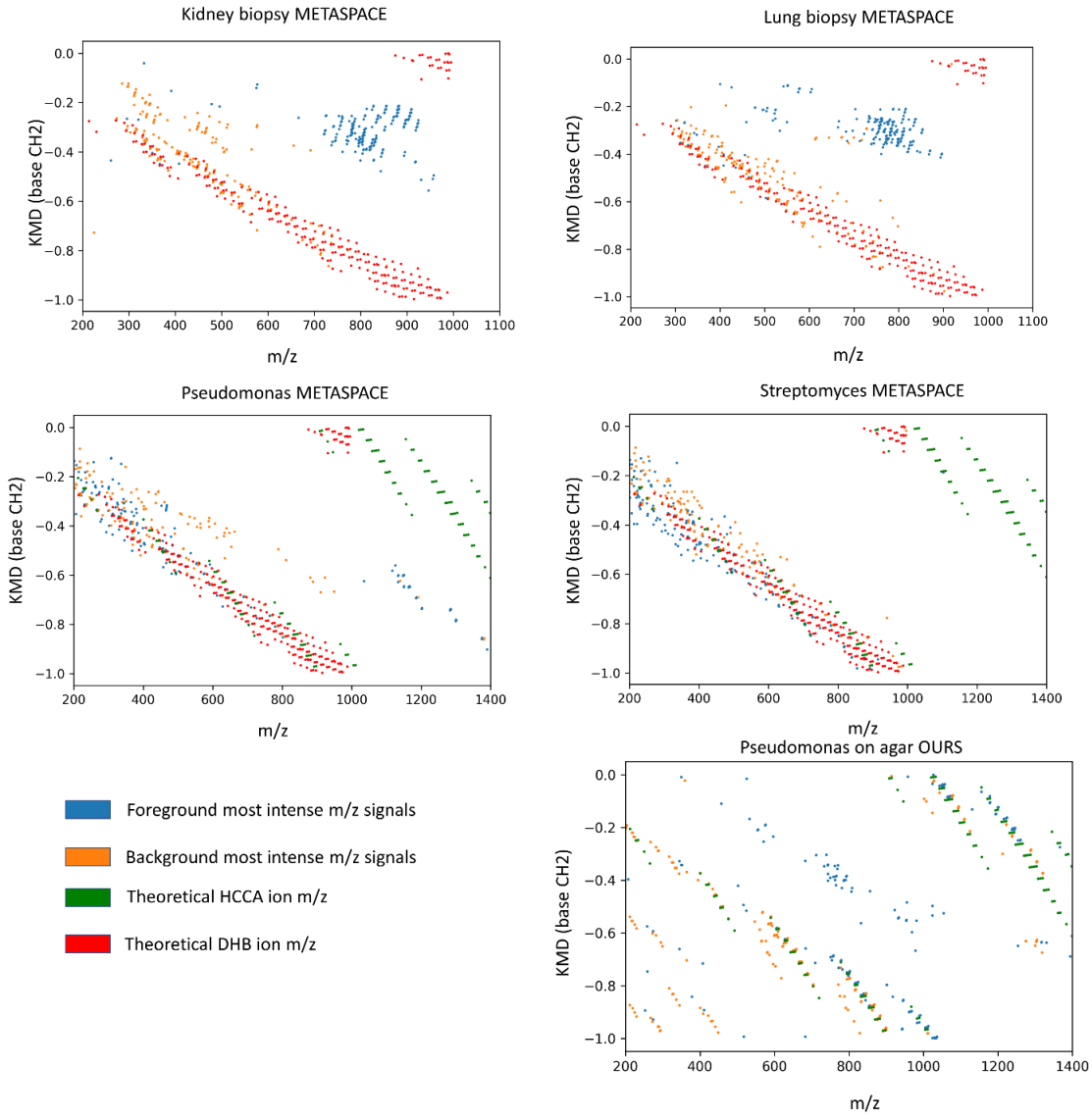


Figure S2. Kendrick plots of the main m/z signals for each class background/foreground and the theoretical MALDI matrix mass patterns of DHB and HCCA. The blue dots correspond to the 200 signals with a higher mean expression in the foreground spectra than the background spectra. The orange dots correspond to the 200 signals with a higher mean expression in the background spectra than the foreground spectra.

S7: R-GCN Implementation Description

We implemented the architecture in Python using the following libraries and versions: pytorch=1.6, torchvision=0.7, cudatoolkit=10.1, and torch-geometric=2.0.2. The network is composed of two RGCN Convolutional layers from torch_geometric.nn.conv (RGCNConv) activated at the output with a ReLU activation function to bring non-linearity to the network. Then, a global max pooling layer is applied on the nodes of the graph feature by feature using global_max_pool from torch_geometric.nn. Then, A 1D batch norm is applied to the pooled features using torch.nn.BatchNorm1d followed by two Dense layers from torch.nn.Linear (the first one is ReLU activated, and the second one is softmax activated to output bounded value for each class) is applied. The number of features in each layer is: 3 (input), 8 (RGCNConv 1), 16 (RGCNConv2), 8 (Dense Layer 1), and 2 (Dense layer 2). The model is trained using the cross-entropy loss, the ADAM optimizer, and a reduce-on-plateau scheduler strategy with an initial learning rate of 10-3 and a division by 10 at each plateau until a learning rate of 10-8 where the training stops.

S8: Performance comparison of different R-GCN parameters

Parameters	Background/foreground	Control and /non-inflammatory NAFLD human HepaRG hepatocyte
Default parameters	0.99344+- 0.0007	0.9384 +- 0.0072
Tolerance 0.0005 Da	0.9946 +-0.0032	0.9314 +- 0.0044
Tolerance 0.0001 Da	**0.9676 +- 0.0043**	**0.8124 +-0.0049**
Tolerance 0.003 Da	**0.9952 +- 0.00154**	0.9406 +- 0.0033
Tolerance 0.007 Da	0.9951 +- 0.0027	0.9354 +- 0.0053
300 Most frequent mass patterns in all peaks	0.9861 +-0.0018	0.9518 +- 0,0068
102 Most frequent mass patterns in all peaks	0.9854 +- 0.0031	0.9186 +- 0.0144
102 Most frequent mass patterns in all annotated peaks	0.9698 +- 0,0069	0.9254 +- 0.01679
300 Most frequent mass patterns in all annotated peaks	0.9938 +-0.00171	**0.9736 +- 0.0116**
From a node, link all nodes that match the same mass difference	0.9930 +- 0.0001	0.9384 +- 0.0072
Mass Defect base H2	0.9586 +- 0.0026	0.925 +- 0.0034
Mass Defect base CH2	0.9746 +- 0.0027	0.9328 +- 0.0026
Multiplier 2	0.9938 +- 0.0019	0.9385 +-0.0051
Multiplier 3	0.9938 +- 0.000	0.9506 +- 0.0051
Removed mass and mass defect as node features	0.9942 +- 0.0022	0.9324 +- 0.0057
Normalization by TIC	0.9865 +- 0.000	0.9218 +- 0,0033
Min-max normalization	0.9890 +- 0.0010	0.9221 +- 0.0072
No normalization	0.9925 +- 0.0026	0.9366 +- 0.0084

Table S3: Comparison of the impact of different hyperparameters of our method (R-GCN) on mass spectra classification performances (balanced accuracies). The models are trained on the hepatocyte and background/foreground datasets. We reduced the dataset sizes by randomly sampling 1/5 of the spectra from the original datasets. The default parameters correspond to the ones described in the manuscript. Each row of the table corresponds to the model performance with a change in one parameter compared to the default. From 2-5, we have modified the mass tolerance that considers whether the graph's two nodes (centroids) are linked. From 6-7, we have generated a list of 300 or 102 mass differences from the most frequent mass differences observed in the training set of spectra. From 7-9, we have generated a list of 300 or 102 mass differences from the most frequent mass differences observed between METASPACE annotated centroids. From 10-11, we have modified the base from which Kendrick mass defect is computed. From 12-13, we have modified the number of features in each layer of the R-GCN by multiplying the default number of features by 2 or 3. The bolded values correspond to the best and worst performances of the model for each dataset.

Different methods for selecting mass differences have been explored, such as computing the most frequent mass differences between the exact mass of the molecular assignment obtained by METASPACE. While this strategy gives better results than the 102 most frequent mass patterns selected in the manuscript, this selection depends on annotated mass features, rendering the method less practical on datasets containing unknown ions, such as the bacterial colony datasets.

S9: 1D-CNN, SVM, RF, LDA, XGBoost pre-processing

We choose the Random Forest (RF), Linear discriminant analysis (LDA), Support vector machine (SVM) implementations from the scikit-learn Python library "RandomForestClassifier" [9], SVC[10], LinearDiscriminantAnalysis, with default parameters. We choose the Python library XGboost [11]. Unlike our method, these algorithms require MSI pre-processing to find all corresponding signals and mass peaks among all mass spectra. Therefore, we first compute an average spectrum from all spectra in the training set and keep the 2000 most intense peaks. Then, for each spectrum given to the algorithm, during training or inference, we look for the corresponding mean spectrum's peaks with a mass tolerance of ±0.001 Da, and return their corresponding intensities within the same mass tolerance, creating a set of 2000 intensity values per spectrum. Finally, the intensities of the spectra are first log-transformed before being normalized by their Total ion count, as described in Seddiki et al.[12].

For the 1D-CNN, we choose the implementation of *variant Lecun*[12]. In this case, the preprocessing consists of dividing the mass range of each mass spectrum into fixed-sized bins. The intensity value of each bin is computed as the sum of intensity values in the spectrum covered by the bin. We keep the same parameters as advised in Seddiki et al.[12] such as a bin size of 0.1 Da as a smaller bin size requires a larger network, which is then simply too big for most GPU servers. Finally, as for all methods, the intensities of the spectra are first log-transformed before being scaled by a min-max scaling between 0 and 1.

S10: Evaluation of unseen data

In this section, we evaluate the classification robustness of the different methods on unseen datasets where multiple previous signal variations co-occur. Moreover, those datasets have very distinct signals as they come from different species analyzed with different MALDI matrices, making them ideal for evaluating the generalizability of the different methods.

S10.1. Experimental setup

To this end, we create two pairs of train/test sets. For the first pair, the train set comprises the METSPACE MSI from human organ biopsies samples (kidney and lung, third and fourth datasets in Figure S1). In contrast, the test set comprises bacterial colonies grown on agar (fifth dataset in Figure S1). For the second pair, we simply inverse the train and test sets. The two pairs of samples have very distinct signals. Indeed, the differences in molecular composition are due, on the one hand, to their distinct metabolisms, cellular composition, and structure (Humans vs bacteria) and, on the other hand, to the use of different MALDI matrices for the acquisitions (HCCA for the bacterial and DHB for the human organ biopsies). As in the previous section, the models are then evaluated by computing the balanced accuracy of the classification BG/FG on 15 models for each method.

Table S4: Performances of the binary classification (BG/FG) for RF, SVM, LDA, XGBoost, 1D-CNN and our RGCN on samples outside their training distribution. Mean performances of 15 models for each method trained and tested on bacterial colonies and organ biopsies.

Balanced accuracy					
Trained on\Tested on	R-GCN (ours)	1D-CNN	RF	SVM	LDA
Bacterial colonies\Organ biopsies	0.823 ± 0.017	0.709 ± 0.042	0.5 ± 0.0	0.5 ± 0.0	0.5 ± 0.0
Organ biopsies\Bacterial colonies	0.790 ± 0.083	0.519 ± 0.035	0.5 ± 0.0	0.5 ± 0.0	0.499 ± 0.0

S10.2. Classification results and analysis

Our R-GCN has very good mean prediction performances, over 80% balanced accuracy, and clearly outperforms the other methods for both pairs of evaluations (Table S2). The other methods perform relatively poorly. In particular, the methods based on peak matching (RF, SVM, LDA, XGBoost) cannot generalize to new datasets outside its training distribution, while 1D-CNN performs moderately well. We have also visualized the image of the predictions over the bacterial colony dataset when the models are trained on the organ biopsy datasets (see Figure S3).

The high classification performance of R-GCN on unobserved data shows that related signals exist between the classes BG/FG of the bacterial colony dataset and those of the human biopsies dataset.

First, the lipid ions are characteristic of the FG class for any organism. Although the lipid ions of different organisms do not have the same masses, they have very similar mass defects and mass ranges, indicating a close atomic composition. Moreover, the mass patterns of the lipid ions are mostly the same between the two organisms as they are classically separated by H_2, C_2H_4, and adduct patterns. However, RF cannot capture these features and is only poorly captured by 1D-CNN. Methods based on peak matching discard mass information, and 1D-CNN has too large bins and too restricted convolution to learn the spaced and accurate patterns of the HRMS spectrum.

We can also analyze the mean class prediction for the bacterial colony dataset when the models are trained on the organ biopsy datasets. As shown in Figure S3, R-GCN correctly classifies most of the BG and FG pixels of Pseudomonas colonies. RF classifies everything as BG, and 1D-CNN classifies almost all pixels as BG. We also mapped the mean probability output for each class. Through this visualization, we can observe which regions of the MSI trigger a model's response for each class. The mean probability maps highlight that RF's models can still not discern the two classes. For example, the bacterial colonies spectra wrongly activate the model for BG. However, this is not the case for 1D-CNN, as it tends to be more discriminating of the bacterial colony spectra (Figure S3).

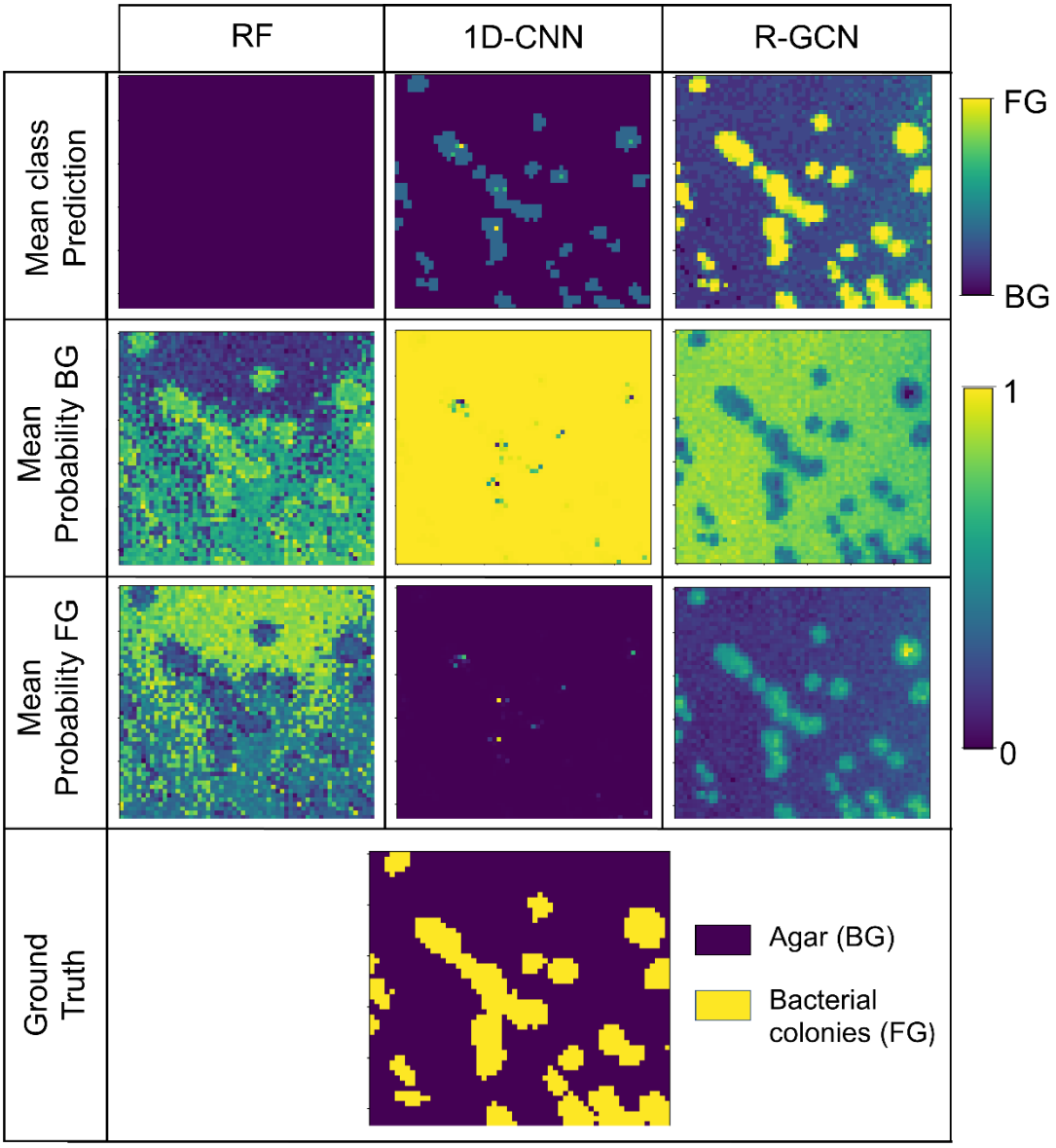


Figure S3: Performances of the binary classification (BG/FG) for RF, 1D-CNN and R-GCN on samples outside their training distribution. The models are trained on organ biopsy datasets and evaluated on a bacterial colony dataset. The first row contains the mean predictions of 15 models for each method over the bacterial colony image. Yellow pixels correspond to pixels most often classified as FG, and black pixels correspond to pixels most often classified as BG. The second and third rows correspond to the mean probability prediction of each class over all pixels. The last row corresponds to the ground truth classes of the bacterial colony dataset.

The better performances of 1D-CNN and R-GCN show that integrating mass features (exact mass, mass defect, mass patterns) is essential for applying unseen mass spectra (generalization process), confirming the results of the previous analysis on signal variation. It is possible that these algorithms can improve their performances when trained on a more diverse set of datasets. However, as MSI data are expensive to generate and signal variations are frequent and diverse, having the capacity to learn on very few datasets is a significant advantage. Finally, only our method can learn and understand molecular similarities between signals based on mass patterns and mass and mass defect information, allowing lipid signals in different organisms to be considered similarly.

S11: Model evaluation details

S11.1

Model	Training Time (sec)	Number of parameters
LDA	2	/
xgboost	3	/
RF	7	/
SVM	33	/
1D-CNN	102	5771170
R-GCN	1367	15746

Table S4: Training time and model complexity for each method on the background/foreground classification task.

This table reports the training time (in seconds) and the number of trainable parameters for each classification model when applied to the background/foreground classification task. All models were trained on the same dataset for a fair comparison. Classical machine learning models (LDA, XGBoost, Random Forest, and SVM) were implemented using the scikit-learn library and do not rely on trainable parameters in the deep learning sense. In contrast, the 1D-CNN and R-GCN are deep learning models implemented in Keras and PyTorch, respectively.

The deep learning experiments were performed on a workstation equipped with an Intel Core i9-9920X CPU (3.50 GHz) and an NVIDIA RTX 2080 Ti GPU (11 GB VRAM). Both the 1D-CNN and R-GCN were trained using GPU acceleration.

While the R-GCN has significantly fewer trainable parameters (~15,746) compared to the 1D-CNN (~5.77 million), it requires considerably more training time. This discrepancy is primarily due to differences in data loading strategies. The 1D-CNN implementation preloads the full dataset into memory before training, allowing efficient batch processing and low I/O latency. In contrast, our current R-GCN implementation loads each spectrum and its associated graph dynamically from an external hard drive during training. This introduces substantial I/O overhead, which becomes the bottleneck in training speed.

We anticipate that optimizing the data pipeline, for example, by preloading all spectra and graph structures into RAM or switching to faster storage, would significantly reduce the R-GCN training time, making it more practical for broader deployment.

S11.2. Class imbalance dataset

We chose to use balanced accuracy to provide a more robust and fair evaluation metric, particularly in the context of class imbalance, which is present in several of our datasets. Unlike standard accuracy, which can be inflated by the majority class, balanced accuracy gives equal weight to each class, ensuring that model performance is not biased toward the more prevalent class. To clarify:

- In the background/foreground dataset, the class distribution is 4005 foreground vs. 5995 background (imbalance).
- In the Pseudomonas dataset, the distribution is 3698 vs. 10472, showing even stronger imbalance.
- The hepatocyte dataset is balanced (5000 vs. 5000), but we maintain the same metric across all experiments for consistency.

S11.3.R-GCN signal interpretability: Ion classification variability

Dataset	Minium accuracy	Maximum accuracy	Average accuracy
Pseudomonas colonies	0.98	0.98	0.98
Streptomyces scabiei (cellobiose)	0.95	0.96	0.96
Streptomyces scabiei (FeCl3)	0.95	0.97	0.96
HepaRG	0.94	0.97	0.96

Table S5: Pairwise classification agreement between five R-GCN models trained with different random seeds for each dataset. Agreement is computed using the average pairwise accuracy across the predicted m/z class labels. The minimum and maximum accuracies across all model pairs are also reported. This analysis demonstrates the consistency of the ion-level classification results across different initializations of the model.

References

(1) Janda, M.; Seah, B. K. B.; Jakob, D.; Beckmann, J.; Geier, B.; Liebeke, M. Determination of Abundant Metabolite Matrix Adducts Illuminates the Dark Metabolome of MALDI-Mass Spectrometry Imaging Datasets. *Anal. Chem.* **2021**, *93* (24), 8399–8407. https://doi.org/10.1021/acs.analchem.0c04720.

(2) Schramm, T.; Hester, Z.; Klinkert, I.; Both, J.-P.; Heeren, R. M. A.; Brunelle, A.; Laprévote, O.; Desbenoit, N.; Robbe, M.-F.; Stoeckli, M.; Spengler, B.; Römpp, A. imzML — A Common Data Format for the Flexible Exchange and Processing of Mass Spectrometry Imaging Data. *J Proteomics* **2012**, *75* (16), 5106–5110. https://doi.org/10.1016/j.jprot.2012.07.026.

(3) Skinner, S. O.; Sepúlveda, L. A.; Xu, H.; Golding, I. Measuring mRNA Copy Number in Individual Escherichia Coli Cells Using Single-Molecule Fluorescent in Situ Hybridization. *Nat. Protoc.* **2013**, *8* (6), 1100–1113. https://doi.org/10.1038/nprot.2013.066.

(4) McCann, A.; Rappe, S.; La Rocca, R.; Tiquet, M.; Quinton, L.; Eppe, G.; Far, J.; De Pauw, E.; Kune, C. Mass Shift in Mass Spectrometry Imaging: Comprehensive Analysis and Practical Corrective Workflow. *Anal. Bioanal. Chem.* **2021**, *413* (10), 2831–2844. https://doi.org/10.1007/s00216-021-03174-1.

(5) La Rocca, R.; Kune, C.; Tiquet, M.; Stuart, L.; Eppe, G.; Alexandrov, T.; De Pauw, E.; Quinton, L. Adaptive Pixel Mass Recalibration for Mass Spectrometry Imaging Based on Locally Endogenous Biological Signals. *Anal. Chem.* **2021**, *93* (8), 4066–4074. https://doi.org/10.1021/acs.analchem.0c05071.

(6) Alexandrov, T. MALDI Imaging Mass Spectrometry: Statistical Data Analysis and Current Computational Challenges. *BMC Bioinf* **2012**, *13* (S16), S11. https://doi.org/10.1186/1471-2105-13-S16-S11.

(7) Rappez, L.; Stadler, M.; Triana, S.; Gathungu, R. M.; Ovchinnikova, K.; Phapale, P.; Heikenwalder, M.; Alexandrov, T. SpaceM Reveals Metabolic States of Single Cells. *Nat. Methods* **2021**, *18* (7), 799–805. https://doi.org/10.1038/s41592-021-01198-0.

(8) Burguet, P.; La Rocca, R.; Kune, C.; Tellatin, D.; Stulanovic, N.; Rigolet, A.; Far, J.; Ongena, M.; Rigali, S.; Quinton, L. Exploiting Differential Signal Filtering (DSF) and Image Structure Filtering (ISF) Methods for Untargeted Mass Spectrometry Imaging of Bacterial Metabolites. *J. Am. Soc. Mass Spectrom.* **2024**, *35* (8), 1743–1755. https://doi.org/10.1021/jasms.4c00129.

(9) Breiman, L. Random Forests. *Mach Learn* **2001**, *45* (1), 5–32. https://doi.org/10.1023/A:1010933404324.

(10) Chang, C.-C.; Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2* (3), 1–27. https://doi.org/10.1145/1961189.1961199.

(11) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: San Francisco California USA, 2016; pp 785–794. https://doi.org/10.1145/2939672.2939785.

(12) Seddiki, K.; Saudemont, P.; Precioso, F.; Ogrinc, N.; Wisztorski, M.; Salzet, M.; Fournier, I.; Droit, A. Cumulative Learning Enables Convolutional Neural Network Representations for Small Mass Spectrometry Data Classification. *Nat Commun* **2020**, *11* (1), 5595. https://doi.org/10.1038/s41467-020-19354-z.

Supporting information of Chapter 6

S1: Molecular annotation workflow

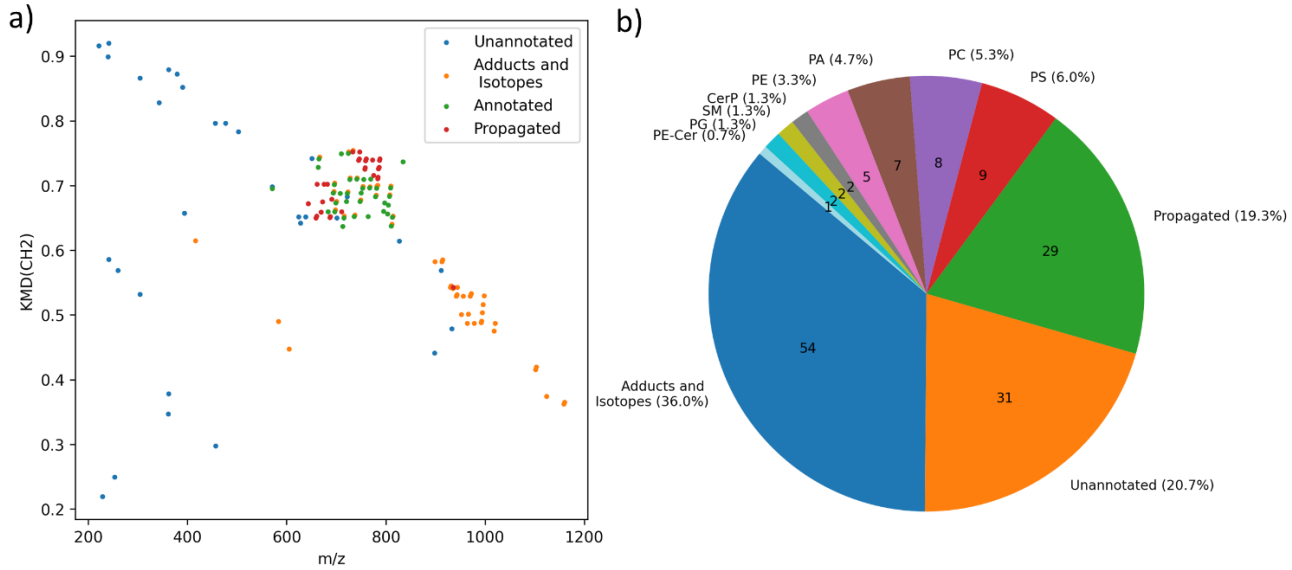


Figure S1. *Curation of MSI ions used in fusion. (a) Kendrick-style overview of the 150 cell-correlated m/z values after spatial filtering (chaos ≥ 0.7) and correlation screening (Spearman p ≥ 0.5). Points are colored by status: annotated lipids (green, < 2 ppm), adducts/isotopes (orange; identified by exact mass deltas), propagated homologs (red; C_2H_4/H_2 spacing from annotated seeds), and unannotated (blue).*
(b) Composition of the curated set: 36 annotated phospholipids, 54 adducts/isotopes, 29 propagated phospholipid-like signals, 31 unannotated.

We detail here the procedure for curating MSI peaks and assigning putative molecular identities prior to model training/evaluation.

1. **Database annotation with METASPACE.**

   The coculture imzML was uploaded to METASPACE with 3 ppm tolerance against SwissLipids (2018-02-02), HMDB (v4), and ChEBI (2018-01), allowing $[M+H]^+$, $[M+Na]^+$, and $[M+K]^+$adducts. At 10% FDR, very few annotations were returned

(SwissLipids: 5; HMDB: 3; ChEBI: 0). To increase coverage, we retained SwissLipids at 20% FDR (125 annotations) while HMDB/ChEBI did not gain additional confident hits.

2. **Remove spatially noisy ion images.**

   To exclude non-informative/artefactual peaks (e.g., matrix splatter, detector noise), we computed the spatial chaos score (Chapter 2) on both monoculture slides and discarded ions with score < 0.7 on either slide. This reduced the working list from 8,694 to 2,141 m/z values.

3. **Keep ions correlated with cell presence.**

   From the manually annotated monoculture subregions, we computed for each MSI pixel the count of microscope pixels labeled as "cell" (MCF-7 ∪ WI-38) within a central crop (crop fraction c = 0.3; see Methods). We then concatenated the two monocultures and computed the Spearman correlation between this cell-area proxy and each ion image. Ions with $\rho \geq 0.5$ were retained, yielding 150 m/z values (Figure S1a).

4. **Map database annotations and apply ppm quality filter.**

   We associated METASPACE assignments to the 150 retained ions and kept only lipid annotations with mass error < 2 ppm. This left 36 phospholipid assignments from SwissLipids; no additional confident metabolite annotations were obtained from HMDB/ChEBI under these filters.

5. **Identify adducts/isotopes among remaining signals.**

   We examined exact mass differences (tolerance ±0.005 Da) corresponding to common relationships:

   -HCCA dimer adduct: 189.04259 Da

   – Na–H: 21.98194 Da

   – K–H: 37.95588 Da

   Peaks matching these deltas to an annotated peak (or to each other) were labeled as adducts/isotopes, accounting for 54 (36%) of the 150 m/z values.

6. **Propagation by homologous series.**

Several unannotated peaks formed series separated by C_2H_4 (28.0313 Da) and/or H_2 (2.0157 Da). We propagated known lipid annotations along these series (combinations of 1–3 × C_2H_4 and 1–3 × H_2), identifying 28 additional signals as putative (unknown-formula) phospholipids (Figure S1a, red).

7. **Unannotated remainder & odd-chain caution.**

   After the steps above, 30 m/z values remained unannotated. We also note that a few database assignments suggested odd total carbon numbers in the acyl chains (e.g., PS C37, PG C41/C43). Because mammalian lipid elongation typically proceeds in C_2 units, we did not propagate CH_2 (14.0157 Da) spacing and treat such odd-chain assignments with caution; they may reflect alternative headgroup assignments, adduct/fragment misassignment, or rarer odd-chain species. Targeted MS/MS would be required to resolve these cases.

Summary: Of the 150 cell-correlated ions, 36 were confidently mapped to phospholipids (< 2 ppm), 54 were classified as adducts/isotopes, 28 were propagated as likely phospholipid homologs, and 30 remained unannotated (Figure S1b).

Associated papers: Table of Contents

Associated papers of Chapter 2

A1- Original publication of “Exploiting Differential Signal Filtering (DSF) and Image Structure Filtering (ISF) Methods for Untargeted Mass Spectrometry Imaging of Bacterial Metabolites.”

Associated papers of Chapter 3

A2- Original publication of “Mass shift in mass spectrometry imaging: comprehensive analysis and practical corrective workflow.”

A3- Original publication of “FT-ICR Mass Spectrometry Imaging at Extreme Mass Resolving Power Using a Dynamically Harmonized ICR Cell with 1ω or 2ω Detection.”

Associated papers of Chapter 4

A4- Original publication of “Rapid Visualization of Chemically Related Compounds Using Kendrick Mass Defect As a Filter in Mass Spectrometry Imaging.”

A5- Original publication of “Organ-resolved lipid mapping in Steatoda nobilis spider model using high-resolution mass spectrometry imaging and Kendrick mass defect analysis.”

Associated papers of Chapter 2

A1- Original publication of “Exploiting Differential Signal Filtering (DSF) and Image Structure Filtering (ISF) Methods for Untargeted Mass Spectrometry Imaging of Bacterial Metabolites.”

ACS Partner Journal

Journal of the American Society for Mass Spectrometry

pubs.acs.org/jasms

Article

Exploiting Differential Signal Filtering (DSF) and Image Structure Filtering (ISF) Methods for Untargeted Mass Spectrometry Imaging of Bacterial Metabolites

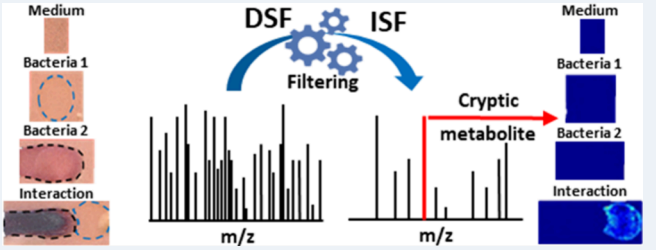
Pierre Burguet, Raphaël La Rocca, Christopher Kune, Déborah Tellatin, Nudzejma Stulanovic, Augustin Rigolet, Johann Far, Marc Ongena, Sébastien Rigali, and Loïc Quinton*

Cite This: *J. Am. Soc. Mass Spectrom.* 2024, 35, 1743−1755

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Matrix-assisted laser desorption/ionization (MALDI) mass spectrometry imaging (MSI) is a label-free technique, producing images where pixels contain mass spectra. The technique allows the visualization of the spatial distribution of (bio)molecules from metabolites to proteins, on surfaces such as tissues sections or bacteria culture media. One particularly exciting example of MALDI-MSI use rests on its potential to localize ionized compounds produced during microbial interactions and chemical communication, offering a molecular snapshot of metabolomes at a given time. The huge size and the complexity of generated MSI data make the processing of the data challenging, which requires the use of computational methods. Despite recent advances, currently available commercial software relies mainly on statistical tools to identify patterns, similarities, and differences within data sets. However, grouping m/z values unique to a given data set according to microbiological contexts, such as coculture experiments, still requires tedious manual analysis. Here we propose a nontargeted method exploiting the differential signals between negative controls and tested experimental conditions, i.e., differential signal filtering (DSF), and a scoring of the ion images using image structure filtering (ISF) coupled with a fold change score between the controls and the conditions of interest. These methods were first applied to coculture experiments involving *Escherichia coli* and *Streptomyces coelicolor*, revealing specific MS signals during bacterial interaction. Two case studies were also investigated: (i) cellobiose-mediated induction for the pathogenicity of *Streptomyces scabiei*, the causative agent of common scab on root and tuber crops, and (ii) iron-repressed production of siderophores of *S. scabiei*. This report proposes guidelines for MALDI-MSI data treatment applied in the case of microbiology contexts, with enhanced ion peak annotation in specific culture conditions. The strengths and weaknesses of the methods are discussed.



INTRODUCTION

Matrix-assisted laser desorption/ionization mass spectrometry imaging (MALDI-MSI) is a label-free technique used to visualize the spatial distribution of molecules,[1−3] such as metabolites,[4] peptides,[5] lipids,[6] or proteins[7] on various surfaces. Information obtained from comprehensive studies based on mass spectrometry (MS) and the determination of the spatial distributions of molecules in samples offer significant knowledge about the molecular characterization of biological samples, ranging from tissue sections (organ biopsies, plant slices, ...) to biopolymer thin films.[8] Over the last decades, MSI methods have been increasingly explored in microbiology to produce images of metabolite synthesis and distribution by various interacting microorganisms, usually grown on a thin-layer of agar based medium.[2,9−11] The molecular "snapshot" of the metabolome of strains grown under controlled experimental conditions allows inferring molecular crosstalk and induced metabolite expression within and between microorganisms. Competition for resources and inhibition or elimination of antagonist species would greatly benefit *in situ* and real-time monitoring, making MALDI-MSI well suited for biomarkers, chemical communication elucidation, or new molecule discovery.[12]

Despite the recent advances of MSI applied in the field of microbiology, concerning the reproducibility of sample preparation, homogeneous matrix deposition or data acquisition,[12] the MSI data generated have a huge memory footprint and are challenging to process and analyze. Actually, each image file contains thousands of spectra (and consequently thousands of pixels), each of which displays thousands of mass m/z signals (peaks), especially when MALDI-MSI and high-resolution mass spectrometry (HRMS) are engaged. These $m/$

Received: March 29, 2024
Revised: June 27, 2024
Accepted: June 28, 2024
Published: July 15, 2024



Downloaded via UNIV DE LIEGE on September 2, 2024 at 17:06:33 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

ACS Publications

© 2024 American Society for Mass Spectrometry. Published by American Chemical Society. All rights reserved.

z signals can have diverse origins. In the low mass range, they are, for instance, mainly due to matrix ions, adducts, and aggregates.[13] The mass spectra also contain signals from salts, agar, and unconsumed nutrients (amino acids, sugars,...) contained in the culture media that further increase the chemical background and may cause ion suppression effects, impairing the detection of ions specifically implied in the investigated metabolic activities of the microorganisms. Most of these molecules can be ionized and desorbed in positive ionization mode as protonated species, or cationized with sodium, ammonium, and potassium adducts, and in various oxidation states due to the ionization processes. In negative ionization mode, the metabolites and their oxidized forms are mainly detected as deprotonated species. Overall, the number of signals contained in each image can easily exceed several thousand m/z peaks, making untargeted and discovery studies by manual data analysis unrealistic even if it could lead to the highlight of a few (sometimes arbitrarily selected) ions of interest.[8] Therefore, the data generated by MSI often require advanced software or computational approaches to be analyzed. Computational tools can handle the complexities of MSI data sets, such as spatial segmentation, component analysis, hypothesis tests, or statistical tools based on colocalized features. They have been developed for identifying patterns, similarities, and differences within data sets. Spatial segmentation appears to be an appropriate approach for assessing similarities between spectra and grouping them into clusters. This allows differentiation of specific areas within the sample, based on their molecular compositions.[14] The so-called colocalized value approach reveals m/z peaks that are primarily localized within a specific region of the image, proving highly beneficial for evaluating the spatial relationship among various molecular species in a sample.[15] Principal component analysis (PCA) remains one of the most popular methods to determine the variance within a data set. These approaches are typically used to reduce the complexity of the data set by identifying patterns and highlighting variations between different samples.[16] Finally, univariate statistical hypothesis and discriminative tests such as t test are suitable to determine whether an m/z value can be used to confidently discriminate between two or more groups.[17]

However, none of the currently available software, including both academic and commercial codes used for analyzing MSI data, provides satisfactory data processing and still requires tedious efforts to refine the processed data by hand, especially for features which are unique to a peculiar data set. This is pertinent across various fields, including microbiology, and of great importance when studying secondary metabolism of bacteria. By modifying the culture conditions, it is possible not only to trigger global changes in the regulation of the metabolic pathways but also to induce the expression of cryptic genes responsible for the synthesis of unknown metabolites.[18−21] The development of our own in-house computational and statistical tools was proposed not only for improving data treatment in the context of microbiology-based studies, but also for speeding up the MSI data analysis while enhancing the quality of the results. Similarly, the development of in-house software by other laboratories has already been demonstrated as significant to improve analytical capabilities, showcasing the broad applicability and potential of custom solutions in scientific research.[22−24]

Here we propose a complete workflow based on 2 algorithms used separately or in combination to filter and extract a maximum of ions of interest with a great confidence. The first one is exploiting the differential signals detected between two or more data sets, differential signal filtering (DSF). The second is a scoring of the ion image structure (i.e., evaluation of the degree of coherence concerning the repartition of the ion image in the whole image) using image structure filtering (ISF) coupled with a fold change score (i.e., how much the score changed between controls and the condition of interest).

We challenged our method using three case studies to demonstrate its robustness and the increased speed of the data analysis. In the first case study, we propose guidelines on how MSI data treatment should be performed in the frame of MALDI-MSI applied to metabolomic response of bacterial coculture in interaction between *Escherichia coli* and *Streptomyces coelicolor* to improve the detection and the identification of ions. The strengths and weaknesses of DSF and ISF methods—applied individually or in combination—are also critically discussed, especially with regard to tedious manual analysis. Finally, two other case studies were applied using our proposed methods: (i) cellobiose-mediated induction of the pathogenicity of *Streptomyces scabiei*, the causative agent of common scab on root and tuber crops, and (ii) iron-repressed production of siderophores of *Streptomyces scabiei*.

METHODS

Chemicals. α-Cyano-4-hydroxycinnamic acid (HCCA), 9-aminoacridine (9-AA), and formic acid (FA) were purchased from Sigma-Aldrich (Belgium). Acetonitrile was LC-MS grade and obtained from Biosolve (Belgium).

Bacterial Strains, Medium, and Culture Conditions. The bacterial strains used in this study are *Streptomyces coelicolor* M145, *Streptomyces scabiei* 87-22, and *Escherichia coli* DH5α. Cultivation media were prepared as follows (for 1 L): International Streptomyces Project number 2 (ISP2) medium (Yeast extract (VWR) 4 g, Bacto Malt Extract (Gibco) 10 g, D(+)-glucose monohydrate (Carl Roth) 4 g, and Bacto Agar (BD Difco) 20 g); Peptone Yeast Malt (PYM, Yeast extract (VWR) 3 g, Bacto Malt Extract (Gibco) 3 g, D(+)-glucose monohydrate (Carl Roth) 10 g, bacteriological peptone (Condalab) 5 g, and Bacto Agar (BD Difco) 20 g); Thaxtomin Defined Medium (TDM, $(NH_4)_2SO_4$ 2 g, $MgSO_4{\cdot}7H_2O$ 0.6 g, K_2HPO_4 2.6 g, sucrose 1,71 g, 1 mL of minor elements solution ($FeSO_4{\cdot}7H_2O$ 1 g, $ZnSO_4{\cdot}7H_2O$ 1 g, $MnCl_2{\cdot}4H_2O$ 1 g, $CaCl_2$ 1 g, in 1 L, and Bacto Agar (BD Difco) 20 g). Coculture experiments were carried out by inoculating 30 μL of spores (10^7 CFU/μL) of *S. coelicolor* M145 in line on a thin layer on ISP2 medium.[16] After 24 h of incubation at 28 °C, 5 μL of an overnight culture of *E. coli* DH5α ($OD_{600\ nm}$ set to 0.1) was spotted in contact with *S. coelicolor* and incubated at 28 °C for 2 days. Induction of the virulome of *S. scabiei* 87-22 was monitored on TDM medium supplemented with sucrose (5 mM), and with either glucose (2.5 mM) or cellobiose (2.5 mM) as an additional carbon source. Finally, the effect of iron ($FeCl_3$ 10 μM) on the production of siderophores produced by *S. scabiei* 87-22 was monitored on Peptone Yeast Malt (PYM) agar medium.

Sample Preparation for Mass Spectrometry Imaging. Microbial cultures grown on the solid medium were excised directly from the agar plate and transferred to a microscope slide (VWR, USA), previously covered with double-sided conductive copper tape (3M, reference no. 1182/25). The

https://doi.org/10.1021/jasms.4c00129
J. Am. Soc. Mass Spectrom. 2024, 35, 1743−1755

slides were air-dried in an oven at 40 °C for 2 h. For MALDI MSI experiments using positive ion mode, a HCCA matrix solution was prepared at 5 mg/mL in 80/20 ACN/water doped with 0.1% FA. Ten layers of matrix were sprayed onto the slides with the SunCollect instrument (SunChrom, Friedrichsdorf, Germany). The first layer was sprayed at a flow rate of 10 μL/min. The flow rate was increased by steps of 10 μL/min after each layer until it reached 60 μL/min. The last four layers of sprayed matrix were applied using our maximum flow rate of 60 μL/min. The procedure of matrix deposition was repeated once again, which means that a total of 20 layers of matrix were sprayed per slide.

FT-ICR Mass Spectrometry. MSI data were acquired on a dual source ESI/MALDI solariX xR 9.4T instrument (Bruker Daltonics, Bremen, Germany) using a mass range of 200−1600 m/z in amplitude mode and a file size of 2 M (FWMH ± 300,000 @ 400m/z). This solariX xR 9.4T is an FT-ICR fitted with a high field supraconductive magnet, dynamically harmonized ICR (ion cyclotron resonance) cell ("paracell"), and 1ω detection capable of extreme mass resolving power (FWMH of 1 million and better at m/z 400) and mass accuracy better than 0.5 ppm. For each mass spectrum (pixel) of the image, the beam focus was set to "small" to reach a lateral resolution of the pixel's image around 100 μm. Twenty laser shots were performed at a repetition rate of 200 Hz. The laser power was tuned to produce a stable ion current between pixels as described elsewhere.[25] In this work, the laser power was set to 14%. Before imaging the sample, the mass spectrometer was externally mass calibrated from 200 m/z to 1600 m/z before each analysis using red phosphorus suspension in pure acetone and spotted directly onto the slides covered with the conductive tape. The laser power was adapted to produce a similar amount of ions (total ion current) than those obtained from the sample. FlexImaging v5 (Bruker Daltonics, Bremen, Germany) software was used to design the imaged areas for MALDI-MSI acquisition, using a pixel step size for the surface raster set to 200 μm. For supplementary sets of parameters of the FT-ICR used for performing MSI, see supplementary data (see Figure S1).

Data Preprocessing and Filtering. The data were first exported into the imzML files[26] of each image with Fleximaging 5.0 (Bruker). Each continuum mass spectrum was then individually (i.e., each pixel) converted to a centroid mass spectrum. The mass spectra of one image of reference were recalibrated using the m/z of HCCA clusters, and then the average spectrum of the reference image was generated; the m/z of the peaks in this spectrum was used to recalibrate all the other images using an algorithm published by La Rocca et al.[27] The MS images have been separately analyzed by different filtration algorithm differential signal filtering (DSF) and image structure filtering (ISF) methods (discussed below). All images reported in this work were generated using SciLs 2016b (Bruker Daltonics, Bremen, Germany). All figures reported in this work will adhere to the SMART reporting standards to enhance the transparency and reproducibility of the experiments.[28]

Differential Signal Filtering (DSF) Method. The DSF algorithm has been integrated in Mass Spectrometry Kendrick Filter (MSKF), an in-house Python3 software suite allowing the filtration of mass spectrometry data based on the Kendrick mass defect.[29] MSKF software reprocesses the imzML data into ICK files. Briefly the ICK files are Python object files that contain a DataFrame of all signals detected in the MSI data conserving the pixel localization (i.e., X- and Y-position), the mass-to-charge m/z, and the signal intensity. All the MSI data in the imzML format were recalibrated using the software from La Rocca and co-workers[27] before conversion into to ICK files, as described elsewhere.[30−32] Noise reduction, based on a minimum intensity threshold, was applied to remove the background noise in each MS spectrum (each pixel). This intensity threshold was set high enough to remove the residual MS noise, but without suppressing the potential weakest signals intensity of the species of interest. After manually reviewing randomly selected pixels on a few images, a minimum intensity of 10^5 counts per scan (cps) was retained (maximum signal intensity in a pixel was around 5×10^7 cps). The MSKF differential signals filtering tool was used to extract differential signals between a set of ICK files (e.g., condition A vs condition B, ...) and generated a new filtered ICK file. This algorithm is based on a signal (m/z peaks) subtraction between data sets (e.g., condition A − condition B − condition C), retaining only the m/z peaks detected in the first data (unique signal in condition A in regards of condition B and condition C) and disregarding their intensity values. The software requests an additional user-defined m/z tolerance window to consider the possible mass shift events and decide if the peaks corresponded to the same ions or not. Additionally, the filtering could also impose an additional parameter defined by a minimum number of pixels that the signal has to occur on the whole image (e.g., the m/z peak need to be detected in more than 100 pixels), minimizing the risk of considering chemical noise due to, e.g., single-pixel hot spot.

Image Structure Filtering (ISF) Method. The ISF algorithm is an in-house software combining a python 3 script and R script. From the imzML recalibrated MSI data set, the mean spectra of the MSI experiments for each case studied were computed. Peak picking was performed within a contiguous mass window of at least 0.01 m/z. In the case where several m/z peaks are found in this window, only the most intense one is conserved. A list of m/z containing the 10^4 most intense m/z peaks detected in the mean spectrum is generated. An ion image is generated for each of these peaks and for each MS image to be compared. This is performed by iterating through each pixel and reporting the maximum intensity value detected within the ±0.01 m/z window around the m/z from the peak list. Once all the MSI pixels have been parsed, a list of intensity values that can be rearranged into an ion image from the MSI pixel coordinates (X- and Y-positions) is obtained. The images generated are scored between 0 and 1 to assess the topological coherence of the signal, also described as structure, using an algorithm that evaluates the level of structure of the ion image. A score of 0 corresponds to a complete absence of structure and a totally random distribution of pixel intensities across the analyzed space, while a score closed to 1 corresponds to a structured image, where large regions of pixels have very close intensity values (see examples in Figure 4). The current approach to compute the level of structure is highly based on the measure of spatial chaos algorithm, developed by Palmer and co-workers,[33] which was initially inspired by the concept of persistent homology.[34] The score can be exploited to evaluate the variation of spatial distribution between 2 ion images by computing the fold change of structure scores (score of the tested condition divided by the score of the control condition).

To provide context to the processing time of the resulting data, the algorithm was run on a single processor workstation

https://doi.org/10.1021/jasms.4c00129
J. Am. Soc. Mass Spectrom. 2024, 35, 1743−1755

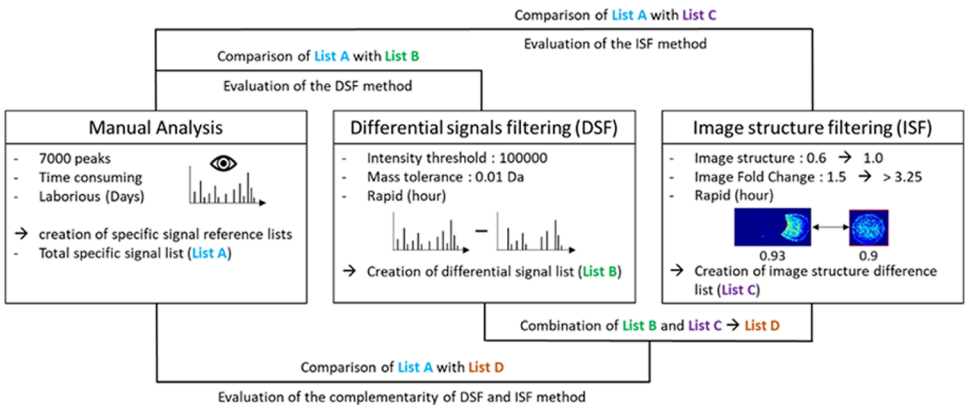


Figure 1. Methodology for filtering and comparing specific signals depending on the filtration method (DSF and ISF). Manual analysis generates List A (blue), which serves as reference lists for methods evaluation. Differential signals filtering (DSF) and image structure filtering (ISF) generates, respectively, Lists B and C, allowing the comparison to be made. Finally, List D is created by combining Lists B and C to evaluate the complementarity of the methods to identify specific signals.

containing an Intel(R) Core(TM) i9-9920X CPU @ 3.50 GHz. For one imzML file, it takes less than 40 min to complete the processing. The algorithm runs independently without requiring supervision, allowing time to focus on other tasks.

Code Availability. Software to perform the image structure filtering is implemented in an R package. The software and installation instructions are available on GitHub at https://github.com/LaRoccaRaphael/ISF.

RESULTS AND DISCUSSION

The currently available software, including both academic and commercial codes, mainly focuses on processing the MS image data through targeted mode or untargeted mode based on statistical analyses of the data (multivariate analysis, data clustering, ROI, etc.), or employs brute force peak annotation (requiring an existing and compatible database). None of these approaches provide satisfactory data processing when applied to the field of chemical communication and chemical interaction between bacteria, fungi, and even plants, particularly in the context of studies to discover new compounds. Indeed, these solutions are not specifically designed for identifying m/z signals that are uniquely present in one data set compared to control samples. This specific capability is essential for accurately distinguishing the unique chemical signals emerging from interactions within these biological systems. Scientists have to then use the basic features of the MSI software such as data visualization, data segmentation, and peak list generation, that need to be manually reviewed and processed. These are tedious and time-consuming tasks where human errors can likely occur due to, among other things, the large (thousands of m/z peaks) peak list to handle, significant chemical background noise, disparity of the cations adducts, or ion suppression effects. Here, the idea is based on the fact that unique chemical signals are emitted exclusively during interaction between two bacterial (or other) species and not in monocultures. So, to isolate these interaction-specific signals, the list of m/z features identified under monoculture conditions (serving as a negative control) is compared from the list obtained under coculture conditions. This makes it possible to distinguish only the differential signals specific to the interaction. Moreover, one of the strengths of MSI lies in its ability to discern between compounds and the random chemical background, leveraging the principle that compounds produced during interactions are not scattered randomly but are organized into structured ionic images. This structured nature of the compound ionic images serves as a foundational concept for isolating signals specific to interactions, thereby enhancing the accuracy of the identified chemical signatures unique to molecular crosstalk.

Methodology. Coculture experiments between *Escherichia coli* and *Streptomyces coelicolor*, two well-known bacterial species, were carried out in 5 replicates. Both species have been extensively studied and therefore benefit from well-established metabolomic databases.[35−37] Furthermore, preliminary assays have revealed that specific compounds were generated as a consequence of the interaction between these two-model species.

The differential signal filtering (DSF) and image structure filtering (ISF) methods were challenged for identifying specific signals in comparison to the manual review of the data set of MS images. A detailed and methodical process is described in Figure 1. The initial step involves an exhaustive manual review of all ions present in the images using SciLs. This manual analysis generates one reference list that contains all the m/z features corresponding to all ions detected in one specific condition, the interactions between the two bacterial strains. The m/z features, grouped in List A, are retained only if the m/z peaks were detected in at least 4 out of the 5 replicates. These manual analyses were conducted for benchmarking purposes, providing a reference to compare and validate the performance of the two methods. The comprehensive nature of manual analysis ensures that we have a robust data set against which to measure the effectiveness of our automated approaches. Thus, the automated workflows described in our study are designed to streamline the process, offering a more scalable solution by reducing the time and effort required for analyzing MS images. We plan next to apply these methods in the two next study cases, allowing us to further validate their effectiveness and demonstrate their practicality in various experimental contexts.

Manual Analysis of the Interaction between *Streptomyces coelicolor* and *Escherichia coli*. In a first phase of the study, we identified in the MS images all (or at least considered as such) the signals (i.e., m/z peaks) specifically induced during the bacterial interaction. To do so, the MS images obtained from the three controls (noninoculated ISP2 medium, and ISP2 inoculated either with *S. coelicolor* or with *E. coli* alone) and from the tested condition (coculture of *S.*

https://doi.org/10.1021/jasms.4c00129
J. Am. Soc. Mass Spectrom. 2024, 35, 1743−1755

coelicolor and *E. coli* on ISP2 medium) were first processed with SciLs software. The average spectrum corresponding to the coculture condition was then investigated to look for signals specifically induced by the interaction between bacteria with a high confidence level. For this purpose, the following criteria were required to postulate the specificity of a signal implied during the interaction:

- The ion must be exclusively present in the mass spectra of the tested condition (interaction).
- The resulting ion image had to be recorded in a minimum of 4 out of 5 biological replicates; the topological coherence of the ion image structures was visually assessed by the operator.

MSI is generally speaking not a quantitative method, especially considering the label-free techniques, despite substantial efforts focusing on tissue sections during the last years.[38−43] In the absence of a quantitative method that could operate directly on agar media, we decided to retain the signals detected only during the interaction between the bacteria. Consequently, we omitted the signals variation linked to the overexpression or repression of the compounds, initiated by the coculture conditions, but in return, we preserved our data from biases and ambiguities.

Figure 2A shows the optical images obtained for five biological replicates of samples from control (ISP2 medium), from monocultures (*E. coli* and *S. coelicolor* grown alone), and from the interaction between the two bacteria. The colonies are delimited using dashed lines for convenience. Figure 2B and C shows the ionic images of two ions, m/z 588.263 and m/z 609.329, which are present in the interaction samples but not in the control conditions. These ions are part of List A (the peak list of all detected ions in the interaction samples but not in the controls). The signal detected at m/z 588.263, secreted by *S. coelicolor*, and m/z 609.329, produced by *E. coli*, were exclusively detected when the two bacteria were cocultured.

About 7000 peaks were present in the average spectrum of the MS images generated in SciLs. The 205 m/z signals were specific to the tested condition when compared against controls (medium and monoculture), and these were compiled into List A which represents about only 3% of all signals manually filtered. It is important to emphasize that, within this list of signals, a large proportion presents highly structured ionic images, offering notable consistency in relation to the samples analyzed. This consistency reflects the specificity of chemical interactions within the cultures studied, but also underlines the reliability of the selection of signals for List A. However, it is also crucial to recognize that some signals within this list are less structured, generally characterized by lower intensity. All the images are present in the supplementary files (Supplementary file_2). Thus, the manually processed data leading to List A constitutes the reference set allowing the benchmarking of our proposed filtering bioinformatics tools in terms of performance and accuracy.

Evaluation of the Differential Signals Filtering (DSF) Method. The DSF method, which drastically speeds up the data filtering, was initially used to extract the ions of interest, appearing in the context of the bacterial interaction between *E. coli* and *S. coelicolor*. The MSI data corresponding to this interaction were filtered with successive subtractions of the MSI data of the blank (noninoculated ISP2 culture medium) and negative controls (monocultures). For example, the signals of the MALDI matrix, present in all data sets, were

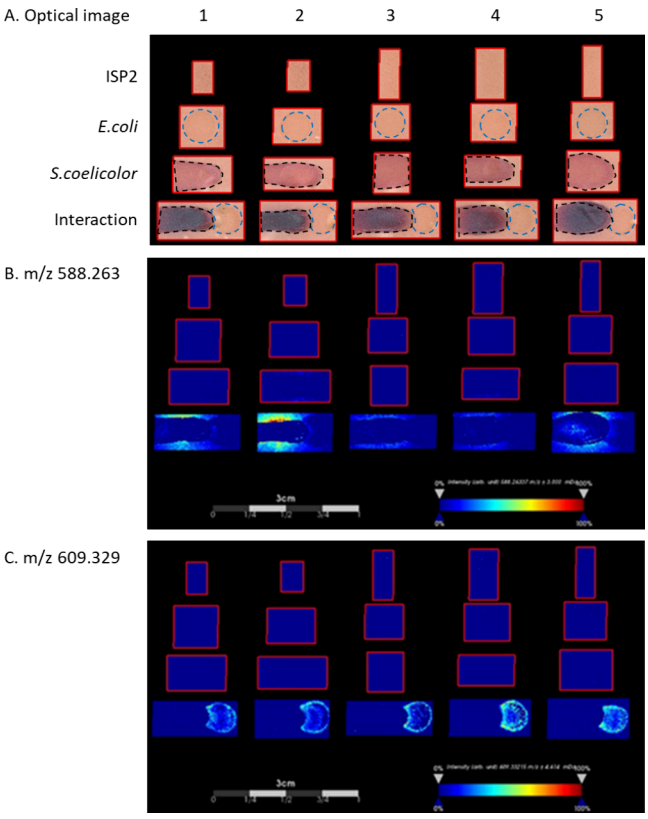


Figure 2. Study of the interaction between *S. coelicolor* and *E. coli* by MALDI mass spectrometry imaging. (A) Optical images obtained for the five biological replicates. Colonies are delimited using dashed lines for convenience. (B) m/z 588.263 and (C) m/z 609.329, representing two of the ions preserved after manual filtration. Scale bars (gray and white) represent the real size of the images. The color bar is coding the relative ion intensity (from blue as 0% to red as 100%) of the extracted ions for each pixel of the ionic images. Bacteria colony limits are represented by dotted blue lines for *E. coli* and black lines for *S.coelicolor*.

automatically removed during the filtering steps. Figure 3 illustrates the signal subtraction processes on a restricted mass range (m/z from 390 to 400) for clarity. After the subtractions of the m/z peaks from the three controls (noninoculated medium and the two monocultures, Figure 3B, D, and F), only signals corresponding to the m/z peaks specific to the interaction between both microorganisms remain (Figure 3G). Each of these peaks resulting from the coculture can then be extracted and their corresponding images reconstructed in SCiLs. The metabolite detected at m/z 394.286 is illustrated as an example (Figure 3H), demonstrating the capacity of the filtering method to highlight interaction-specific signals.

Each of the five biological replicates generates its own mass lists, containing exclusively the compounds produced during the bacterial interaction for a given replica. To maintain consistency with the manual analysis approach, the same criteria of selection was kept and only the m/z signals retrieved in a minimum of four replicas were retained for further data analysis. This method resulted in the selection of 107 signals, constituting List B (Figure 1) (Supplementary file_1). From these 107 m/z peaks, 87 are also among the 205 signals of List A (total specific signal list).

https://doi.org/10.1021/jasms.4c00129
J. Am. Soc. Mass Spectrom. 2024, 35, 1743−1755

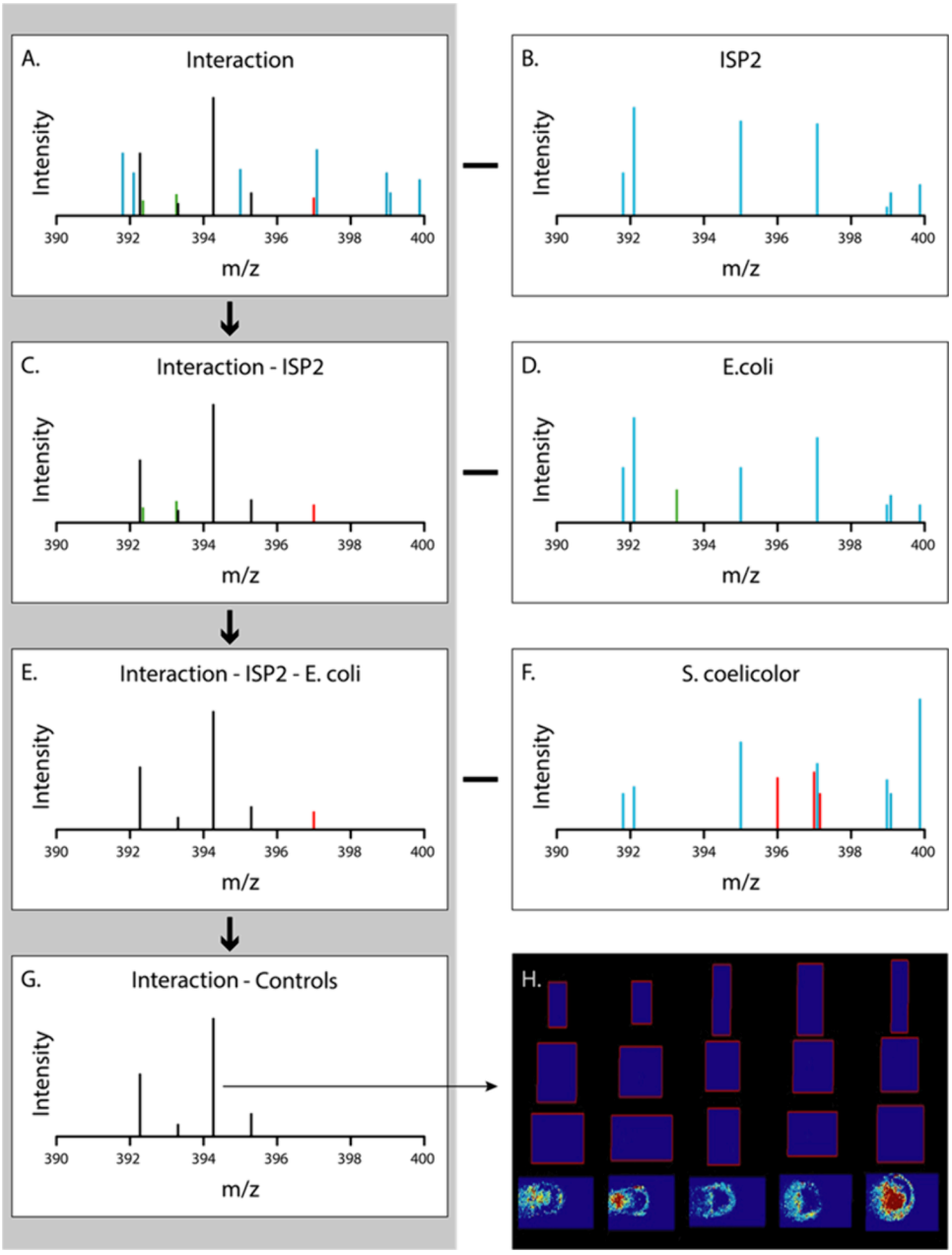


Figure 3. Overview of the differential signal filtering method processing. (A) Zoom-in of the MSI average mass spectrum between m/z 390 and 400 of the interaction between *S. coelicolor* and *E. coli.* (B,D,F) Mass spectra of the control (no bacterial interactions) conditions, i.e., the noninoculated ISP2 culture medium, the monoculture of *E. coli* on ISP2, and the monoculture of *S. coelicolor* on ISP2, respectively. (C,E,G) Processed mass spectra after successive subtraction of controls presented in panels (B), (D), and (F), respectively. Signals (m/z peaks) are colored based as their function of origin: blue, culture medium ISP2; green, *E. coli* monoculture; red, *S. coelicolor* monoculture; black, signal specific to bacterial coculture (interaction). (H) MSI image of a specific signal (394.286 m/z) obtained after filtration where the relative intensity of the signal is color coded from blue (0%) to red (100%).

Remember that the DSF method applied an intensity signal threshold and a mass shift tolerance during the filtration of the MSI data (see Methods): the less intense ions could be lost due to the applied threshold, and closely related m/z values from 2 different MSI data set to be subtracted could be considered as a unique and same feature by the algorithm included in the homemade MSKF software, depending on the user settings. The introduction of false negatives and false positives can be avoided using a better evaluation of the signal threshold with regard to the background noise in the MSI data and the reduction of the applied mass tolerance. The range of the mass tolerance window has to be considered with regard to the MS instrumentation, MS methods, and MS performance during the acquisition of the MSI data. The mass tolerance can be greatly reduced, but it requires high quality data such as that delivered by MALDI-Orbitrap and MALDI FT-ICR instruments fitted with a high field supraconductive magnet, dynamically harmonized ICR cell ("paracell"), and 1ω or 2ω detection, which limit the mass shift and provide sub-ppm or better mass accuracy (sub-ppm mass accuracy ensures 6 significant digits of the floating-point m/z value). However, it is noteworthy that data realignment or recalibration serves as

https://doi.org/10.1021/jasms.4c00129
J. Am. Soc. Mass Spectrom. 2024, 35, 1743−1755

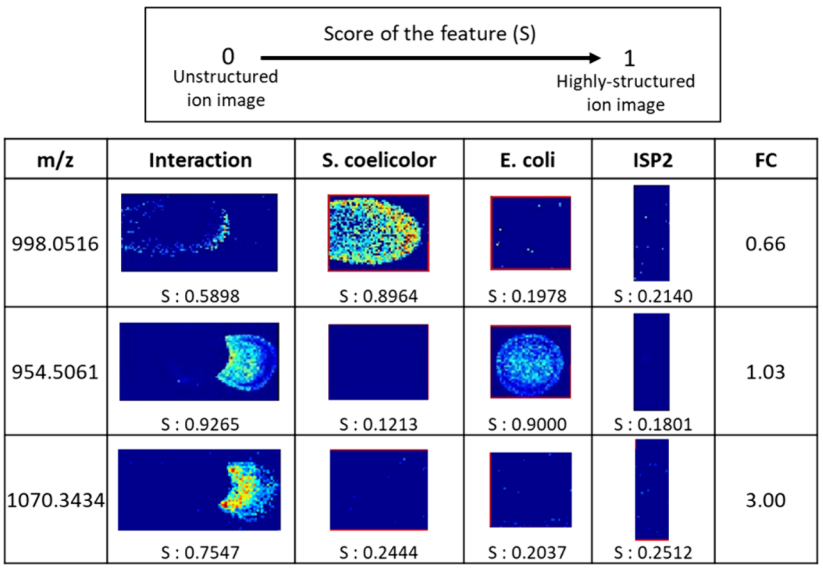


Figure 4. Overview of the image structure filtering method. The legend ranging from 0 to 1 shows that scores close to 0 correspond to an unstructured ion image and 1 to a highly structured ion image. To illustrate this, three m/z values were picked (m/z 998.052, 954.506 and 1070.343) and the images and scores (below the image) of every condition (Interaction, *S. coelicolor*, *E. coli* and ISP2) for one biological replicate are shown. Finally, Fold Change score (FC) is computed by comparing scores of the interaction over the scores of the highest control condition.

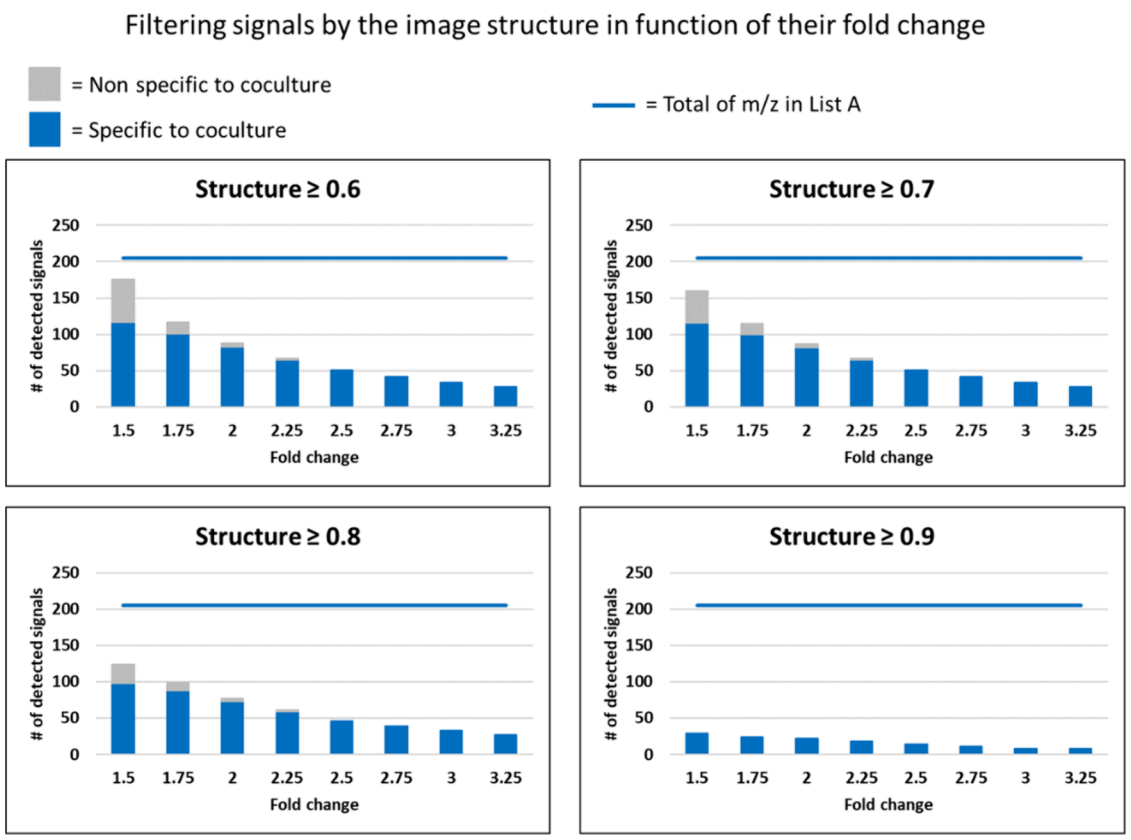


Figure 5. Histograms comparing the total specific m/z depending on the image structure in function of the fold change. The blue line represents the number of signals in List A. The histograms show the number of signals (# of m/z) retained as a function of the parameters applied (fold change of structure). The blue part of each histogram indicates the signals and thus the m/z present in List A, while the gray parts show the signals considered as false positives.

an effective strategy to minimize the mass shift effect or improve the mass accuracy, respectively.[44] The evaluation of the spectrum noise background is evaluated by manual evaluation of the average noise around dominant m/z peaks observed in some ion image pixels randomly selected from the MS image.

Evaluation of the Image Structure Filtering Method (ISF). The second filtration method adopted is based on an approach particularly well adapted to MSI data set. This

https://doi.org/10.1021/jasms.4c00129
J. Am. Soc. Mass Spectrom. 2024, 35, 1743−1755

method, based on the topological coherence of the image (also defined as structure of the image), assigns a score to the structure of each feature (i.e., an extracted m/z value from the ionic image). See in the material and methods for details. Briefly the score tends to 0 in the absence of structure and reaches a maximum of 1 for a very structured ion image. For illustrating this, Figure 4 shows the structure scores of three different m/z values in a biological replica. These scores are then used to calculate a new parameter, called the "fold change structure", representing the ratio of the highest score of the interaction to the highest score obtained in the controls. This calculation is performed on the 10^{+4} most intense features, generating a table including all masses (m/z peaks) according to two criteria: the image structure of the interaction and the fold change (FC).

Figure 4 additionally shows the ionic images of selected m/z values to illustrate the varying fold changes that can be obtained by comparing their structural scores. Indeed, for the first m/z 998.052, a fold change (FC) of 0.66 is calculated. As this score is below 1, the image structure is better in at least one of the negative controls, here *Streptomyces coelicolor*, compared to that of the interaction (coculture). This suggests that the metabolite is potentially less abundant in the coculture condition compared to the monoculture (bacterium grown separately). Indeed, the intensity usually reflects the abundance of the metabolites and can contribute to higher image structure, as it often affects the topological coherence of the signal. For m/z 954.506, a FC of 1.03 is calculated, which indicates no significant change in the abundance of this metabolite between the sample resulting from the interaction experiment and the bacterial monoculture. Indeed, the associated images do not show a notable difference, confirming that the interaction does not affect significantly the relative intensity of the detected ion corresponding to this metabolite. For m/z 1070.343, an FC value of 3.00 is reached, suggesting an induced production of this compound during the coculture, which is indeed validated by the visual evaluation of the images.

The underlying hypothesis of this filtering method is that if a signal is specific to a condition being tested, it will have a high structure score, and the ratio of fold change (FC) will indicate the condition to which the improvement of the structure score is associated. In evaluating these 10,000 features, we can retain only signals with a structure score for the condition tested of our interest in regard of to their fold change.

Thus, Figure 5 shows the results obtained using the ISF filtering method, conserving the images with structure scores higher than 0.6, 0.7, 0.8, and 0.9 only. The blue line indicates the number of peaks classified into the reference Lists A, representing thus the target to achieve. Considering a structure score greater or equal than 0.9 and a fold change greater or equal than 1.5, only 29 significant signals remain, with all of them initially included in the List A. This drastic structure fold change score exemplifies how the methodology can automatically identify peaks exclusively specific to the interaction condition. By tuning the structure and fold change parameters (for instance, signals with a structure score greater than 0.7 and a fold change greater than 1.5), the number of signals obviously increases. For these new arbitrarily selected parameters, 160 significant signals are identified (List C, see Supplementary file_1), among which 115 were already present in List A (*list of specific signals determined with a manual analysis*). These signals account for 56% of all candidate ions characterized by a production response unique to the bacterial interaction. This filtration method is therefore of great interest not only because it automatically extracts most of the significant signals from the huge amount of data, but also because it allows the parameters to be modulated by the operator, offering a flexible and efficient approach to the analysis process.

Obviously, when the fold change exceeds 1, the number of ions specific to the bacterial interaction increases, suggesting that a higher FC effectively filters out false positives. In the examples shown in Figure 5, a FC score of below than 2.25 makes false positives signals appear (signals in gray in Figure 5). Furthermore, when the base score of the structure is high (0.8 and 0.9), the total number of ions specific to the coculture is of course lower but strongly relevant. Here is now the trade-off for users: Choosing higher structure scores ensures reliable results but could lead to the exclusion of relevant ions that would have been selected at lower thresholds. Conversely, lower structure scores can identify a larger number of ions, but this also includes a higher likelihood of detecting false positives, thus necessitating careful evaluation of the MSI data.

The ISF method thus exhibits certain limitations, particularly regarding false negatives and false positives, due mainly to too selective signal criteria. On one hand, false negatives can be highlighted for two primary reasons. First, the calculation of the structural fold change, a critical parameter for distinguishing significant m/z, relies on comparing the highest structural score of the interactions among the 5 replicas against the highest score of all the control experiments. If a structural score in a control is abnormally high, it can lead to false negatives. Second, the ISF method was designed to retain the 10,000 most intense signals. This constraint could induce biases by ignoring minor m/z peaks in the MS data. On the other hand, several factors also explain the occurrence of false positives. In contrast to false negatives, if the structure scores in the controls are too low, the ratio between the structure scores can lead to false positives. Moreover, the current ISF method does not consider our last criteria which is an ion having a structure should appear for 4 of our 5 replicates. For example, the compound detected at m/z 600.3204 was computed with decent ISF and FC scores but was detected in only 3 of our 5 biological replicates (see Figure S2). This compound was assigned as false positive, maybe wrongly.

Synergy between the Differential Signal Filtering and the Image Structure Filtering Methods. Ultimately, our aim was to analyze whether these two filtering methods were complementary and this led to a more efficient filtering of the detected features in the MSI data set. The 107 m/z signals obtained by the differential subtraction filtration (DSF) were merged with the 160 masses obtained by the image structure filtration (ISF) by applying a threshold structure parameter greater than 0.7 and a fold change greater than 1.5. This combination resulted in a total of 204 m/z signals, referred to as List D. Among these 204 m/z signals in List D, 139 were also found in List A (specific signal list, manual filtering) (Supplementary file_1).

Moreover, more of these specific signals occurred if isotopologues were also considered. The disparity in signal intensity between the ion at the monoisotopic mass and the respective isotopologues typically results in a lower percentage of covering of List A by the different filtering methods. Indeed, isotopologues have a greater chance to exhibit lower structural score via the ISF method or are most likely to be excluded by

https://doi.org/10.1021/jasms.4c00129
J. Am. Soc. Mass Spectrom. 2024, 35, 1743−1755

the threshold applied in the DSF method compared with the corresponding monoisotopic ion. It is important to note that the DSF-ISF combined approach filtered and retrieved, within 3−4 h, 79.0% of the signals specific to the case study, whereas the manual data filtering required days of continuous workdays. Among these ions identified through the filtering process, this mainly corresponds to signals presenting a certain intensity with ionic images possessing notable characteristics, such as a relative abundance around or inside the colonies and diffusion from the colonies. The ions missing from this filtration compared to reference List A, primarily, correspond to signals of low intensities, which typically have a weaker structure and are generally less visually interpretable, highlighting the efficiency and precision of the DSF-ISF approach in prioritizing signals with significant biological relevance.

Figure 6 proposes a Venn diagram illustrating the synergic effect of both filtering method. Combined DSF-ISF filtering

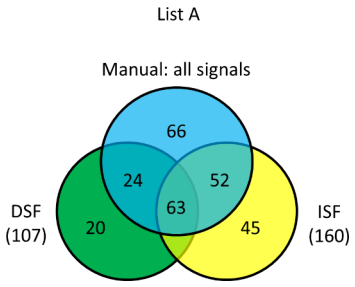


Figure 6. Venn diagrams compile the two new methods (DSF an ISF) of filtration in comparison with manual data analysis (all signals from the averaged mass spectrum in the MSI data).

shares 63 common features with List A, while additionally providing their own masses. The 24 and 52 signals unique to the DSF and ISF lists, respectively, are not false positives, as they are present in List A. These signals could only be detected by one of the filtering methods, highlighting the need for orthogonal methods to extract the maximum information from the complex data obtained from mass spectrometry images. While the filtering methods have proven efficient in reducing the data from several thousand features to only a couple of hundred m/z, they are not perfect. Some unique features might be lost due to parameters such as the intensity threshold applied in the DSF method or the fold change threshold chosen in the ISF method. It does not feel like there is a specific trend in the unique features found by the two methods. This underscores the importance of using complementary approaches to ensure comprehensive data analysis.

However, the 20 and 45 m/z signals that are fully unique to the DSF and ISF lists, respectively, are indeed considered false positives. This can be explained by several aspects of the filtering methods. For the DSF method, the false positives can arise from the signal threshold applied during filtration. While, on the other hand, in the ISF method, false positives may result from a too low structure score in the control samples or failing to meet one of the criteria, which requires that an ion with a structure should appear in 4 out of our 5 replicates (see Figure S2).

Challenging the Methodology with Well-Described Case Studies. A major challenge in modern microbiology is to identify the environmental elicitors and nutrients that modulate the expression of genes associated with the secondary/specialized metabolism of bacteria and fungi. Finding these elicitors and the characterization of the metabolomic response could also lead to the identification of novel molecules associated with the numerous cryptic biosynthetic gene clusters (BGCs). In that context, we challenged the DFS and ISF methods on two well-described culture conditions known for either activating or repressing the production of specialized metabolites. The production of these molecules is a costly process that requires tight expression control of BGCs for which transcription is activated or repressed upon sensing specific environmental signals. The two chosen on/off models are (i) cellobiose-mediated induction of the pathogenic lifestyle of *Streptomyces scabiei*, the causative agent of common scab on root and tuber crops, and (ii) iron-repressed production of siderophores of *S. scabiei*.

First Case Study: Cellobiose-Mediated Induction of the Pathiogenicity of S. scabiei. Thaxtomin phytotoxins are the main virulence determinants of *Streptomyces scabiei*. Their production is mandatory for pathogenicity and is induced by the transport of the cellulose byproducts: cellobiose and cellotriose.[45,46] Deflandre and co-workers[47] recently studied the composition of the virulome of *S. scabiei* in response to cellobiose. This study revealed by RNA sequencing that, in addition to the BGC responsible for producing thaxtomins, the expression of nine cryptic BGCs was activated by cellobiose.

We applied both DSF and ISF approaches to evaluate how they facilitate the identification of thaxtomins and to possibly discover candidate ions of new metabolites of cryptic BGCs for which production would also be under the control of cellobiose.

S. scabiei was grown on agar medium to permit the bacteria to colonize the entire surface of the plate. The production of BGCs was induced by diffusion of cellobiose from a soaked square-shaped piece made of paper deposited directed on the surface of the agar medium. Finally, water- and glucose-soaked papers were used as negative controls (Figure 8A). The combined DSF-ISF method allowed retention of 393 ions from a total of more than 4000 ions. Most of them were identified to be specific to the cellobiose condition (Figure 7 and Supplementary file_3−4). Among these signals, various ionization adducts of thaxtomins A and B were easily identified (Figure 7B,C), which confirmed that our filtering method allowed retention of ions of molecules regarded as positive controls for cellobiose-induced metabolites.

This synergistic data filtering also revealed specific signals that could not be attributed to any known metabolite produced by *S. scabiei* upon cellobiose-mediated induction, such as the metabolites detected at m/z 934.220 ($[M + Na]^+$) and m/z 950.194 ($[M + K]^+$) (Figure 7D). These metabolites could actually be a combination of cellobiose and HCCA matrix, with some sodium/potassium adducts indicating a complex interplay of molecular interactions and ionization processes to the conditions under study.

Second Case Study: The Iron-Mediated Repression Production of Siderophores of S. scabiei. Our second case study consisted of examining how iron regulates metabolism with *S. scabiei*, specifically looking at siderophore production in relation to iron presence in the medium. Although many siderophores of *S. scabiei* have already been identified (desferrioxamines, pyochelin, scabichelin/turgichelin), others, not yet described as siderophores, have been predicted through genome mining.[47] Siderophore production is classically controlled by the availability of iron; iron depleted environment triggers their biosynthesis while iron overload represses

https://doi.org/10.1021/jasms.4c00129
J. Am. Soc. Mass Spectrom. 2024, 35, 1743−1755

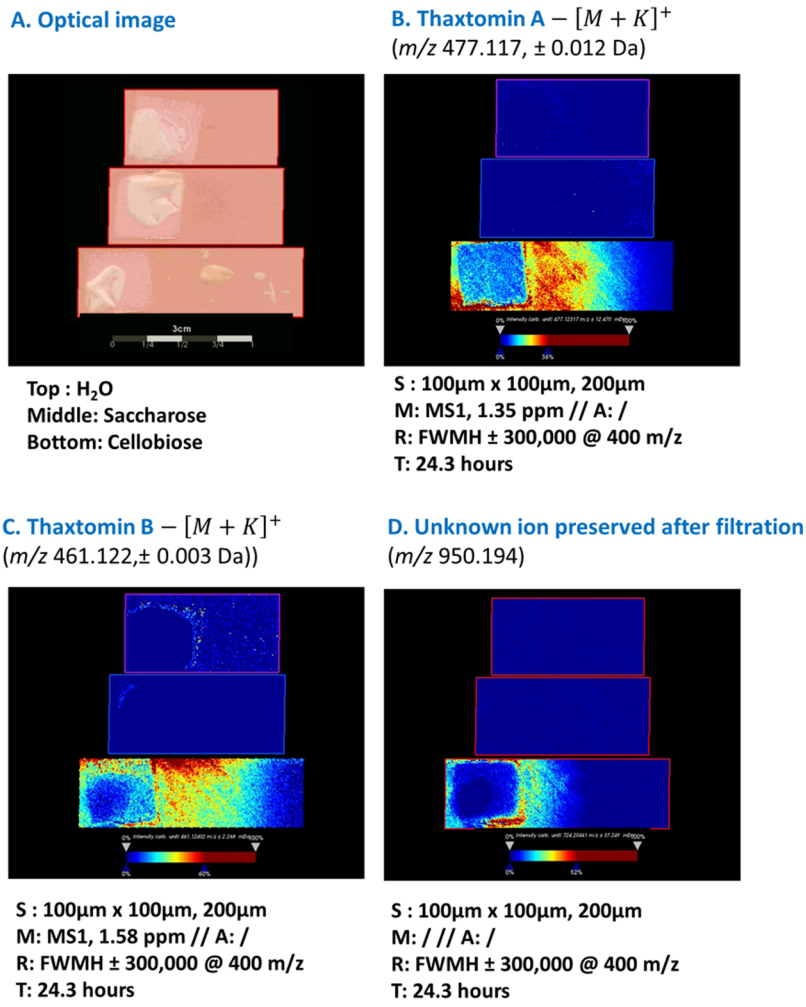


Figure 7. (A) Optical image of *Streptomyces scabiei* grown on minimal agar medium, covering the entire, with a piece of paper soaked with water, saccharose, or cellobiose. (B) 477.117 *m/z* corresponding to [Thaxtomin A + K]$^+$, (C) 461.122 *m/z* corresponding to [Thaxtomin B + K]$^+$, and (D) 950.19 *m/z*. corresponding to an unknown ion appearing preserved after filtration.

their production.[48] In this context, we evaluated the combined DSF-ISF approach by MALDI mass spectrometry imaging (Figure 8) for screening the presence of features associated with siderophores productions by *S. scabiei*, cultivated under conditions either triggering or repressing the production of iron chelating molecules. The strain was cultivated on peptone−yeast−malt extract (PYM) medium without or with iron ($FeCl_3$ 10 μM). Applying the DSF-ISF filtering method led to a total of 749 selected signals (out of more than 6,000 features) that disappear in the presence of iron. A majority of these signals display the expected contrasted production pattern, i.e. no signal in the presence of iron (Supplementary file_5−6). Among them, ions corresponding to *S. scabiei* siderophores desferrioxamine B (Figure 8B), desferrioxamine E (Figure 8C), and scabichelin (Figure 8D) were easily detected.

Moreover, we observe ions such as *m/z* 498.266 (Figure 8E) and *m/z* 514.240 corresponding, respectively, to a not yet identified compound with a sodium and potassium adduct, which both display iron-repressed production patterns similar to the siderophore produced by *S. scabiei* but cannot be related to any known siderophore of this strain when looking in database such as NPatlas[49] or StreptomeDB.[35] The distinctive ion signatures and their iron-repressed production patterns, similar to known siderophores yet unidentifiable in current databases, suggest these compounds could represent novel siderophores synthesized by *S. scabiei*. This possibility underlines the importance of further investigation into these compounds.

CONCLUSIONS AND PERSPECTIVES

In this study, two computational methods for filtering mass spectrometry imaging (MSI) data were introduced and evaluated, with a particular focus on untargeted microbiological applications. The differential signal filtering (DSF) and image structure filtering (ISF) methods were meticulously assessed through coculture experiments and the appropriate negative controls, revealing their significant potential in streamlining the data analysis process of MSI, especially in the context of molecule discovery and microbiological applications. Our method was evaluated with regard to a set of data filtered manually, which could itself includes biases due to human errors. Additionally, we used "only" one manually filtered data set as a reference data set to benchmark our method. Confronting the method to a larger set of trusted reference data would be desirable. If we could use more data sets with high confidence levels of the annotated compounds, we are envisioning a better refinement of the user-defined

https://doi.org/10.1021/jasms.4c00129
J. Am. Soc. Mass Spectrom. 2024, 35, 1743−1755

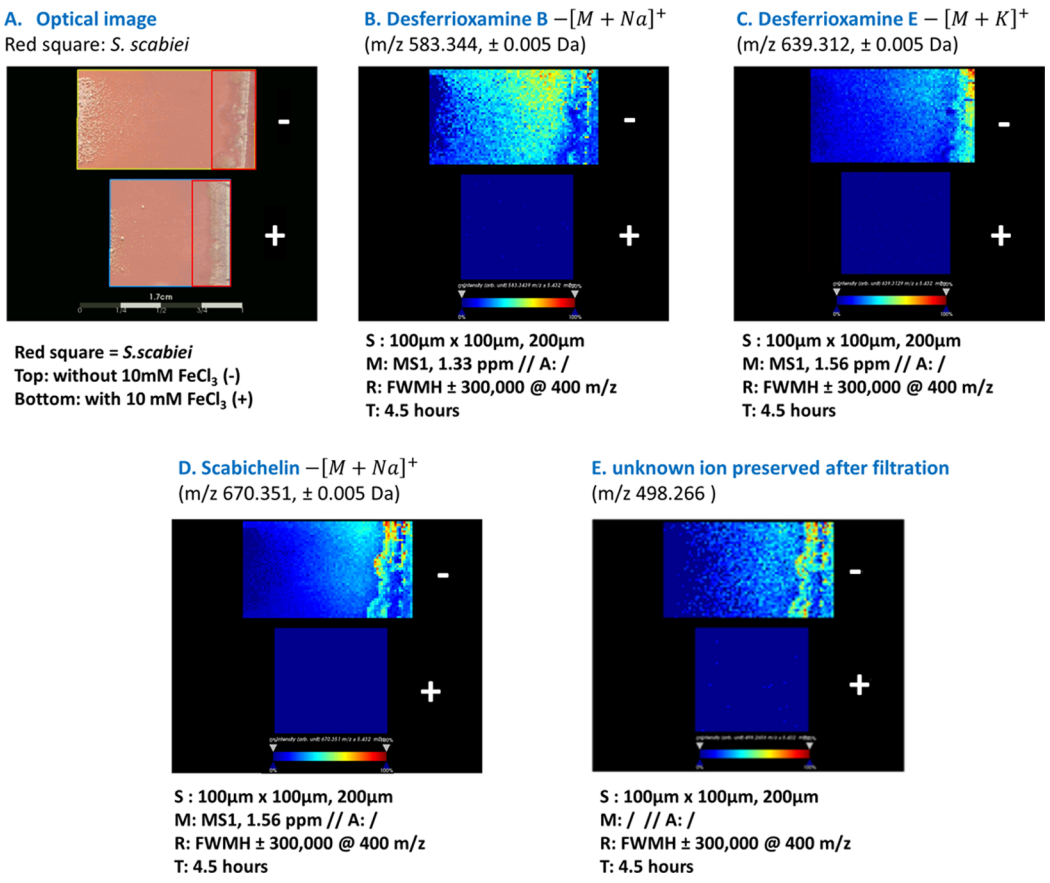


Figure 8. (A) Optical image of *Streptomyces scabiei* cultivated on peptone−yeast−malt extract (PYM) supplemented without (top) and with 10 μM $FeCl_3$ (bottom). (B) m/z 583.344 corresponding to $[Desferrioxamine\ B + Na]^+$ ion, (C) m/z 639.312 corresponding to $[Desferrioxamine\ E + K]^+$ ion, (D) m/z 670.351 corresponding to $[Scabichelin + Na]^+$ ion, and (E) m/z 498.266 corresponding potentially to an unknown siderophore with a sodium adduct.

parameters used by our algorithms and limit as much as possible the presence of false positives and false negatives.

The efficiency of these filtration methods has been demonstrated, particularly when used concomitantly, which allowed notably rapid screening and delineation of specific signals in atypical biological samples for mass spectrometry imaging. The time-saving aspect and easy operation to process these MSI data with regard to manual processing or other multivariate statistical or data clustering tools (see Figure S3) included in other software suite are therefore most welcome. In the absence of a quantitative *in situ* method by MALDI MSI applied for such agar-based culture medium, no induction or repression factor is yet accessible. This work offers new insights and perspectives in the study of microbiological samples, such as the secondary metabolism of bacteria, and data filtration of complex data can be particularly helpful for the adduct deconvolution of signals to identify metabolites, with an example of metabolite identification such as lyso-phosphatidylethanolamine 16:0 (see Figures S4−S6).

We aim to further enhance the capabilities of these computational methods. The development pipeline includes the implementation of isotope deconvolution, a feature that will further assist users during the metabolite identification process. This addition is anticipated to offer an even deeper understanding of the complex data generated by MSI, potentially broadening the scope of potential applications.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jasms.4c00129.

Parameters for the solariX XR 9.4T to perform MSI; example of some false positive obtained by ISF; data treatment using ROC between class 1 (controls) and class 2 (interaction between *E. coli* and *S. coelicolor*); filtration enables better adduct deconvolution of signals to identify metabolites; images of lyso-phosphatidylethanolamine 16:0 with different adduct and sharing similar image; experimental m/z and possible monoisotopic mass considering different adducts scenario where M1 = 475.2651 Da and M2 = 453.2848, and identification of lyso-phosphatidylethanolamine 16:0 by LC-MS/MS (PDF)

Comparison between different modes of analysis (manual, DSF, ISF) (XLSX)

False positives at different m/z values (starting at 218.993) (ZIP)

DSF cumulative intensity and ISF image values starting at m/z 203.032 (XLSX)

False positives at different m/z values (starting at 203.293) (ZIP)

DSF cumulative intensity and ISF image values starting at m/z 155.669 (XLSX)

https://doi.org/10.1021/jasms.4c00129
J. Am. Soc. Mass Spectrom. 2024, 35, 1743−1755

False positives at different m/z values (starting at 155.669) (ZIP)

AUTHOR INFORMATION

Corresponding Author

Loïc Quinton − *Mass Spectrometry Laboratory, MolSys Research Unit, University of Liège, 4000 Liège, Belgium;* orcid.org/0000-0001-8153-9590; Email: loic.quinton@uliege.be

Authors

Pierre Burguet − *Mass Spectrometry Laboratory, MolSys Research Unit, University of Liège, 4000 Liège, Belgium*

Raphaël La Rocca − *Mass Spectrometry Laboratory, MolSys Research Unit, University of Liège, 4000 Liège, Belgium;* orcid.org/0000-0003-3388-3041

Christopher Kune − *Mass Spectrometry Laboratory, MolSys Research Unit, University of Liège, 4000 Liège, Belgium*

Déborah Tellatin − *InBioS − Center for Protein Engineering, University of Liège, 4000 Liège, Belgium*

Nudzejma Stulanovic − *InBioS − Center for Protein Engineering, University of Liège, 4000 Liège, Belgium;* orcid.org/0000-0003-3243-4973

Augustin Rigolet − *Gembloux Agro-Bio Tech, University of Liege, 5030 Gembloux, Belgium*

Johann Far − *Mass Spectrometry Laboratory, MolSys Research Unit, University of Liège, 4000 Liège, Belgium;* orcid.org/0000-0003-1208-6262

Marc Ongena − *Gembloux Agro-Bio Tech, University of Liege, 5030 Gembloux, Belgium*

Sébastien Rigali − *InBioS − Center for Protein Engineering, University of Liège, 4000 Liège, Belgium*

Complete contact information is available at: https://pubs.acs.org/10.1021/jasms.4c00129

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors are grateful for the financial support from the University of Liège by the mean of the Concerted Action Research MIND (2020-2023) and for the FEDER BIOMED HUB (number 2.2.1/996) that funded the MALDI FT-ICR SolariX XR and the TimsTOFpro.

REFERENCES

(1) Ho, Y. N.; Shu, L. J.; Yang, Y. L. Imaging Mass Spectrometry for Metabolites: Technical Progress, Multimodal Imaging, and Biological Interactions. *Wiley Interdiscip Rev. Syst. Biol. Med.* **2017**, *9* (5), e1387 DOI: 10.1002/wsbm.1387.

(2) Feucherolles, M.; Frache, G. MALDI Mass Spectrometry Imaging: A Potential Game-Changer in a Modern Microbiology. *Cells* **2022**, *11* (23), 3900.

(3) Zaima, N.; Hayasaka, T.; Goto-Inoue, N.; Setou, M. Matrix-Assisted Laser Desorption/Ionization Imaging Mass Spectrometry. *Int. J. Mol. Sci.* **2010**, *11* (12), 5040−5055.

(4) Sturtevant, D.; Lee, Y. J.; Chapman, K. D. Matrix Assisted Laser Desorption/Ionization-Mass Spectrometry Imaging (MALDI-MSI) for Direct Visualization of Plant Metabolites in Situ. *Curr. Opin Biotechnol* **2016**, *37*, 53−60.

(5) Mitra, A.; Kant Choubey, S.; Das, R.; Rout, P.; Sridhar, T. S.; Mishra, D.; Kumar Mandal, A. Differential Expression of Proteins in Human Prostate Cancer Tissues Probed by MALDI Imaging Mass Spectrometry. *medRxiv*, 2020. DOI: 10.1101/2020.10.05.326686.

(6) Carter, C. L.; McLeod, C. W.; Bunch, J. Imaging of Phospholipids in Formalin Fixed Rat Brain Sections by Matrix Assisted Laser Desorption/Ionization Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2011**, *22* (11), 1991−1998.

(7) Rauser, S.; Marquardt, C.; Balluff, B.; Deininger, S. O.; Albers, C.; Belau, E.; Hartmer, R.; Suckau, D.; Specht, K.; Ebert, M. P.; Schmitt, M.; Aubele, M.; Hŏfler, H.; Walch, A. Classification of HER2 Receptor Status in Breast Cancer Tissues by MALDI Imaging Mass Spectrometry. *J. Proteome Res.* **2010**, *9* (4), 1854−1863.

(8) Alexandrov, T. MALDI Imaging Mass Spectrometry: Statistical Data Analysis and Current Computational Challenges. *BMC Bioinformatics* **2012**, *13* (Suppl 16), S11.

(9) Watrous, J. D.; Dorrestein, P. C. Imaging Mass Spectrometry in Microbiology. *Nat. Rev. Microbiol* **2011**, *9* (9), 683−694.

(10) Gonzalez, D. J.; Haste, N. M.; Hollands, A.; Fleming, T. C.; Hamby, M.; Pogliano, K.; Nizet, V.; Dorrestein, P. C. Microbial Competition between Bacillus Subtilis and Staphylococcus Aureus Monitored by Imaging Mass Spectrometry. *Microbiology (N Y)* **2011**, *157* (9), 2485−2492.

(11) Dunham, S. J. B.; Ellis, J. F.; Li, B.; Sweedler, J. V. Mass Spectrometry Imaging of Complex Microbial Communities. *Acc. Chem. Res.* **2017**, *50* (1), 96−104.

(12) Yang, J. Y.; Phelan, V. V.; Simkovsky, R.; Watrous, J. D.; Trial, R. M.; Fleming, T. C.; Wenter, R.; Moore, B. S.; Golden, S. S.; Pogliano, K.; Dorrestein, P. C. Primer on Agar-Based Microbial Imaging Mass Spectrometry. *J. Bacteriol.* **2012**, *194* (22), 6023−6028.

(13) Baquer, G.; Sementé, L.; García-Altares, M.; Lee, Y. J.; Chaurand, P.; Correig, X.; Ràfols, P. RMSIcleanup: An Open-Source Tool for Matrix-Related Peak Annotation in Mass Spectrometry Imaging and Its Application to Silver-Assisted Laser Desorption/Ionization. *J. Cheminform* **2020**, *12* (1), 1−13.

(14) Lagarrigue, M.; Lavigne, R.; Guével, B.; Palmer, A.; Rondel, K.; Guillot, L.; Kobarg, J. H.; Trede, D.; Pineau, C. Spatial Segmentation and Metabolite Annotation Involved in Sperm Maturation in the Rat Epididymis by MALDI Imaging Mass Spectrometry. *Journal of Mass Spectrometry* **2020**, *55* (12), e4633.

(15) Mittal, P.; Price, Z. K.; Lokman, N. A.; Ricciardelli, C.; Oehler, M. K.; Klingler-Hoffmann, M.; Hoffmann, P. Matrix Assisted Laser Desorption/Ionization Mass Spectrometry Imaging (MALDI MSI) for Monitoring of Drug Response in Primary Cancer Spheroids. *Proteomics* **2019**, *19* (21−22), 1−4.

(16) Wang, S.; Bai, H.; Cai, Z.; Gao, D.; Jiang, Y.; Liu, J.; Liu, H. MALDI Imaging for the Localization of Saponins in Root Tissues and Rapid Differentiation of Three Panax Herbs. *Electrophoresis* **2016**, *37* (13), 1956−1966.

(17) Keller, C.; Maeda, J.; Jayaraman, D.; Chakraborty, S.; Sussman, M. R.; Harris, J. M.; Ané, J. M.; Li, L. Comparison of Vacuum Maldi and Ap-Maldi Platforms for the Mass Spectrometry Imaging of Metabolites Involved in Salt Stress in Medicago Truncatula. *Front Plant Sci.* **2018**, *9* (August), 1−15.

(18) Barrios-González, J. Secondary Metabolites Production. *Current Developments in Biotechnology and Bioengineering* **2018**, 257−283.

(19) Fouillaud, M.; Dufossé, L. Microbial Secondary Metabolism and Biotechnology. *Microorganisms* **2022**, *10* (1), 123.

(20) Craney, A.; Ahmed, S.; Nodwell, J. Towards a New Science of Secondary Metabolism. *J. Antibiot.* **2013**, *66* (7), 387−400.

(21) Rigali, S.; Anderssen, S.; Naômé, A.; van Wezel, G. P. Cracking the Regulatory Code of Biosynthetic Gene Clusters as a Strategy for Natural Product Discovery. *Biochem. Pharmacol.* **2018**, *153*, 24.

(22) Bemis, K. D.; Harry, A.; Eberlin, L. S.; Ferreira, C.; Van De Ven, S. M.; Mallick, P.; Stolowitz, M.; Vitek, O. Cardinal: An R Package for Statistical Analysis of Mass Spectrometry-Based Imaging Experiments. *Bioinformatics* **2015**, *31* (14), 2418−2420.

(23) Race, A. M.; Palmer, A. D.; Dexter, A.; Steven, R. T.; Styles, I. B.; Bunch, J. SpectralAnalysis: Software for the Masses. *Anal. Chem.* **2016**, *88* (19), 9451−9458.

(24) Janda, M.; Seah, B. K. B.; Jakob, D.; Beckmann, J.; Geier, B.; Liebeke, M. Determination of Abundant Metabolite Matrix Adducts

Illuminates the Dark Metabolome of MALDI-Mass Spectrometry Imaging Datasets. *Anal. Chem.* **2021**, *93* (24), 8399−8407.
(25) Tiquet, M.; La Rocca, R.; Kirnbauer, S.; Zoratto, S.; Van Kruining, D.; Quinton, L.; Eppe, G.; Martinez-Martinez, P.; Marchetti-Deschmann, M.; De Pauw, E.; Far, J. FT-ICR Mass Spectrometry Imaging at Extreme Mass Resolving Power Using a Dynamically Harmonized ICR Cell with 1ω or 2ω Detection. *Anal. Chem.* **2022**, *94* (26), 9316−9326.
(26) Schramm, T.; Hester, A.; Klinkert, I.; Both, J. P.; Heeren, R. M. A.; Brunelle, A.; Laprévote, O.; Desbenoit, N.; Robbe, M. F.; Stoeckli, M.; Spengler, B.; Römpp, A. ImzML - A Common Data Format for the Flexible Exchange and Processing of Mass Spectrometry Imaging Data. *J. Proteomics* **2012**, *75* (16), 5106−5110.
(27) La Rocca, R.; Kune, C.; Tiquet, M.; Stuart, L.; Eppe, G.; Alexandrov, T.; De Pauw, E.; Quinton, L. Adaptive Pixel Mass Recalibration for Mass Spectrometry Imaging Based on Locally Endogenous Biological Signals. *Anal. Chem.* **2021**, *93* (8), 4066−4074.
(28) Xi, Y.; Sohn, A. L.; Joignant, A. N.; Cologna, S. M.; Prentice, B. M.; Muddiman, D. C. SMART: A Data Reporting Standard for Mass Spectrometry Imaging. *Journal of Mass Spectrometry* **2023**, *58* (2), 1−5.
(29) Kune, C.; McCann, A.; Raphaël, L. R.; Arias, A. A.; Tiquet, M.; Van Kruining, D.; Martinez, P. M.; Ongena, M.; Eppe, G.; Quinton, L.; Far, J.; De Pauw, E. Rapid Visualization of Chemically Related Compounds Using Kendrick Mass Defect As a Filter in Mass Spectrometry Imaging. *Anal. Chem.* **2019**, *91* (20), 13112−13118.
(30) Hustin, J.; Kune, C.; Far, J.; Eppe, G.; Debois, D.; Quinton, L.; De Pauw, E. Differential Kendrick's Plots as an Innovative Tool for Lipidomics in Complex Samples: Comparison of Liquid Chromatography and Infusion-Based Methods to Sample Differential Study. *J. Am. Soc. Mass Spectrom.* **2022**, *33* (12), 2273−2282.
(31) Müller, W. H.; Verdin, A.; Kune, C.; Far, J.; De Pauw, E.; Malherbe, C.; Eppe, G. Dual-Polarity SALDI FT-ICR MS Imaging and Kendrick Mass Defect Data Filtering for Lipid Analysis. *Anal Bioanal Chem.* **2021**, *413* (10), 2821−2830.
(32) McCann, A.; Kune, C.; La Rocca, R.; Oetjen, J.; Arias, A. A.; Ongena, M.; Far, J.; Eppe, G.; Quinton, L.; De Pauw, E. Rapid Visualization of Lipopeptides and Potential Bioactive Groups of Compounds by Combining Ion Mobility and MALDI Imaging Mass Spectrometry. *Drug Discov Today Technol.* **2021**, *39*, 81−88.
(33) Palmer, A.; Phapale, P.; Chernyavsky, I.; Lavigne, R.; Fay, D.; Tarasov, A.; Kovalev, V.; Fuchser, J.; Nikolenko, S.; Pineau, C.; Becker, M.; Alexandrov, T. FDR-Controlled Metabolite Annotation for High-Resolution Imaging Mass Spectrometry. *Nat. Methods* **2017**, *14* (1), 57−60.
(34) Edelsbrunner, H.; Harer, J. Persistent Homology—a Survey. *Contemp. Math.* **2008**, 257−282.
(35) Moumbock, A. F. A.; Gao, M.; Qaseem, A.; Li, J.; Kirchner, P. A.; Ndingkokhar, B.; Bekono, B. D.; Simoben, C. V.; Babiaka, S. B.; Malange, Y. I.; Sauter, F.; Zierep, P.; Ntie-Kang, F.; Günther, S. StreptomeDB 3.0: An Updated Compendium of Streptomycetes Natural Products. *Nucleic Acids Res.* **2021**, *49* (D1), D600−D604.
(36) Guo, A. C.; Jewison, T.; Wilson, M.; Liu, Y.; Knox, C.; Djoumbou, Y.; Lo, P.; Mandal, R.; Krishnamurthy, R.; Wishart, D. S. ECMDB: The E. Coli Metabolome Database. *Nucleic Acids Res.* **2012**, *41* (D1), D625−D630.
(37) Sajed, T.; Marcu, A.; Ramirez, M.; Pon, A.; Guo, A. C.; Knox, C.; Wilson, M.; Grant, J. R.; Djoumbou, Y.; Wishart, D. S. ECMDB 2.0: A Richer Resource for Understanding the Biochemistry of E. *Coli. Nucleic Acids Res.* **2016**, *44* (D1), D495−D501.
(38) Vandenbosch, M.; Mutuku, S. M.; Mantas, M. J. Q.; Patterson, N. H.; Hallmark, T.; Claesen, M.; Heeren, R. M. A.; Hatcher, N. G.; Verbeeck, N.; Ekroos, K.; Ellis, S. R. Toward Omics-Scale Quantitative Mass Spectrometry Imaging of Lipids in Brain Tissue Using a Multiclass Internal Standard Mixture. *Anal. Chem.* **2023**, *95* (51), 18719−18730.
(39) Han, Y.; Qu, X.; Geng, H.; Wang, L.; Zhu, Z.; Zhang, Y.; Cui, X.; Lu, H.; Wang, X.; Chen, P.; Wang, Q.; Sun, C. Isotope-Coded On-Tissue Derivatization for Quantitative Mass Spectrometry Imaging of Short-Chain Fatty Acids in Biological Tissues. *Anal. Chem.* **2023**, *95* (48), 17622−17628.
(40) Holbrook, J. H.; Kemper, G. E.; Hummon, A. B. Quantitative Mass Spectrometry Imaging: Therapeutics & Biomolecules. *Chem. Commun.* **2024**, *60*, 2137−2151.
(41) Nagano, E.; Odake, K.; Shimma, S. An Alternative Method for Quantitative Mass Spectrometry Imaging (q-MSI) of Dopamine Utilizing Fragments of Animal Tissue. *Mass Spectrometry* **2023**, *12* (1), A0128.
(42) Unsihuay, D.; Mesa Sanchez, D.; Laskin, J. Quantitative Mass Spectrometry Imaging of Biological Systems. *Annu. Rev. Phys. Chem.* **2021**, *72*, 307−329.
(43) Lamont, L.; Hadavi, D.; Viehmann, B.; Flinders, B.; Heeren, R. M. A.; Vreeken, R. J.; Porta Siegel, T. Quantitative Mass Spectrometry Imaging of Drugs and Metabolites: A Multiplatform Comparison. *Anal Bioanal Chem.* **2021**, *413* (10), 2779−2791.
(44) Eriksson, J. O.; Sánchez Brotons, A.; Rezeli, M.; Suits, F.; Markó-Varga, G.; Horvatovich, P. MSIWarp: A General Approach to Mass Alignment in Mass Spectrometry Imaging. *Anal. Chem.* **2020**, *92* (24), 16138−16148.
(45) Jourdan, S.; Francis, I. M.; Kim, M. J.; Salazar, J. J. C.; Planckaert, S.; Frère, J. M.; Matagne, A.; Kerff, F.; Devreese, B.; Loria, R.; Rigali, S. The CebE/MsiK Transporter Is a Doorway to the Cello-Oligosaccharide-Mediated Induction of Streptomyces Scabies Pathogenicity. *Sci. Rep* **2016**, *6* (May), 1−12.
(46) Francis, I. M.; Bergin, D.; Deflandre, B.; Gupta, S.; Salazar, J. J. C.; Villagrana, R.; Stulanovic, N.; Ribeiro Monteiro, S.; Kerff, F.; Loria, R.; Rigali, S. Role of Alternative Elicitor Transporters in the Onset of Plant Host Colonization by Streptomyces Scabiei 87−22. *Biology (Basel)* **2023**, *12* (2), 234.
(47) Deflandre, B.; Stulanovic, N.; Planckaert, S.; Anderssen, S.; Bonometti, B.; Karim, L.; Coppieters, W.; Devreese, B.; Rigali, S. The Virulome of Streptomyces Scabiei in Response to Cello-Oligosaccharide Elicitors. *Microb Genom* **2022**, *8* (1), 000760.
(48) Pecoraro, L.; Wang, X.; Shah, D.; Song, X.; Kumar, V.; Shakoor, A.; Tripathi, K.; Ramteke, P. W.; Rani, R. Biosynthesis Pathways, Transport Mechanisms and Biotechnological Applications of Fungal Siderophores. *Journal of Fungi* **2022**, *8* (1), 21.
(49) Van Santen, J. A.; Jacob, G.; Singh, A. L.; Aniebok, V.; Balunas, M. J.; Bunsko, D.; Neto, F. C.; Castan∼o-Espriu, L.; Chang, C.; Clark, T. N.; Cleary Little, J. L.; Delgadillo, D. A.; Dorrestein, P. C.; Duncan, K. R.; Egan, J. M.; Galey, M. M.; Haeckl, F. P. J.; Hua, A.; Hughes, A. H.; Iskakova, D.; Khadilkar, A.; Lee, J. H.; Lee, S.; Legrow, N.; Liu, D. Y.; Macho, J. M.; McCaughey, C. S.; Medema, M. H.; Neupane, R. P.; O'Donnell, T. J.; Paula, J. S.; Sanchez, L. M.; Shaikh, A. F.; Soldatou, S.; Terlouw, B. R.; Tran, T. A.; Valentine, M.; Van Der Hooft, J. J. J.; Vo, D. A.; Wang, M.; Wilson, D.; Zink, K. E.; Linington, R. G. The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Cent Sci.* **2019**, *5* (11), 1824−1833.

https://doi.org/10.1021/jasms.4c00129
J. Am. Soc. Mass Spectrom. 2024, 35, 1743−1755

Associated papers of Chapter 3

A2- Original publication of “Mass shift in mass spectrometry imaging: comprehensive analysis and practical corrective workflow.”

Analytical and Bioanalytical Chemistry
https://doi.org/10.1007/s00216-021-03174-1

RESEARCH PAPER

Mass shift in mass spectrometry imaging: comprehensive analysis and practical corrective workflow

Andréa McCann[1] · Sophie Rappe[1] · Raphaël La Rocca[1] · Mathieu Tiquet[1] · Loïc Quinton[1] · Gauthier Eppe[1] · Johann Far[1] · Edwin De Pauw[1] · Christopher Kune[1]

Received: 14 October 2020 / Revised: 22 December 2020 / Accepted: 13 January 2021
© Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

MALDI mass spectrometry imaging (MSI) allows the mapping and the tentative identification of compounds based on their *m/z* value. In typical MSI, a spectrum is taken at incremental 2D coordinates (pixels) across a sample surface. Single pixel mass spectra show the resolving power of the mass analyzer. Mass shift, i.e., variations of the *m/z* of the same ion(s), may occur from one pixel to another. The superposition of shifted masses from individual pixels peaks apparently degrades the resolution and the mass accuracy in the average spectrum. This leads to low confidence annotations and biased localization in the image. Besides the intrinsic performances of the analyzer, the sample properties (local composition, thickness, matrix deposition) and the calibration method are sources of mass shift. Here, we report a critical analysis and recommendations to mitigate these sources of mass shift. Mass shift 2D distributions were mapped to illustrate its effect and explore systematically its origin. Adapting the sample preparation, carefully selecting the data acquisition settings, and wisely applying post-processing methods (i.e., *m/z* realignment or individual *m/z* recalibration pixel by pixel) are key factors to lower the mass shift and to improve image quality and annotations. A recommended workflow, resulting from a comprehensive analysis, was successfully applied to several complex samples acquired on both MALDI ToF and MALDI FT-ICR instruments.

Keywords Mass spectrometry imaging · FT-ICR · Mass shift · Data recalibration · Time of flight · Mass accuracy

Introduction

Mass spectrometry imaging (MSI) is a powerful mass spectrometry (MS) method to visualize 2D molecular species distribution in a sample. A mass spectrometry image consists of a set of mass spectra obtained from a thin section of the sample moved step-by-step in a regular grid under an ionization source. Each spot, also called pixel, contains a whole mass spectrum. A common method to produce ions at different spots is the use of laser-based ion source, such as laser desorption ionization (LDI), surface-assisted laser desorption ionization (SALDI), and especially matrix-assisted laser desorption ionization (MALDI). Desorption electrospray ionization (DESI) and secondary ion mass spectrometry are also popular ion sources fitting well with mass spectrometry imaging [1]. Less frequently, images are obtained in microscopic mode using a position sensitive detector [2]. In scanning MSI methods, all the individual mass spectra (i.e., from all pixels) are compiled to produce 2D maps that can be reconstructed for each *m/z* [3] or for families of compounds [4].

The use of maximum mass accuracy and sufficient mass resolving power allows a molecular formula to be assigned to each peak, as long as the peak is free of isobaric interferences and detected with a sufficient sensitivity [5], while MS/MS strategies are usually required to resolve atomic connectivity and to identify the ions with more confidence [6–8]. Any additional descriptors, such as the collision cross section (i.e., from ion mobility MS) can help in confirming the identification without requiring MS/MS steps [9–11].

Peak annotation in mass spectrometry imaging will therefore rely primarily on the *m/z* values and isotopic ratios. MSI gains its full power when accurate mass determination at high

Published in the topical collection *Mass Spectrometry Imaging 2.0* with guest editors Shane R. Ellis and Tiffany Porta Siegel.

Andréa McCann and Sophie Rappe are co-first authors of this paper.

Andréa McCann and Sophie Rappe contributed equally to this work.

✉ Christopher Kune
c.kune@uliege.be

[1] Mass Spectrometry Laboratory, MolSys Research Unit, Department of Chemistry, University of Liège, Allée du Six Août, 11 – Quartier Agora, 4000 Liège, Belgium

Published online: 30 January 2021

resolving power can be obtained on the whole image. This allows the optimum use of a common step in MSI; the construction of an average spectrum from the individual pixel spectra. The average spectrum facilitates the obtention of a quick global view of *m/z* distributions and the evaluation of the data quality. Peak annotation can then be performed from the average spectrum and reliable *m/z* images can be reconstructed. The individual spectrum quality of a pixel corresponds to the performances of the MS analyzer, such as time of flight (ToF), Fourier transform ion cyclotron resonance (FT-ICR), and Orbitrap mass spectrometers. However, the exact mass may be shifted from one pixel to another leading to peak broadening in the average spectrum. Both instrument and sample characteristics can affect the mass accuracy from one pixel to another. Mass analyzers are subjected to physical limitations linked to their ion optics, which need to be properly considered for successful mass spectrometry imaging experiment [12–19]. In ToF instruments, a variation in sample height may affect the ion trajectory and consequently the mass accuracy. In FT-ICR instruments, a high ion density may cause space charge effects in the ICR cell, resulting in mass shifts. In the presence of mass shift, higher tolerance levels in terms of mass accuracy are required to reconstruct the relevant image of a targeted ion, increasing the risk of isobaric interference and miss-annotated peaks.

Here, we present a systematic study aimed at understanding, visualizing, reducing, and correcting mass shift to alleviate an important limitation in MSI. The results are presented to provide users general recommendations to improve mass accuracy of the MSI data by preventing or, at least, by reducing the experimental mass shift between pixels. To achieve this purpose, we have pinpointed the main key factors causing mass shift according to the type of MALDI MS instruments: sample thickness for the MALDI ToF (linear geometry) and sample heterogeneity for MALDI FT-ICR fitted with the dynamically harmonized ICR cell. Nonetheless, other MALDI MS instruments used for MALDI mass spectrometry imaging will also benefit from these recommendations. Our recommendations cover sample preparation to the optimization of MS parameters for data acquisition. We also propose data post-processing methods to perform the *m/z* realignment and *m/z* recalibration of the spectra pixel by pixel. Finally, we have applied our recommendations for various complex and problematic samples where we have favorably improved the mass accuracy in the MSI average spectrum leading to improved peak annotations.

Materials and methods

Material

MALDI matrix α-cyano-4-hydroxycinnamic acid (HCCA) MALDI matrix 2,5-Dihydroxybenzoic acid (DHB), trifluoroacteic acid (TFA), formic acid, ethyl acetate, chloroform, and absolute ethanol were purchased from Sigma-Aldrich (Overijse, Belgium). Acetonitrile and methanol were LC-MS grade from Biosolve (Valkenswaard, Netherlands). Poly(ethylene oxide) monomethylether (CH_3O-$(CH_2O)_n$-OH) having an average nominal mass of 750 g/mol and sodium chloride were purchased from Sigma-Aldrich (Overijse, Belgium). Select Agar™ under powder form was obtained from Invitrogen (Thermo Fisher Scientific, Waltham, MA, USA). Standards Ceramide Sphingoid mix II (25 μM) were purchased from Avanti polar lipid (from local exclusive redistributor Sigma-Aldrich). Indium tin oxide (ITO)-coated glass slides were purchased from Bruker Daltonics (Bremen, Germany). n-Hexane was purchased from LGC (Queens Roads, Teddington, Middlesex, UK).

Sample preparation for MALDI MSI

PEO synthetic samples PEO stock solution was prepared by dissolving PEO in ACN 10 μM NaCl to reach a concentration of 100 μM. A synthetic sample imitating heterogeneous sample in terms of molecular composition using different concentrations of PEO was built. The PEO solution was diluted twice 10 times in series in ACN to obtain three PEO solutions of 100 μM, 10 μM, and 1 μM (assuming non-dispersed distribution of PEO having a molecular weight of 750 g/mol). Each PEO solution was then sprayed onto different regions of an ITO glass slide with the SunCollect instrument (SunChrom, Friedrichsdorf, Germany). For each region, 10 layers of PEO were sprayed at a linear gradient of flow rate from 10 to 60 μL/min. A purposely designed synthetic sample for the study of topology variation effect was prepared using 1.5% agar powder diluted in H_20 spiked with 5 mM NaCl and PEO stock solution to reach a final PEO concentration of 10 μM. The liquid agar solution was then poured in 10-cm plastic petri dishes with various volumes of liquid agar (10 mL, 15 mL, 20 mL, 25 mL, depicted as T1, T2, T3, and T4 respectively) to obtain different agar thicknesses. The agar was then left to cool until solidified. Small rectangles of agar were cut and fixed onto 75×25-mm^2 ITO-coated conductive glass slides previously covered with double-sided copper tape (StructureProbe INC, West Chester, PA, USA). The ITO slides were placed under vacuum at 70 mbar until the agar dried.

Cryosectioning Mouse brain tissue (kindly provided by the Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience from Maastricht University) and salmon brain (from a local retailer) homogenate slices destined to mass spectrometry imaging were prepared using a cryomicrotome (Cryostar NX70, Thermo Fischer Scientific, Waltham, MA, USA) with temperatures set at − 20 °C for the sample holder and − 10 °C for the SEC 35e razor blade. Mouse brain tissue slices were cut at a thickness of 10 μm while salmon brain homogenates were cut at 16 μm.

Additionally, some salmon brain homogenate slices were cut after inducing bevel-like slices in the sample holder horizontal axis where the thickest part was estimated to be 50 ± 10 μm. All slices were thawed-mounted on ITO glass slides. The samples were then dried in a vacuum desiccator for 15 min or until visible dryness.

Bacterial strains, medium, and culture conditions The strains used for this study were *Bacillus velezensis GA1* and *Pseudomonas sp. CMR12a*. Five micoliters of both bacterial cell suspension (OD 600 nm = 0.1) was spotted at 10-mm distance on a semisolid agar-based root exudates mimicking medium [20] and incubated for 48 h at 25 °C. Microbial colonies on agar and region of interest were cut from the Petri dish and transferred to the target ITO plate, previously covered with double-sided conductive copper tape. This assembly was then placed in a vacuum desiccator at 70 mbar until dryness overnight [21].

Matrix deposition HCCA matrix was sprayed on PEO samples, bacteria sample, and brain samples (mouse brain tissue and salmon brain homogenates). HCCA matrix was mixed in 70:30 ACN:water v:v with 0.2% TFA to reach a final concentration of 5 mg/mL. In total, 10 layers of HCCA matrix were sprayed onto the ITO slides with the SunCollect instrument (SunChrom, Friedrichsdorf, Germany). The first layer was sprayed at a flow rate of 10 μL/min. Flow rate was increased by 10 μL/min after each layer until reaching 60 μL/min.

Standard lipid samples with different crystallization methods The heterogeneous crystallization sample was obtained by diluting the standard Ceramide Sphingoid mix II (25 μM) tenfold in an ethanol-chloroform mixture (50:50 v:v). The DHB matrix was diluted at 100 mg/mL in an ethyl acetate-hexane-formic acid mixture (98.9:1:0.1 v:v:v). One microliter of diluted lipids standard (Ceramide Sphingoid mix II) was spotted onto an ITO slide. One microliter of matrix solution was then spotted onto the sample and mixed thoroughly with the dried droplet crystallization method. Samples were air-dried at room temperature. The sizes of crystals were determined using an inverted optical microscope Olympus IX 81 (Olympus, Aartselaar, Belgium). Sample topology was investigated by scanning electron microscopy using the XL30 FEG-ESEM instrument (Field Electron and Ion Company, Hillsboro, OR, USA). Two homogeneous crystallization samples were prepared alongside by spotting 1 μL of diluted lipids standard on each ITO slide. A 10 mg/mL solution of matrix DHB diluted in an ethyl acetate-hexane-formic acid mixture (98.9:1:0.1 v:v:v) was then sprayed onto the first ITO slide with the HTX TM-Sprayer (HTX Technologies, LLC, Chapel Hill, NC, USA). In total, 25 layers of matrix solution were applied at a flow rate of 5 μL/min. Fifty milligrams of DHB matrix dissolved in 2 mL of methanol was also applied by sublimation using the HTX Sublimator (HTX technologies, Chapel Hill, NC, USA) to the second ITO slide. The solvent was evaporated during preheating of the chamber at 65 °C and the sublimation of DHB occurred at 160 °C under vacuum for 160 s.

MALDI MSI methods

Bacteria co-culture, synthetic PEO, and agar samples Mass spectrometry images were produced in positive ionization mode from *m/z* 300 to 2000 using either FT-ICR instrument (solariX XR 9.4T, Bruker Daltonics, Bremen, Germany) fitted the dual ESI/MALDI ion sources and SmartBeam laser or ToF instrument (rapifleX, Bruker Daltonics, Bremen, Germany) fitted with the smartbeam™ 3D laser. Laser parameters (laser power, shots number, and repetition rate) were optimized on both instruments to ensure the correct detection of the minimum concentration of PEG (1 μM) at *S/N* above 3 according to the FlexImaging 5.0 software (Bruker). Both instruments were calibrated from *m/z* 300 to 2000 in the same conditions as the mass spectrometry images using red phosphorous freshly suspended in acetone and directly spotted onto each ITO glass slide. Pixel step size for the surface raster was set to 200 μm.

Bacteria imaging was performed with a raster of 100 μm using 20 laser shots set and repetition rate of 5 kHz for both the solariX XR and rapifleX instruments. Laser power was set at 40% on the rapifleX instrument and at 70% on the solariX XR instrument. Two other images were acquired with either 200 or 2000 laser shots to estimate the effect of mass shift dependence in regard to the amount of the injected ions per packet.

Mouse brain tissue and salmon brain homogenate samples The mouse brain tissue images were acquired on the solariX XR FT-ICR MS using a lateral laser resolution of 60 μm and laser parameters of 400 shots at 1000 Hz with a power of 70%. Salmon brain homogenate images were acquired on the rapifleX ToF MS and on the solariX XR FT-ICR MS instruments using a raster width of 100 μm. FT-ICR laser parameters were set as 20 shots at 200 Hz with a laser power of 70%. The rapifleX laser parameters were set to 500 laser shots at 5 kHz.

Lipid standard sample and co-crystallization experiments Data acquisitions of lipid samples were performed using the rapifleX in positive mode ionization at *m/z* 100 to 1200. Laser power was set to 60%. The MSI data were acquired with a spatial resolution of 25 μm for the dried droplet sample and 10 μm for sprayed matrix and matrix sublimation deposition, using 1000 laser shots per pixel at 10 kHz frequency.

Data processing

Data were reprocessed using FlexImaging 5.0 (Bruker Daltonics, Bremen, Germany) and SCiLS (Bruker Daltonics,

Bremen, Germany). MSI data recalibration was carried out using our in-house algorithm [22] (see also Supplementary Information (ESM)). The image of the mass shift dispersion and computation of the maximum mass shift of the dataset were performed using in-house Python software (using Numpy, Scipy, Tkinter, and Matplotlib numerical packages) written by C. Kune [4] and modified to fit this purpose.

Result and discussions

Mass shift visualization

The difference between the experimental accurate *m/z* values of a target ion and its theoretical value (computed from its molecular formula) is called the mass accuracy. Variations of the mass accuracy may occur from one pixel to another due to the limitations of the instrument, which will be diagnosed in this study on experimental conditions. Those mass accuracy variations affect the average mass spectrum, extending the width of *m/z* distribution and thus apparently decreasing the mass resolving power below the instrument specifications. In addition, the apex of the peaks in the average spectrum may differ from those of the individual spectra, due to the use of a unique calibration covering the whole MSI data. Consequently, the peak apex in the average spectrum may deviate significantly from the expected *m/z*, resulting in poor accuracy compared to manufacturing specifications. The peak apex of the *m/z* distribution from the average mass spectrum is often used to estimate the global mass accuracy of a MSI data without considering mass shift effect. Both mass accuracy and mass shift in the average spectrum should be considered as two criteria for evaluating the quality of a MSI data. Since the average mass spectrum is very frequently used as the starting point for extracting reconstructed *m/z* images from MSI data, it is crucial to first understand the contribution of the experimental settings for the optimization of these two criteria before their correction involving data post-processing.

Two worst-case scenarios illustrating mass shift issues on MSI data analysis are reported in Fig. 1. The first example (Fig. 1a) concerns the signal monitoring of the Na^+ adduct of Lactosyl C_{12} ceramide ([Lactosyl C_{12} + Na]$^+$, *m/z* 828.5444), present in the Ceramide Sphingoid mix II standard lipid solution co-crystallized in DHB matrix. The mass spectrometry images were acquired using a rapifleX MALDI ToF instrument. The crystallization of the sample was deliberately intended to produce different crystal shapes. As expected, one observes the formation of matrix crystals exhibiting different sizes from a few micrometers to 200–300 μm wide and approximately 40 μm of height variations, determined with an optical inverted microscope. The mass accuracy in the average mass spectrum (around 100 ppm) and the final resolving power were very poor (2500 FWHM at *m/z* 828.5444)). This mass accuracy is far below the expected

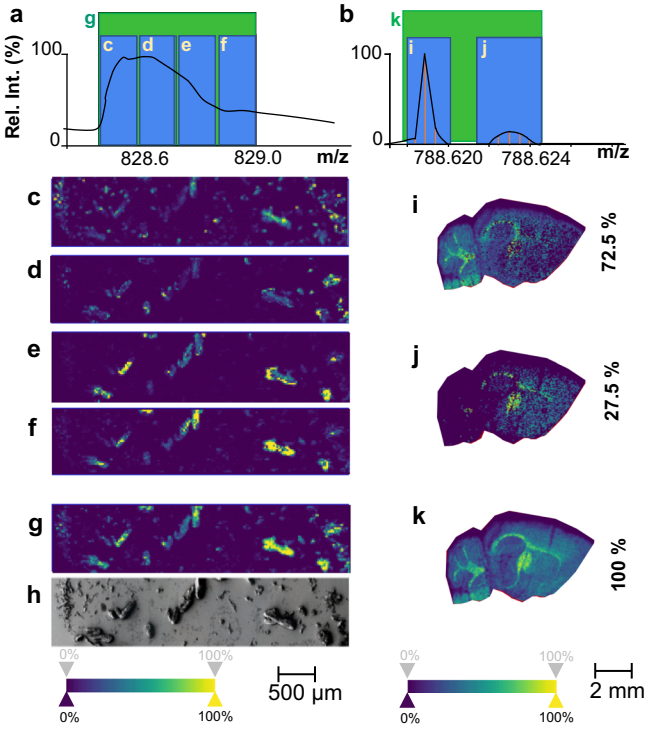


Fig. 1 Ion images reconstructed from *m/z* selection in the average MS spectrum for MSI data subjected to mass shift. **a** Distribution of Lactosyl C12 ceramide with sodium adduct (*m/z* 828.5444) in a dried droplet of standard lipid solution with DHB matrix. An electron microscopy image is displayed for the dried droplet of standard lipid solution (h) to show the high heterogeneity of the sample crystallization. **b** Distribution of protonated PC 36:1 (*m/z* 788.6164) in mouse brain tissue. For both samples, the partial images reconstructed from *m/z* selections (c, d, e, f for sample a and i, j for sample b) and the global images (g and k) are reported showing the importance of the *m/z* tolerance for relevant MSI image reconstruction

manufacturer's specification of the rapifleX's ToF analyzer, which under optimal condition and in reflectron mode, should provide a mass accuracy less than or equal to 5 ppm and a resolving power greater than or equal to 13,000 FWHM for Bradykinin 2-9 at *m/z* 904. Different consecutive *m/z* ranges (0.1400 u) were taken from the broad peaks in the average spectrum and corresponding images were reconstructed. The combination of all the *m/z* areas within a single broad peak allows reconstructing of the full image. In this example, [Lactosyl C_{12} + Na]$^+$ is detected within a relatively large range of *m/z* values (about 0.5 u) depending on the sample localization. The whole MS distribution of [Lactosyl C_{12} + Na]$^+$ (Fig. 1g) must be taken into account to capture the entire map of this compound. It is interesting to note that the MS images reconstructed by sectioning different *m/z* values (Fig. 1c, d, e and f) highlight different crystals based on their size, suggesting a correlation between crystals size and mass shifts (this will be discussed in the section dedicated to sample thicknesses). Mass shifts could also be observed, in a lesser extent (0.005 u), using the solariX XR FT-ICR instrument, as reported in Fig. 1b (mouse brain tissue section MSI). The investigated ion in this study has been putatively identified as the protonated phosphatidylcholine PC 36:1

(m/z = 788.6164) after a bulk structure search on the LIPID MAPS Structure Database (LMSD) [23]. The effect of mass shift reported in this figure is even more pronounced than the previous one since two apparently distinct m/z peaks were observed for a single ion. The reconstructed ion images from each MS peak (selections i and j in Fig. 1b) lead to partial but complementary images with a pixel recovery score of 72.5% and 27.5% respectively. Here again, the whole signal of protonated PC 36:1 should be taken into account to reconstruct the entire ion image of this tissue section (k in Fig. 1b).

These two examples show that the mass shift impairs the mass resolving power in the average spectrum, and results in a potentially compromised ion annotation. On the FT-ICR example, individual spectra comply with the expected resolution, but the mass measurement accuracy may differ. The degraded mass resolving power in the average mass spectrum suggests therefore that this occurs upon summation of the individual spectra. On the ToF example, some individual spectra showed broader peaks than expected by the instrument resolution, suggesting additional factors influencing the average MS spectrum resolution.

Before attempting any correction, mass shift visualization by mapping the variation of the experimental m/z peaks for all the pixels is a fast and easy way to figure how severe could be this mass shift in your mass spectrometry imaging data.

Mass shift amplitude can be estimated from the variation of the measured m/z value of the same ion in each pixel. For this purpose, the measured m/z value in each pixel should be characterized as the m/z peak apex or directly obtained from centroid MS spectra (i.e., bar charts obtained after peak-picking process). This visualization of the mass shift can be computed for any MS analyzer. Because mass shift estimation relies on centroid m/z distribution only, the instrument resolving power does not affect the calculation. Mass shift value can be plotted in a 2D heatmap (i.e., mass shift value in function of the XY position of the pixel) allowing the visualization of the mass shift in function of the localization in the sample. The mass shift heatmaps presented in this work have been generated with an in-house software. Briefly, this software performed a peak picking in all pixels of the MSI data. The mass shift is calculated according to Eq. 1 in a very similar fashion as the mass accuracy calculation between the centroid peak of the computed pixel and the average mass spectrum. The main difference with the conventional mass accuracy equation relies on the chosen mass reference value. This can be the theoretical m/z value of a target ion or an arbitrary m/z value of this peak, e.g., the smallest m/z value reported for this peak, as long as the chosen mass reference is explicitly and initially defined by the users. The output of this script is a list of mass shifts (i.e., mass shift according to the pixel identifier) and a mass shift heatmap (i.e., mass shift plotted according to the XY position in the sample). This script also allows the extraction of the peak width (e.g., at 10% of its absolute intensity of full width at half maximum) to monitor the MS peak width evolution from pixel to pixel in addition to mass shifts.

$$\text{masserror}_{\text{rel.}}(\text{ppm}) = \frac{m/z_{\text{exp.}} - m/z_{\text{exact}}}{m/z_{\text{exact}}} \times 10^6 \quad (1)$$

Mass shift understanding

Sample composition variation

Unlike ToF, ICR instruments are ion trap mass analyzers. FT-ICR instruments infer the m/z ratio of ions by measuring the angular frequency of the ions under high magnetic field due to the Lorentz force. The angular frequencies of the ions depend solely on the m/z and the magnetic field strength. Ion detection is carried out by the mirror current of the spinning ions in proximity with the detection plates of the ICR cell after excitation by the excitation plates. However, the trapped ions inside the ICR cell may be subject to space charge effects, affecting the angular frequencies of the ions.

Space charge effects on the mass shift during MSI experiments by MALDI FT-ICR were studied using various concentrations of PEO sprayed onto an ITO glass slide and covered with HCCA matrix. Ion count and mass shift evolution of the most intense PEO ion $[PEO17+Na]^+$ (m/z = 803.4616) were monitored according to the different zones of the sample slide with both MALDI ToF and MALDI FT-ICR, and are summarized in Table 1. The mass shift (in ppm) was calculated according to Eq. 1, using the peak apex in the average mass spectrum of 1 μM $[PEO17+Na]^+$ as m/z reference.

As expected, Fig. S1 (see ESM) and Table 1 illustrate that the increase of PEO concentration leads to an increase of the total ion current (TIC), especially for the MALDI FT-ICR instruments (i.e., solariX XR instrument). MSI experiment performed with the solariX XR 9.4T (see Table 1 and ESM Fig. S1) shows that the m/z values are shifted to higher values when the TIC increase. A linear correlation between the TIC and the shift of the measured mass was observed in perfect agreement with the report from E. Nikolaev and coworkers [14]. A significant mass shift of 2 ± 0.1 ppm (specification of the solariX XR is 0.5 ppm RMS) is observed between the m/z signal in the zone with the highest PEO concentration (PEO concentration = 100 μM, TIC = 1.79×10^8 cps) and the reference zone (PEO concentration = 1 μM, TIC = 8.35×10^7 cps). Moderate mass shift is also observed at the intermediate concentration level (PEO concentration = 10 μM, mass shift = 0.5 ± 0.1 ppm, TIC = 9.98×10^7 cps). Contrary to the solariX XR experiment, no correlation between mass shift and TIC was observed in MALDI ToF MSI data. This confirms that the m/z measurement is almost unaffected by the space charge effect in the ToF analyzer. This model confirms that the space charge effect is an inherent issue for high-resolution instrument trapping ions such

Table 1 Effect of concentration of polyethylene oxide methyl ether on the generated total ion current (TIC) on the experimental mass shift. Mass reference was (PEO) $[PEO17+Na]^+$ (m/z = 803.4616). MALDI ToF was a rapifleX and the MALDI FT-ICR was a solariX XR 9.4T. See related Fig. S1 in ESM

Mass analyzers		Polyethylene oxide methyl ether concentration					
		TIC (cps)			Mass shift (ppm)		
		1 μM	10 μM	100 μM	1 μM	10 μM	100 μM
ToF	Average	116	115	171	0.00	1.51	1.68
	Standard deviation	±3	±3	±4	±1.13	±0.64	±0.67
FT-ICR	Average	8.35×10^7	9.98×10^7	1.79×10^8	0.00	0.37	2.03
	Standard deviation	$\pm5.45\times10^5$	$\pm6.78\times10^5$	$\pm1.10\times10^6$	±0.07	±0.05	±0.06

as FT-ICR devices. Nevertheless, the correlation between the mass shift and the TIC is instrument-dependent and must be determined for each type of device (e.g., TIC effect on mass shift could be different on various FT-ICR instruments especially when using a different supra-conducting magnet).

Sample thickness

ToF mass analyzers are scanning instruments, meaning that the ions are transmitted through the mass analyzer to determine their m/z ratios. In these instruments, the mass is calculated from the ion velocity (specifically, the time to travel a given distance) when they are accelerated to a common kinetic energy in a vacuum. Any deviation from this acceleration will modify the measured m/z value. Delayed extraction associated with the use of a reflectron makes ToF instruments fast and efficient mass analyzers. However, the kinetic energy provided by the ions is critically affected by the position in the source of the surface of the sample to be ionized [24]. Variations in the sample's surface topology will produce packets of ions with different kinetic energies. To analyze the contribution of the surface topology to the mass shift, we used polyethylene oxide methyl ether (PEO) polymer embedded in various thicknesses of agar medium that we dried and recovered with HCCA matrix before mass spectrometry imaging.

Rough surface samples are commonly known to cause a loss in mass accuracy [16, 17, 25] and a drop in ionization efficiency [26, 27]. Mass shifts for the $[PEO17+Na]^+$ (m/z = 803.4616) were calculated (in ppm) from Eq. 1 (using the average m/z value of $[PEO17+Na]^+$ in the thinnest zone of the sample, depicted as "T1" as the reference value). The calculated mass shifts are then presented in Table 2 (see ESM Fig. S1), for both MALDI ToF and MALDI FT-ICR instruments. In a ToF instrument, the topology effect is particularly visible, with a relative mass shift reaching up to − 350 ppm. The observed trends showed that the higher the sample surface, the smaller the measured m/z value, as already mentioned elsewhere [24].

In the MALDI FT-ICR instrument, a relatively limited mass shift is observed (± 0.5 ppm, i.e., under the instrument specification), independently of the sample surface. In this experiment, the small range of intensity variation (i.e., from 6.81×10^7 to 7.42×10^7 cps) seems to be low enough to avoid mass shift due to significant space charge effect (note that these values are only applicable for the FT-ICR instrument used, i.e., Bruker's solariX XR). From these results, one can assume that sample topology has a negligible impact on the mass shift when using FT-ICR instruments.

It is however important to highlight that mass shift is not the only consequence of sample topology. Independently of the instrument used, the thickness of the sample also affects

Table 2 Effect of the surface topology of imaged agar samples using mass spectrometry on the experimental mass shift (reference mass = theoretical mass of polyethylene oxide methyl ether (PEO) $[PEO17+Na]^+$ (m/z = 803.4616). MALDI ToF was a rapifleX and the MALDI FT-ICR was a solariX XR 9.4T. Thicknesses of the PEO embedded in agar were ranged as T4 > T3 > T2 > T1. See ESM Fig. S1

Mass analyzers		Sample height of agar sample containing polyethylene oxide methyl ether (T4 > T3 > T2 > T1)							
		TIC (cps)				Mass shift (ppm)			
		T1	T2	T3	T4	T1	T2	T3	T4
ToF	Average	792	695	441	332	0.00	−58.02	−256.20	−374.80
	Standard deviation	± 10	± 9	± 6	± 5	± 1.29	±1.4	± 1.9	± 1.8
FT-ICR	Average	7.40×10^7	7.32×10^7	6.81×10^7	7.42×10^7	0.00	−0.06	0.41	0.5
	Standard deviation	$\pm3.32\times10^5$	$\pm2.45\times10^5$	$\pm3.62\times10^5$	$\pm3.24\times10^5$	± 0.05	± 0.05	± 0.02	± 0.01

the signal intensity. Indeed, the sample surface is no longer in the focal plane of the laser, a reduction of the photon density occurs resulting in a decrease of the apparent ionization efficiency. This loss of sensibility could eventually be mitigated by increasing the laser energy above the threshold energy required for the sample ionization [28]. An uneven surface can also cause a loss in lateral resolution, as the laser ablation spot size may be modified. Up to now, there is no commercially available MALDI source with a pixel to pixel autofocus, but the Spengler group has shown in 2017 that images on irregular surfaces can be achieved without height-related signal artifacts by adjusting sample stage's *z*-position based on a topographic image of the sample [29].

In Fig. 2, we are presenting the heatmaps of the mass shift from a lipid ion at *m/z* 828.5514 (putatively identified as [PC 38:6+Na]$^+$ according to Lipid MAPS LSMD) present in a non-flat sample (beveled slice of salmon brain homogenate, depicted "beveled" sample as shown in Fig. 2a) and a flat sample (flat slice of salmon brain homogenate, depicted "flat" sample in Fig. 2b) acquired on the rapifleX (MALDI ToF) and the solariX XR (MALDI FT-ICR). The *m/z* distribution of this ion is free of isobaric interferences in both MS instrument. The mass shifts have been calculated according to Eq. 1 using the exact mass of this lipid (*m/z* = 828.5514). In the MALDI ToF data, the "flat" sample shows a rather stable mass shift (Fig. 2f); however, the mass shift calculated from the "beveled" sample (Fig. 2d) is shown to be greater in the pixels located at the lower height of the "beveled" sample (mass shift was estimated at +180 ppm). In the FT-ICR MS data, the mass shift remains below 2 ppm and shows no correlation with sample height. These results demonstrate that flat samples are required for accurate MSI data acquisition involving

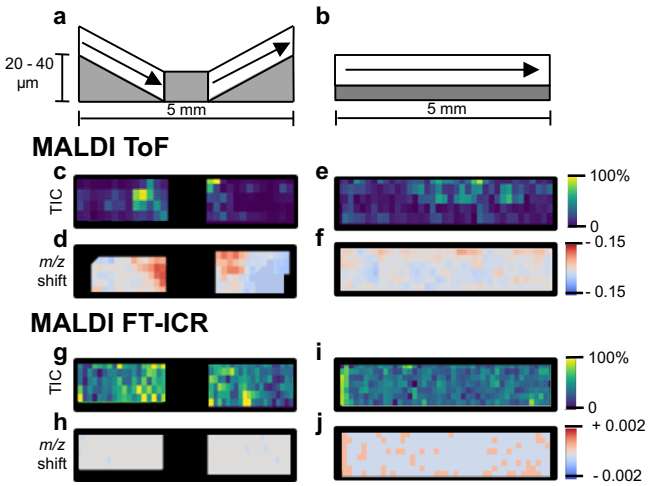


Fig. 2 Beveled (a) and flat (b) samples topographies representation. TIC heatmap for ToF ((c) for "beveled" and (e) for "flat" samples) and FT-ICR ((g) for "beveled" and (i) for "flat" samples) acquisitions. Mass shift heatmaps for *m/z* 828.5514 in ToF ((d) for "beveled" and (f) for "flat" samples) and FT-ICR ((h) for "beveled" and (j) for "flat" samples) MSI data

MALDI ToF, especially when the MALDI source is coupled with a linear ToF mass analyzer. Alternatively, the surface topology contribution is less critical for FT-ICR instrument, which ensures a limited variation of the TIC variation from one pixel to another.

In some cases, sample flatness is not enough to prevent mass shift occurring. Indeed, the mode of matrix deposition affects the homogeneity of the matrix crystals [30, 31] and may induce mass shift as well. As previously shown in Fig. 1a, the heterogeneity of the matrix crystals can induce a significant mass shift for the signal of [Lactosyl C12 ceramide + Na]$^+$ acquired on a MALDI ToF. This mass shift could result from topology variation caused by a heterogeneous distribution of crystals of different sizes and thicknesses. Considering individual spectra of selected pixels, the importance of the mass shift mainly depends on the crystal morphology on the sample (Fig. 3a). In order to investigate the effect of the matrix crystal size on the mass shift, we considered additional matrix deposition methods, such as spraying the matrix on the sample using the HTX sprayer (Fig. 3b) or matrix sublimation using the HTX Sublimator (Fig. 3c). Images obtained by electron and inverted optical microscopy of these samples (see ESM Fig. S2) show improved crystal formation and homogeneity of the matrix deposition under these two matrix deposition techniques. MS peaks extracted from several randomly selected pixels show the reduction of mass shift. Compared to the dried droplet method, matrix sublimation and, to a lesser extent, HTX matrix spraying resulted in better mass resolving power in the average spectrum by reducing the mass shift from one pixel to another.

Nonetheless, the calculated mass accuracy from the peak apex of [Lactosyl C12 ceramide + Na]$^+$ (*m/z* = 828.5444) in the average spectrum is still relatively low (mass errors are estimated at 100 ppm, − 170 ppm, and − 300 ppm in Fig. 3a, b, and c respectively). Since each image was independently and externally calibrated from red phosphorus, these mass errors suggest differences in topology (*Z*-axis) between the calibrating substance and the sample surface.

Interestingly, the crystal heterogeneity could also influence the MS peak width within a single pixel (Fig. 3a, pixels depicted 6, 7, 8, 9, and 10). As the loss in mass resolving power has already occurred at the pixel level, the mass resolving power in the average MS spectrum is inherently affected. In ToF instruments, several MS spectra are cumulated to generate the MS spectrum recorded for the pixel position. In other words, MS spectra recorded in each pixel is already a combination of several MS scans. As a consequence, the MS peak widening due to mass shift can also be observed in a pixel spectrum, as it is the case here. This observation can be explained by (i) the micro-heterogeneity sampled by the laser beam and/or (ii) the crystal morphology changes due to the successive laser ablations [32–34], and tend to be less

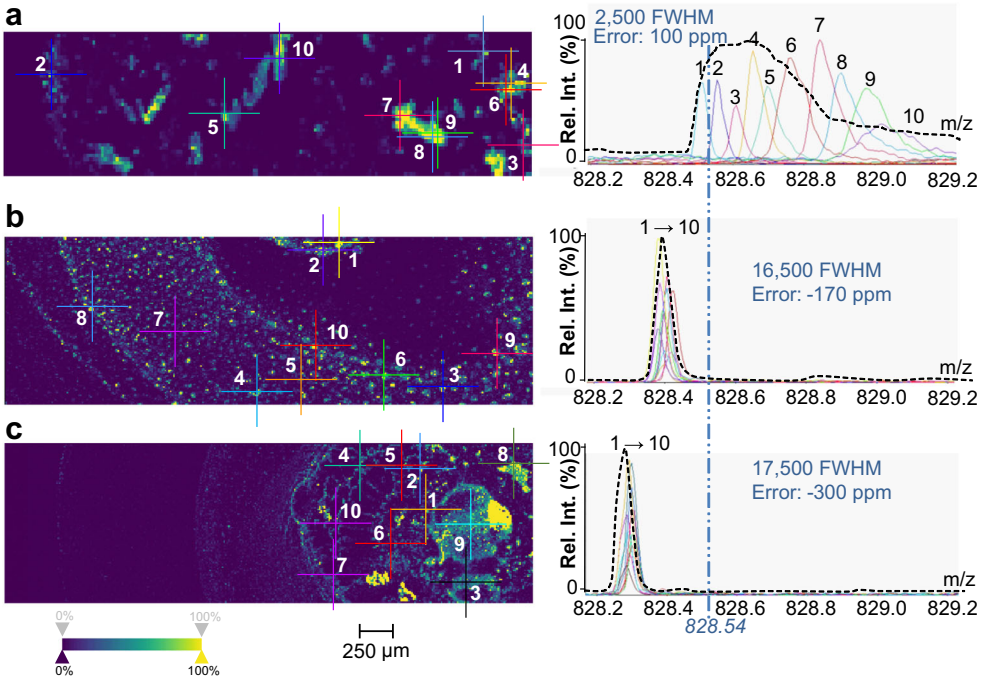


Fig. 3 [Lactosyl C12 ceramide + Na]$^+$ (theoretical m/z = 828.5444) mass spectrometry images acquired with the rapifleX MALDI ToF instrument using different DHB matrix deposition processes. **a** Dried droplet method, **b** HTX Sprayer, **c** HTX Sublimator. Color-coded crosses on the heatmaps correspond to their different individual spectrum at the right of the ion intensities heatmaps. Black dashed line represents the [Lactosyl C12 ceramide + Na]$^+$ distribution in MSI average spectrum. The theoretical m/z value of [Lactosyl C12 ceramide + Na]$^+$ is represented by a blue vertical dashed line. Each image acquisition was externally and independently calibrated using red phosphorus spots

significant in samples with homogeneous crystallization (Fig. 3b and c). These observations suggest that MS peak broadening can occur in parallel to MS peak shifting in MALDI ToF instrument during the MSI acquisition due to the laser ablation of the matrix. MS peak widening in a single pixel was not observed in FT-ICR instrument since only one MS acquisition is performed for each pixel. A peak width monitoring on each pixel could be performed to highlight crystal morphology variations and/or micro-heterogeneities at the pixel scale.

Mass shift prevention

Considerations for performing a robust MS calibration for MSI analyses

Mass shift is related, inter alia, to the sample composition (affecting the space charge effect) and to the sample surface topology. Since these sample properties directly influence the physical quantities measured by the MS instrument, it is recommended to calibrate the instrument with calibrating substances in similar experimental conditions to the sample wherever possible [35, 36]. The MS calibration of heterogeneous MSI samples is thus quite challenging since the sample composition and local heterogeneity is not known before performing the mass spectrometry images of the current sample. Pre-acquisition of a similar sample, such as adjacent or contiguous tissue section, would appear to be promising but there is no guarantee that the mass spectrometry image of this sample will provide a correct external calibration of the target sample. Theoretically one can consider implementing as many external calibration points as required to fit the local properties of the imaged sample at the pixel scale. From a practical point of view, this would be almost impossible to achieve and extremely time-consuming and would probably increase the risk of miss calibrating the MSI data. A workaround would be to apply an internal MS calibration by adding calibrating substances or using some known signal of the sample. Attempts may be made to use matrix ions as calibrating points but this would lead to poor calibration due to the lack of a mass range covering. It is also possible to use the sample signals as an internal MS calibration, such as the sample lipid signals [37], but this would require knowing the sample composition in advance. On the other hand, the addition of a calibrating substance to the sample can increase the competition for ionization between the calibrating compounds and the intrinsic molecules of the sample to be imaged by mass spectrometry. Worse yet, this competition between the calibrating compounds and the molecules from the sample is expected to vary from pixel to pixel, making the MSI interpretation impossible. In the absence of an internal standard, a correct MS calibration should be performed with the same acquisition parameters as the sample to obtain signals with comparable intensities, and at the same sample height.

Mass shift prevention due to unstable TIC

As previously shown, mass shift depends on several imaging parameters and is not constant from one pixel to another. It is however possible to restrain the mass shift even in complex samples by astutely tuning the acquisition parameters of the instrument, especially the MALDI source parameters.

Bacterial-interaction samples are well suited to highlight sample composition and topology variation effects on mass accuracy.

To this end and for each experiment, mass shift and peak widths were calculated using the mycosubtilin ion $[C_{49}H_{76}N_{12}O_{14}+Na]^+$, m/z = 1079.5496 as the reference ion, and presented as a heatmap in Fig. 4, thanks to our visualization software.

The images acquired on the ToF instrument show a similar mass shift, independently of the number of laser shots applied (2000 laser shots Fig. 4a and 20 laser shots Fig. 4b). Mass shift varies from −200 to +200 ppm, and appears to be superimposed with the bacterial colonies localization. On both colonies (e.g., *Bacillus velezensis GA1* and *Pseudomonas Sp. CMR12a*), where the sample is expected to be the thickest, a negative mass shift of −200 ppm is observed, in compliance with the data obtained from the various PEO embedded agar thicknesses and the brain homogenate samples. Peak width distribution is also relatively similar between these two data (Fig. 4f, g), by showing larger peak width in *Pseudomonas Sp. CMR12a* near environment. This increasing peak width in this specific region would suggest the presence of inhomogeneous matrix crystals, potentially reflecting the sample heterogeneity.

The mass shift observed in the FT-ICR data ranges from 0 to 50 ppm and appears to be correlated with the laser shots number. When using 2000 laser shots per pixel, a mass shift around 50 ppm was observed (Fig. 4c), which deviates considerably from the manufacturer's specification (± 0.5 ppm). By reducing the number of laser shots (Fig. 4d and e), the TIC was stabilized and the space charge effect became negligible. As a result, the mass shift was drastically reduced, until approaching the variation of about ± 0.5 ppm. Reducing the number of laser shots thereby demonstrate significant improvements regarding the mass shift. On the FT-ICR images, the selected reference ion is visible on every pixel of the image in the MSI data acquired with 2000 laser shots (Fig. 4c and h). When decreasing the laser shots number, the reference ion signal appears to be suppressed below the limit of detection, and is no longer visible for some pixels, left in blank on the image (see *Pseudomonas SP. CMR12a* area, Fig. 4e and j). The peak width distribution on the FT-ICR image does however remain constant (Fig. 4h, i, and j) as expected for this instrument where only one MS scan is performed per pixel.

These results show that mass shift effects are present on both ToF and FT-ICR instruments, but with different origins (either sample topology or space charge effect variations for ToF or FT-ICR analyzers respectively). However, mass shift seems to be controlled under a fine-tuning of the acquisition parameters on the FT-ICR instrument [38]. Our recommendation for FT-ICR analyzers is then to optimize the MALDI source settings to stabilize the amount of ion injected to the mass analyzer in order to limit the space charge effect during the m/z determination. This optimization will be performed according to the type and the specification of the mass analyzers (ToF, type of ICR cell

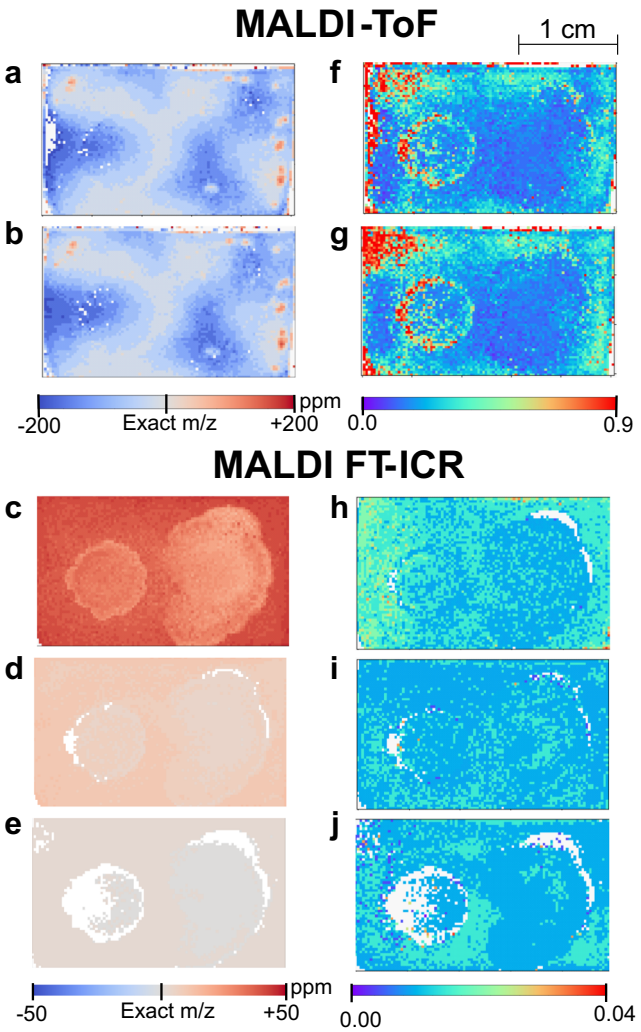


Fig. 4 Heatmaps of mass shifts (a, b, c, d, and e) and peak distribution widths (f, g, h, i, and j) for MSI of a co-culture of *Bacillus velezensis GA1* and *Pseudomonas sp. CMR12a*, using the mycosubtilin ion $[C_{49}H_{76}N_{12}O_{14}+Na]^+$, m/z = 1079.5496 as the m/z of reference. Images were acquired on a rapifleX MALDI ToF using 2000 shots (a and f) or using 20 shots (b and g) and on a solariX XR MALDI FT-ICR using 2000 shots (c and h), 200 shots (d and i), or 20 shots (e and j). The pixels where the ion m/z = 1079.5496 was not detected are left in blank

and power of the supra-conducting magnet FT-ICR, or any other critical specification of the instrument used for mass spectrometry imaging), and by monitoring the signal-to-noise ratio to avoid any loss of signal of interest.

Mass shift prevention due to heterogeneous sample topology

As discussed in the mass shift understanding section, topology variation is the main factor influencing mass shift in linear TOF instruments. Limiting the topology variation should therefore be considered to increase the accuracy and the resolution of MSI analyses with these instruments. When possible, sample preparation should ensure flat samples (as illustrated in Fig. 2). If available, stage tilt correction should also be

performed before acquisition to avoid sample height variation due to misalignment of the sample [16]. As illustrated above, a homogenous matrix deposition technique is also recommended to avoid sample topology variation caused by the matrix layer. For not flat samples (i.e., when sample topology variations cannot be reduced), sample topology variation can be compensated by acquisition strategies involving source settings adaptation to the sample height (e.g., pixel to pixel autofocus) [29]. As a last resort, the use of TOF instruments with orthogonal source (i.e., reduced kinetic energy dispersion) or trap instrument (e.g., Orbitrap, FT-ICR) could be considered since the sample topology does not influence the *m/z* measurement on these devices.

Mass shift reduction by data post processing

Mass shift amplitude in MSI data can be controlled through sample preparation and acquisition parameters optimization. When these recommendations cannot be met, post-processing methods can be applied (i) to mitigate mass shifts effects (data realignment or data recalibration) and (ii) to correct a low mass accuracy (data recalibration).

Mass spectra realignment allows comparison of all spectra contained in an MSI dataset and uses statistical tools to reassemble shifted *m/z* values to a unique value. However, this unique *m/z* value can significantly differ from the accurate *m/z* value of the ion if the initial calibration is not correct. Diversely, mass spectra recalibration will recalibrate the data based on specific signals from the sample [39, 40], resulting in a diminution of mass shifts between pixels and an improvement of the mass accuracy. However, recalibration methods depend on specific signals and experimental conditions applied to the sample (e.g., use of Ag or Au particles, rich peptide signals, *m/z* peak annotation assumption based on the state of the art) and are not applicable to the majority of MSI experiments. Recently, a new method developed by La Rocca et al. [22] introduced a more versatile approach to recalibrate MSI data based on mass shift estimation. This is applicable to different mass analyzers and on different samples with a specific focus on lipid and metabolite signals. This approach aims to recalibrate each pixel individually based on a list of potential internal calibrants generated from other annotated MSI data or databases.

In this section, we have applied this recalibration algorithm on three MSI data of bacteria co-culture obtained with different acquisition conditions on the MALDI FT-ICR (see Fig. 5c, e, and g, using 2000, 200, and 20 laser shots respectively) and on the MALDI TOF (see Fig. 5a using 20 laser shots). Applied recalibration parameters are detailed in the ESM. A minimum of 10 matches (between experimental signals and calibrating ion lists) per pixel was set to properly recalibrate the pixel correctly. Below 10 matches, the pixel is not recalibrated and removed

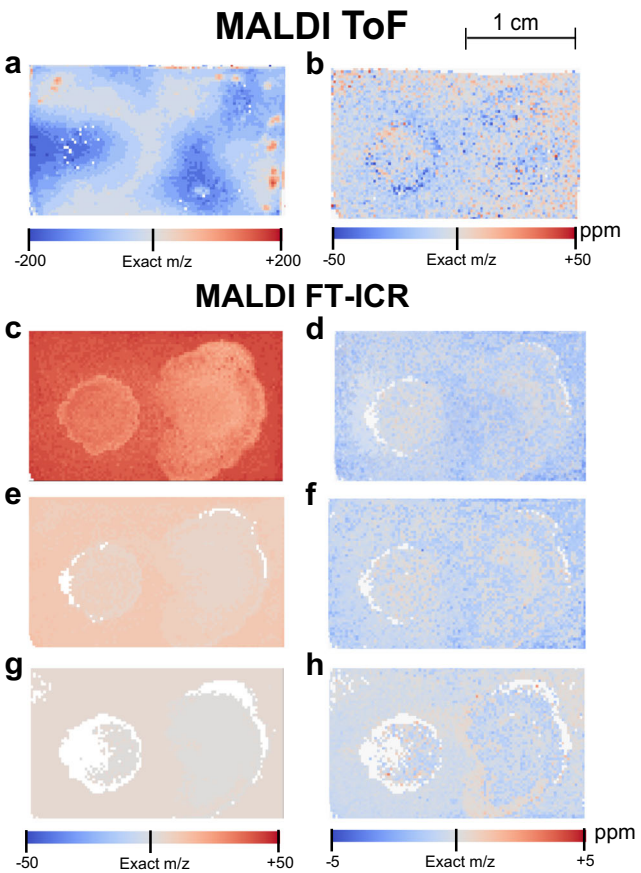


Fig. 5 Comparison of mass shift heatmaps. The figures to the left are the RAW MSI data and to the right the post-reprocessed (realignment + recalibration) MSI data. Mass shifts were calculated using the ion of mycosubtilin as a reference ($[C_{49}H_{76}N_{12}O_{14}+Na]^+$ (*m/z* = 1079.5496). **a** and **b** correspond to the original and recalibrated MALDI ToF data respectively. MALDI FT-ICR MSI data acquired with different source settings, **c** and **d** are for the original and recalibrated data using 2000 laser shots respectively, **e** and **f** are for the original and recalibrated data using 200 laser shots respectively, and g and h are for the original and recalibrated data using 20 laser shots respectively). All pixels where a minimum of 10 matches for the recalibration was not reached or where the ion *m/z* = 1079.5496 was not detected were left in blank

from the MSI data (represented with blank pixels in a reconstructed image).

The mass shift comparison of the original data to the recalibrated MSI data in Fig. 5 allows the investigation of the algorithm performance to reduce the mass shift in function of the initial quality of the MSI data. The recalibration performance was estimated by the percentage of pixels where the measured mass shift was below 1 ppm, 3 ppm, and 5 ppm for the FT-ICR instrument and below 10 ppm, 15 ppm, and 20 ppm for the ToF instrument as shown in Table 3, according to the theoretical mass from our local lipopeptides database. Before recalibration by our software, extreme values of mass shift rose up to ± 50 ppm for FT-ICR when a higher number of laser shots was set (Fig. 5c). This significant mass shift was reduced by decreasing the number of laser shots (Fig. 5e and g).

Table 3 Percentage of pixels with mass accuracy below 1, 3, and 5 ppm for MSI data acquired with MALDI ToF (rapifleX) and MALDI FT-ICR (solariX XR) and below 10, 15, and 20 ppm uniquely for ToF MSI (rapifleX). The errors were computed with the theoretical *m/z* 1079.5496 from mycosubtilin for bacteria co-culture experiments and with *m/z* 828.5514 (PC 38:6) for salmon brain homogenate experiments

	Original MSI data				Recalibrated MSI data			
MALDI FT-ICR	**< 5 ppm**	**< 2 ppm**	**< 1 ppm**	**< 0.5 ppm**	**< 5 ppm**	**< 3 ppm**	**< 1 ppm**	**< 0.5 ppm**
Bacteria 2000 laser shots	0.00	0.00	0.00	**0.00**	100.00	91.81	24.81	10.53
Bacteria 200 laser shots	0.14	0.14	0.00	0.00	100.00	99.98	69.03	37.21
Bacteria 20 laser shots	100.00	31.49	27.11	8.00	99.98	99.98	88.01	65.38
Homogenate "Beveled"	100.00	94.06	90.50	0.00	100.00	100.00	94.06	84.16
Homogenate "Flat"	100.00	98.34	75.93	0.28	100.00	100.00	100.00	75.52
MALDI ToF	**< 20 ppm**	**< 15 ppm**	**< 10 ppm**	**< 5 ppm**	**< 20 ppm**	**< 15 ppm**	**< 10 ppm**	**< 5 ppm**
Bacteria 2000 laser shots	25.42	0.00	0.00	0.00	92.03	84.21	66.08	36.01
Bacteria 20 laser shots	25.65	0.00	0.00	**0.00**	92.90	84.61	66.16	36.01
Homogenate "Bevel"	2.40	2.40	0.48	0.48	46.42	45.66	23.02	12.01
Homogenate "Flat"	0.67	0.67	0.00	0.00	98.50	94.67	79.00	44.33

The acquisition of the same sample in the ToF instrument showed that mass shift reached up to ± 200 ppm (Fig. 5a). After the application of the recalibration algorithm, the extreme values of mass shift were reduced to ± 5 ppm for the three MSI data acquired on the FT-ICR instrument (Fig. 5d, f, and h). The mass shift was reduced to below 1 ppm for 80% of the pixels after recalibration when using the optimized MALDI source settings with 20 laser shots. For the ToF instrument, the recalibration led to mass shifts below 15 ppm for 80% of the pixels, a fair improvement over the initial 200 ppm obtained in the MSI raw dataset. This significant improvement of the post-processing distribution is provided in Fig. 5d, f, and h. Overall and independently of the selected instrument, the smaller the initial mass shift, the better the recalibration efficiency. Nonetheless, it is important to highlight that the number of non-recalibrated pixels is higher when the laser shots are low, potentially due to a signal drop below the limit of detection. Consequently, a compromise can now be found between an acceptable mass shift (improved number of annotated peaks) and an acceptable ionization efficiency (improved number of detected peaks) in mass spectrometry imaging experiments. The same conclusions were obtained on the salmon brain homogenate (see ESM Fig. S3 and Table S1)

These results strongly support the effectiveness of the data recalibration to reduce the mass shift effect since all recalibrated MSI data exhibit smaller mass shift than original data. However, the performance of this algorithm depends on the initial quality of the MSI data. Consequently, it is recommended to control the mass shift at every step of the MSI workflow to ensure the best MSI post-processing recalibration. Furthermore, it is important to maintain sufficient signals to perform the recalibration of the whole MSI data.

This additional gain in accuracy after data recalibration has also been supported by the drastic increase of annotation (compound identification) on METASPACE [41] after data recalibration for all the investigated MSI data (see ESM Fig. S4).

Application MSI suggested workflow

Based on our results and on the literature, best quality MSI data with a controlled mass shift can be obtained when following the recommendations summarized in Table 4.

We applied this workflow to a mouse brain tissue section to show the quality of mass spectrometry images that can now be

Table 4 General recommendations for mass shift visualization, diagnostic, and reduction according to the type of instrument

Type of instrument	Mass shift sources	Mass shift visualization and diagnostic	Mass shift reduction/correction strategy
FT-ICR	Sample heterogeneity	Mass shift is correlated to the total ion count (TIC)	TIC should be monitored (e.g., reducing laser shots number or applying automatic gain control)
TOF	Sample thickness variation	Mass shift is correlated to sample topology	Sample flatness verification Stage tilt correction Source settings adaptations pixel by pixel (if possible)
	Inhomogeneous matrix deposition	Pixels peak width increases	Matrix deposition should be optimized

Table 5 Mass accuracies of the two monitored lipids (*m/z* 739.4675 identified as PA 36:2+K^+ and 838.6164 identified as HexCer 40:1;O3+K^+) during the mass spectrometry imaging of mouse tissue section showed in Fig. 6 acquired on a solariX XR 9.4T MALDI FT-ICR before and after post-processing recalibration using our in-house software

	Original RAW MSI data						Post-processing recalibrated MSI data					
Mass accuracy (ppm)	> 5	> 2	> 1	> 0.75	> **0.5**	> 0.25	> 5	> 2	> 1	> 0.75	> **0.5**	> 0.25
m/z	Percentage of pixels where the detected ions get a mass accuracy better than the specified range[a]											
739.4675	99.61	87.70	82.49	14.70	14.70	0.00	100.00	100.00	99.98	96.40	83.41	49.15
838.6164	100.00	96.91	14.54	14.54	14.54	0.00	100.00	99.84	99.39	86.37	83.25	55.86

[a] Pixels where the ions were not detected are excluded for the percentage calculation

0.5 ppm is the mass accuracy specification of the solariX XR FT-ICR according to the manufacturer in normal operation

routinely achieved with the solariX XR 9.4T mass spectrometers. Table 5 shows the percentage of pixels in different ranges of mass accuracies (from 0.25 to more than 5 ppm). Two lipid ions, PA 36:2 with a potassium adduct ($C_{39}H_{73}O_8K$, *m/z* = 739.4675) and HexCer 40:1;O3 ($C_{46}H_{89}NO_9K$, *m/z* = 838.6164) were monitored to illustrate this purpose. These ions have been identified thanks to the high mass accuracy achieved throughout the entire image and the contribution of the isotopic fine structure obtained by the dynamically harmonized ICR cell of the solariX XR (see ESM Figs. S5 and S6). The ions show a different distribution in the mouse brain tissue (Fig. 6b and c for [PA 36:2+K]$^+$ and [HexCer 40:1;O3+K]$^+$ respectively). However, their calculated mass shift is not related to their distribution (Fig. 6d and e for [PA 36:2+K]$^+$ and [HexCer 40:1;O3+K]$^+$ respectively), but to the areas where the TIC is the highest (Fig. 6a, corresponding to the fiber tracts region). As reported in Table 5, the [PA 36:2+K]$^+$ ion, mainly distributed in the low shifted regions (i.e., where the TIC is low), has more than 80% of its pixels with an accuracy better than 1 ppm before recalibration process. Diversely, the ion [HexCer 40:1;O3+K]$^+$, mainly distributed in the high shifted regions (i.e., where the TIC is high), presents less than 15% of its pixels with an accuracy better than 1 ppm. After the recalibration method, more than 80% of the pixels show a mass accuracy better than 0.5 ppm, and the mass shift distribution no longer corresponds to the TIC for both lipids (Fig. 6f and g for [PA 36:2+K]$^+$ and [HexCer 40:1;O3+K]$^+$ respectively).

Conclusion

This study presented a critical analysis of experimental parameters allowing a significant reduction of the mass shift occurring in MSI analysis. The workflow was optimized on two types of instruments commonly used in MALDI MSI (i.e.,

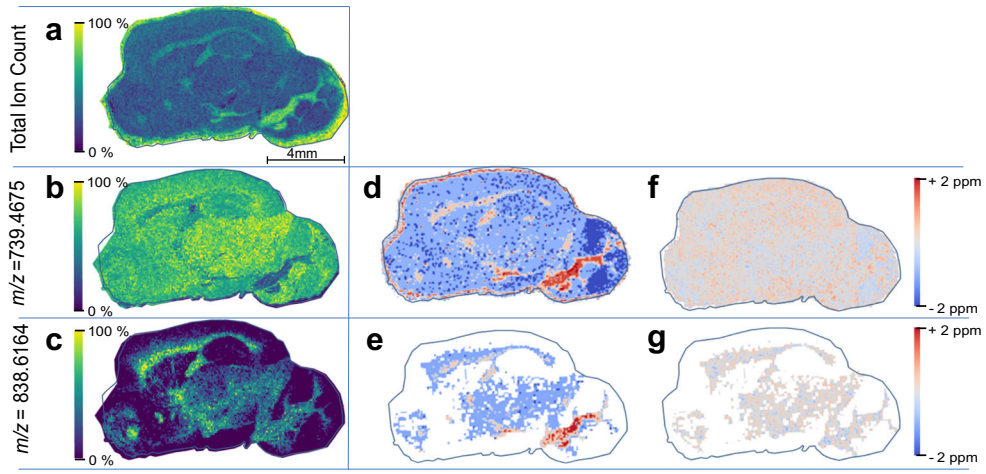


Fig. 6 Mass spectrometry imaging of mouse tissue section acquired on a solariX XR 9.4T MALDI FT-ICR using our recommended optimized workflow (i.e., flat tissue section, homogeneous HCCA matrix deposition, 20 laser shots par pixel for TIC (total ion current) stabilization) at nominal resolving power FWHM higher than 220,000 at *m/z* 800. **a** is the image of the TIC. **b** and **c** depict mass spectrometry images of [PA 36:2+K]$^+$ and [HexCer 40:1;O3+K]$^+$ respectively (black pixel in (**c**) corresponds to pixel where [HexCer 40:1;O3+K]$^+$ is not detected). **d** and **f** depict mass shifts heatmaps of [PA 36:2+K]$^+$before and after post-processing recalibration respectively. **e** and **g** are the mass shifts heatmaps of [HexCer 40:1;O3+K]$^+$ before and after post-processing recalibration respectively

rapifleX and solariX XR 9.4T from Bruker) but could be extended to others. The narrowing of the *m/z* distribution, thanks to mass shift reduction coupled with internal recalibration, allowed a better mapping of compounds and enabled more confident signal annotations. The origins of mass shift can differ from one MS instrument to another. In the MALDI ToF instrument, the sample topology directly influences the kinetic energy given to the ions leading to *m/z* measurement deviations. We reported that the sample preparation can be optimized (i.e., improving the sample flatness and the homogeneity of the matrix deposition) to reduce mass shift. The main factor influencing mass shift in FT-ICR is the variation of the total ion current (TIC) leading to Coulomb expansion of the ion cloud in the FT-ICR cell. Maintaining the TIC values below a given value and therefore avoiding Coulomb expansion during the acquisition facilitates the drastic reduction of the mass shift in these MS instruments. The optimization of the sample preparation and the MS instrument parameters can easily be applied in general MSI workflow.

Whenever possible, the use of orthogonal sources for TOF instruments (i.e., reducing the kinetic energy dispersion) and automatic gain control for FT-ICR-MS instruments (i.e., controlling the ions injected in the FT-ICR cell) are also highly recommended to minimize mass shift in MSI. Meta-analysis of data by bio-informatic tools suggests indeed that Orbitrap should be less impacted by the mass shift [22]. However, the metadata of the acquisition methods are still scarce and the MSI parameter influences on mass shift remain to be investigated. Recent data obtained in our lab on biological sample obtained on a MALDI ion mobility ToF (unpublished results obtained on timstof FleX prototype, Bruker, Bremen, Germany) strongly suggest that the effect of surface topology on mass shifts is negligible since the kinetic energy dispersion practically no longer exists after the ion mobility section of the instrument.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00216-021-03174-1.

Acknowledgments The authors would like to thank the University of Maastricht (The Netherlands) and Dr. S. Ellis for the use of the HTX Sprayer and the HTX Sublimator, Pr. M. Ongena and Dr. A. Argüelles Arias, Microbial Processes and Interactions (MiPI) of Gembloux Agro Biotech (Belgium), for the bacteria samples. D. Van Kruining and Pr. P. Martinez Martinez, Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University (The Netherlands), are kindly acknowledged for providing the mouse brain tissue samples. CAREM platform (University of Liège, Belgium) is thanked for performing electron microscopy acquisition of the crystal samples.

Funding This work was funded by the Excellence Of Science Program, FNRS F.R.S (Rhizoclip EOS30650350), by the European Union Horizon 2020 research and innovation program under grant agreement No. 731077 and by interreg's EURLIPIDS project (R-8598). The MALDI ToF RapifleX and the MALDI FT-ICR SolariX XR were funded by FEDER BIOMED HUB Technology Support (number 2.2.1/996).

Compliance with ethical standards

All experiments were done with permission of the Committee on Animal Welfare, Maastricht University, according to Dutch governmental rules.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Amstalden van Hove ER, Smith DF, Heeren RM. A concise review of mass spectrometry imaging. J Chromatogr A. 2010;1217(25):3946–54. https://doi.org/10.1016/j.chroma.2010.01.033.
2. Luxembourg SL, Mize TH, McDonnell LA, Heeren RMA. High-spatial resolution mass spectrometric imaging of peptide and protein distributions on a surface. Anal Chem. 2004;76(18):5339–44. https://doi.org/10.1021/ac049692q.
3. Caprioli RM, Farmer TB, Gile J. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. AnalChem. 1997;69(23):4751–60. https://doi.org/10.1021/ac970888i.
4. Kune C, McCann A, La Rocca R, Arias AA, Tiquet M, Van Kruining D, et al. Rapid visualization of chemically related compounds using Kendrick mass defect as a filter in mass spectrometry imaging. Anal Chem. 2019;91(20):13112–8. https://doi.org/10.1021/acs.analchem.9b03333.
5. Brown SC, Kruppa G, Dasseux JL. Metabolomics applications of FT-ICR mass spectrometry. Mass Spectrom Rev. 2005;24(2):223–31. https://doi.org/10.1002/mas.20011.
6. McLafferty FW. Tandem mass spectrometry. Science. 1981;214(4518):280–7. https://doi.org/10.1126/science.7280693.
7. de Hoffmann E. Tandem mass spectrometry: a primer. J Mass Spectrom. 1996;31(2):129–37. https://doi.org/10.1002/(sici)1096-9888(199602)31:2<129::Aid-jms305>3.0.Co;2-t.
8. Griffiths WJ, Jonsson AP, Liu S, Rai DK, Wang Y. Electrospray and tandem mass spectrometry in biochemistry. Biochem J. 2001;355(Pt 3):545–61. https://doi.org/10.1042/bj3550545.
9. Kanu AB, Dwivedi P, Tam M, Matz L, Hill HH Jr. Ion mobility–mass spectrometry. J Mass Spectrom. 2008;43(1):1–22. https://doi.org/10.1002/jms.1383.
10. Lanucara F, Holman SW, Gray CJ, Eyers CE. The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics. Nature Chemistry. 2014;6(4):281-94. https://doi.org/10.1038/nchem.1889
11. Lapthorn C, Pullen F, Chowdhry BZ. Ion mobility spectrometry-mass spectrometry (IMS-MS) of small molecules: separating and assigning structures to ions. Mass Spectrom Rev. 2013;32(1):43–71. https://doi.org/10.1002/mas.21349.
12. Gorshkov MV, Nikolaev EN. Optimal cyclotron radius for high resolution FT-ICR spectrometry. Int J Mass Spectrom Ion Process. 1993;125(1):1–8. https://doi.org/10.1016/0168-1176(93)80012-4.
13. Easterling ML, Mize TH, Amster IJ. Routine part-per-million mass accuracy for high- mass ions: space-charge effects in MALDI FT-ICR. Anal Chem. 1999;71(3):624–32. https://doi.org/10.1021/ac980690d.
14. Nikolaev EN, Kostyukevich YI, Vladimirov GN. Fourier transform ion cyclotron resonance (FT ICR) mass spectrometry: theory and simulations. Mass Spectrom Rev. 2016;35(2):219–58. https://doi.org/10.1002/mas.21422.
15. Phelan VV, Liu WT, Pogliano K, Dorrestein PC. Microbial metabolic exchange–the chemotype-to-phenotype link. Nat Chem Biol. 2012;8:26.

16. Malys BJ, Piotrowski ML, Owens KG. Diagnosing and correcting mass accuracy and signal intensity error due to initial ion position variations in a MALDI TOFMS. J Am Soc Mass Spectrom. 2018;29(2):422–34. https://doi.org/10.1021/jasms.8b05757.
17. Berghmans E, Van Raemdonck G, Schildermans K, Willems H, Boonen K, Maes E, et al. MALDI mass spectrometry imaging linked with top-down proteomics as a tool to study the non-small-cell lung cancer tumor microenvironment. Methods Protoc. 2019;2(2):44. https://doi.org/10.3390/mps2020044.
18. Gemperline E, Chen B, Li L. Challenges and recent advances in mass spectrometric imaging of neurotransmitters. Bioanalysis. 2014;6(4):525–40. https://doi.org/10.4155/bio.13.341.
19. McDonnell LA, Heeren RMA. Imaging mass spectrometry. Mass Spectrom Rev. 2007;26(4):606–43. https://doi.org/10.1002/mas.20124.
20. Nihorimbere V, Cawoy H, Seyer A, Brunelle A, Thonart P, Ongena M. Impact of rhizosphere factors on cyclic lipopeptide signature from the plant beneficial strain Bacillus amyloliquefaciens S499. FEMS Microbiol Ecol. 2012;79(1):176–91. https://doi.org/10.1111/j.1574-6941.2011.01208.x.
21. Debois D, Ongena M, Cawoy H, De Pauw E. MALDI-FTICR MS imaging as a powerful tool to identify Paenibacillus antibiotics involved in the inhibition of plant pathogens. J Am Soc Mass Spectrom. 2013;24(8):1202–13. https://doi.org/10.1007/s13361-013-0620-2.
22. La Rocca R, Kune C, Tiquet M, Stuart L, Alexandrov T, De Pauw E, et al. Using biological signals for mass recalibration of mass spectrometry imaging data. ChemRxIv. 2020. https://doi.org/10.26434/chemrxiv.12901679.v1.
23. Sud M, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, et al. LMSD: LIPID MAPS structure database. Nucleic Acids Res. 2007;35(Database issue):D527–32. https://doi.org/10.1093/nar/gkl838.
24. Ievlev AV, Belianinov A, Jesse S, Allison DP, Doktycz MJ, Retterer ST, et al. Automated interpretation and extraction of topographic information from time of flight secondary ion mass spectrometry data. Sci Rep. 2017;7(1):17099. https://doi.org/10.1038/s41598-017-17049-y.
25. Yang JY, Phelan VV, Simkovsky R, Watrous JD, Trial RM, Fleming TC, et al. Primer on agar-based microbial imaging mass spectrometry. J Bacteriol. 2012;194(22):6023–8. https://doi.org/10.1128/jb.00823-12.
26. Buchberger AR, DeLaney K, Johnson J, Li L. Mass spectrometry imaging: a review of emerging advancements and future insights. Anal Chem. 2018;90(1):240–65. https://doi.org/10.1021/acs.analchem.7b04733.
27. Peukert M, Matros A, Lattanzio G, Kaspar S, Abadía J, Mock HP. Spatially resolved analysis of small molecules by matrix-assisted laser desorption/ionization mass spectrometric imaging (MALDI-MSI). New Phytol. 2012;193(3):806–15. https://doi.org/10.1111/j.1469-8137.2011.03970.x.
28. Guenther S, Koestler M, Schulz O, Spengler B. Laser spot size and laser power dependence of ion formation in high resolution MALDI imaging. Int J Mass Spectrom. 2010;294(1):7–15. https://doi.org/10.1016/j.ijms.2010.03.014.
29. Kompauer M, Heiles S, Spengler B. Autofocusing MALDI mass spectrometry imaging of tissue sections and 3D chemical topography of nonflat surfaces. Nat Methods. 2017;14(12):1156–8. https://doi.org/10.1038/nmeth.4433.
30. Gabriel SJ, Schwarzinger C, Schwarzinger B, Panne U, Weidner SM. Matrix segregation as the major cause for sample inhomogeneity in MALDI dried droplet spots. J Am Soc Mass Spectrom. 2014;25(8):1356–63. https://doi.org/10.1007/s13361-014-0913-0.
31. Pei X-L, Liu X-N, Du J-L, Gong C, Xu X. MALDI-MS imaging of lipids in corn using a flexible ultrasonic spraying device as matrix deposition method. Int J Mass Spectrom. 2020;455:116373. https://doi.org/10.1016/j.ijms.2020.116373.
32. Fournier I, Tabet JC, Bolbach G. Irradiation effects in MALDI and surface modifications: part I: sinapinic acid monocrystals. Int J Mass Spectrom. 2002;219(3):515–23. https://doi.org/10.1016/S1387-3806(02)00704-2.
33. Fournier I, Marinach C, Tabet JC, Bolbach G. Irradiation effects in MALDI, ablation, ion production, and surface modifications. Part II: 2,5-dihydroxybenzoic acid monocrystals. J Am Soc Mass Spectrom. 2003;14(8):893–9. https://doi.org/10.1016/S1044-0305(03)00347-7.
34. McDonnell LA, Mize TH, Luxembourg SL, Koster S, Eijkel GB, Verpoorte E, et al. Using matrix peaks to map topography: increased mass resolution and enhanced sensitivity in chemical imaging. Anal Chem. 2003;75(17):4373–81. https://doi.org/10.1021/ac034401j.
35. Moskovets E, Karger BL. Mass calibration of a matrix-assisted laser desorption/ionization time-of-flight mass spectrometer including the rise time of the delayed extraction pulse. Rapid Commun Mass Spectrom. 2003;17(3):229–37. https://doi.org/10.1002/rcm.898.
36. Xiang B, Prado M. An accurate and clean calibration method for MALDI-MS. J Biomol Tech. 2010;21(3):116–9.
37. Hart PJ, Francese S, Claude E, Woodroofe MN, Clench MR. MALDI-MS imaging of lipids in ex vivo human skin. Anal Bioanal Chem. 2011;401(1):115–25. https://doi.org/10.1007/s00216-011-5090-4.
38. Tiquet M, La Rocca R, Van Kruining D, Martinez-Martinez P, Eppe G, De Pauw E, et al. Mass spectrometry imaging using dynamically harmonized FT-ICR at million resolving power: rationalizing and optimizing sample preparation and instrumental parameters. ChemRxIv. 2020. https://doi.org/10.26434/chemrxiv.13013900.v1.
39. Cox J, Michalski A, Mann M. Software lock mass by two-dimensional minimization of peptide mass errors. J Am Soc Mass Spectrom. 2011;22(8):1373–80. https://doi.org/10.1007/s13361-011-0142-8.
40. Boskamp T, Lachmund D, Casadonte R, Hauberg-Lotte L, Kobarg JH, Kriegsmann J, et al. Using the chemical noise background in MALDI mass spectrometry imaging for mass alignment and calibration. Anal Chem. 2020;92(1):1301–8. https://doi.org/10.1021/acs.analchem.9b04473.
41. Palmer A, Phapale P, Chernyavsky I, Lavigne R, Fay D, Tarasov A, et al. FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. Nat Methods. 2017;14(1):57–60. https://doi.org/10.1038/nmeth.4072.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Associated papers of Chapter 3

A3- Original publication of “FT-ICR Mass Spectrometry Imaging at Extreme Mass Resolving Power Using a Dynamically Harmonized ICR Cell with 1ω or 2ω Detection.”

analytical chemistry

pubs.acs.org/ac

Article

FT-ICR Mass Spectrometry Imaging at Extreme Mass Resolving Power Using a Dynamically Harmonized ICR Cell with 1ω or 2ω Detection

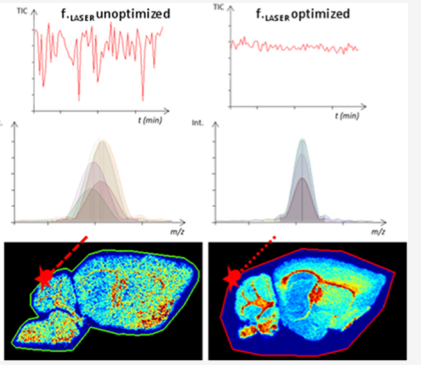
Mathieu Tiquet, Raphaël La Rocca, Stefan Kirnbauer, Samuele Zoratto, Daan Van Kruining, Loïc Quinton, Gauthier Eppe, Pilar Martinez-Martinez, Martina Marchetti-Deschmann, Edwin De Pauw, and Johann Far*

Cite This: *Anal. Chem.* 2022, 94, 9316−9326

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: MALDI mass spectrometry imaging (MALDI MSI) is a powerful analytical method for achieving 2D localization of compounds from thin sections of typically but not exclusively biological samples. The dynamically harmonized ICR cell (ParaCell) was recently introduced to achieve extreme spectral resolution capable of providing the isotopic fine structure of ions detected in complex samples. The latest improvement in the ICR technology also includes 2ω detection, which significantly reduces the transient time while preserving the nominal mass resolving power of the ICR cell. High-resolution MS images acquired on FT-ICR instruments equipped with 7T and 9.4T superconducting magnets and the dynamically harmonized ICR cell operating at suboptimal parameters suffered severely from the pixel-to-pixel shifting of m/z peaks due to space-charge effects. The resulting profile average mass spectra have depreciated mass measurement accuracy and mass resolving power under the instrument specifications that affect the confidence level of the identified ions. Here, we propose an analytical workflow based on the monitoring of the total ion current to restrain the pixel-to-pixel m/z shift. Adjustment of the laser parameters is proposed to maintain high spectral resolution and mass accuracy measurement within the instrument specifications during MSI analyses. The optimized method has been successfully employed in replicates to perform high-quality MALDI MS images at resolving power (FWHM) above 1,000,000 in the lipid mass range across the whole image for superconducting magnets of 7T and 9.4T using 1 and 2ω detection. Our data also compare favorably with MALDI MSI experiments performed on higher-magnetic-field superconducting magnets, including the 21T MALDI FT-ICR prototype instrument of the NHMFL group at Tallahassee, Florida.



1. INTRODUCTION

Matrix-assisted laser desorption/ionization mass spectrometry imaging (MALDI MSI) has emerged as a label-free analytical method for monitoring the relative abundance (despite severe limitations due to suppression effects) and spatial distribution of a wide variety of analytes, especially for biological samples.[1−3] To properly distinguish isobaric compounds[4] inherent to the complexity of biological samples, a high resolving power at full width at half-maximum ($R.P._{FWHM}$ > 300,000 at 400m/z) and a reliable mass measurement accuracy (MMA) are required in the absence of an upstream separation method (such as ion mobility). These performances are commonly achieved by a Fourier transform mass analyzer such as Fourier transform-ion cyclotron resonance (FT-ICR).[5] The Bruker dual ion source ESI/MALDI FT-ICR (solariX and scimaX) is a hybrid instrument equipped with multipoles, a quadrupole, and a collision cell for precursor ion selection and fragmentation. Ions produced by electrospray (ESI) and MALDI are also accumulated in the multipole region to prepare the ion packet to be introduced into the ICR cell. Consequently, almost any combination of MALDI laser settings is compatible with any transient time for mass spectra acquisition. Recent developments introduced by Nikolaev et al.[6−8] led to the dynamically harmonized ICR cell commercialized by Bruker in the solariX XR and scimaX XR brand FT-ICR mass spectrometers under the name ParaCell. This new cell offers the highest mass resolving power (R.P.) currently achievable for such instruments[9] (around 1,000,000 in the lipid mass range in broadband mode), and mass measurement accuracy typically is in the sub-ppm range. These improvements drastically increase the confidence level in the determination of the empiric formula of precursor ions, especially when including the fine isotopic structure.[10]

Received: February 14, 2022
Accepted: May 9, 2022
Published: May 23, 2022



Downloaded via UNIV DE LIEGE on July 5, 2022 at 16:01:31 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

© 2022 The Authors. Published by American Chemical Society

https://doi.org/10.1021/acs.analchem.2c00754
Anal. Chem. 2022, 94, 9316−9326

Table 1. Sets of Parameters Used in the Original and Reoptimized Methods[a]

		solariX XR		scimaX 2XR 1 or 2ω	
parameters	(unit)	original	reoptimized	original	reoptimized
laser focus[b]	%	98	80	93	85
laser shots	(#shots)	600	[2; 10]	400	6
laser frequency	(Hz)	1000	#shots × 10	1000	60
sweep excitation power	(%)	22	[16; 18]	20	18
front & back trap plate	(V)	1.5	1.35	3	3.06
analyzer entrance	(V)	−10	−10	−10	−10
side kick	(V)	5	[6; 10]	0.2	3
side kick offset	(V)	−1.0	−1.5	−1.5	−1.5
time of flight	(ms)	1.2	1.2	1.0	0.7[c]

[a]The laser power was adjusted to get the lower power possible when the TIC signal was reaching 5×10^8 cps. Values in brackets show a working range. [b]Small and medium laser focus for solariX XR and scimaX 2XR, respectively. [c]Time of flight set at 0.7 ms for the 2ω acquisition for 16 M data point only.

Moreover, the introduction[11−13] and the experimental application[12,14] of 2ω detection drastically improved the mass R.P. and the scan duration of the ICR transient signals.

To obtain the highest quality of mass spectrometry images (MSI) in terms of mass R.P. and lateral resolution, each step of the imaging workflow has to be properly optimized. The experimental optimization of the FT-ICR-MS(I) instrument[15] was studied by Carlos Afonso and Abdellah Tebani's group. The sample preparation affects the ionization efficiency as well as the local diffusion of the analytes in tissue sections. Different experimental parameters were evaluated extensively in the literature including slice thickness,[16−18] matrix and solvent selection,[19−23] and optimization of the automatized matrix deposition.[24] When using the recommended parameters and optimized methods intended for the previous ICR design, the "Infinity Cell", the best performance in terms of $R.P._{FWHM}$ and mass accuracy was far from instrument specifications due to abnormally large mass shifts. From an instrumental point of view, significant deviations in the number of injected ions between scan events heavily affect the global performance of the Paracell. Pixel-to-pixel fluctuations of the ion current during the MSI experiments cause a nonrepeatable space-charge effect between pixels in regard to the MALDI MS calibration procedure. In general, this phenomenon can be corrected using a lock-mass calibration during acquisition[25,26] which would ideally require several homogeneously distributed analytes. These targets could be added before matrix deposition at the risk of inducing more or less severe ion-suppression effect(s) and lateral diffusion. An alternative is to use postacquisition recalibration software,[27] which however can be time-consuming due to format conversion and computational steps depending on the size and format of the dataset.

This study reports optimized instrument conditions to mitigate the abnormal mass shifts observed during high/extreme-resolution MALDI FT-ICR MSI fitted with the ParaCell. We present here such an optimization on sample preparation and acquisition parameters to produce MS images at $R.P._{FWHM}$ at least better than 500,000 at m/z 800 (better than 1 million at m/z: 400) in broadband mode using a solariX XR 9.4T and a scimaX 7T 2XR.

2. MATERIAL AND METHODS

2.1. Chemicals. Acetone and methanol (HPLC grade) were obtained from Biosolve (Valkenswaard, Netherlands). Trifluoroacetic acid (TFA, 99%), α-cyano-4-hydroxycinnamic acid (α-HCCA, purity 97%), and red phosphorus (>97%) were purchased from Sigma-Aldrich (Taufkirchen, Germany). The internal standard SPLASH LipidoMIX containing deuterated lipids from different families was purchased from Avanti Polar Lipids (Alabaster, Alabama) via Sigma-Aldrich.

2.2. Animal Handling. Transgenic mice were purchased from Dr. Mary Jo LaDu (University of Illinois at Chicago) and bred in-house at MHeNs at Maastricht University as described elsewhere.[28] In short, human-APOE4 knock-in mice in which the mouse APOE gene was replaced by human APOE were crossbred with 5xFAD mice (Jackson laboratory) carrying human familial Alzheimer's disease mutations PSEN1 and APP to obtain E4FAD mice with increased Aβ peptide production.[28,29] Female E4FAD mice over 6 months of age were sacrificed by CO_2 inhalation, and then brains were extracted. The mice brains were cut across the sagittal midline, immediately fresh-frozen in liquid nitrogen, and subsequently stored at −80 °C. For transportation, samples were placed on dry ice and transferred to the University of Liège to be long-term stored and conserved again at −80 °C before further handling. All procedures were approved by the Animal Welfare Committee of Maastricht University (no. AVD107002015177) and were performed according to Dutch federal regulations for animal protection.

Natural AB-type zebrafish were bred by the Groupement Interdisciplinaire de Génoprotéomique Appliquée (GIGA) at ULiège under the supervision of Pr. Marc Muller. The aquarium water was thermostated at 28 °C with a circadian cycle of 14 h of light and 10 h of darkness. One-month-old fish were first anesthetized by adding tricaine mesylate to a concentration of 0.04% and then increased to 0.16% to induce cardiac arrest. The fish were then embedded in gelatin (350 $mg{\cdot}mL^{-1}$) and stored at −80 °C for at least 24 h. All procedures were approved by the Animal Welfare Committee of the University of Liège (no. 20−2284) and were performed according to Belgian federal regulations for animal protection.

2.3. Tissue Sectioning. Sectioning was performed on a CryoStar NX70 (Thermo Fisher Scientific, Massachusetts) set at −20 °C. SEC35e low-profile razor blades (Thermo Fisher Scientific, Massachusetts) were employed at −15 °C during the sectioning. Mouse brain and zebrafish whole-body sagittal slices were sectioned at medium thicknesses of 14 and 8 μm to keep a good amount of material for ionization[30] while easing the collection of seriated slices. Cryosections were thaw-mounted onto indium tin oxide (ITO)-coated conductive glass slides (Bruker Daltonics, Bremen, Germany).

https://doi.org/10.1021/acs.analchem.2c00754
Anal. Chem. 2022, 94, 9316−9326

2.4. Matrix Coating. Prior to matrix deposition, samples were dried in a vacuum desiccator for 15 min or until no visible wetness was observable. Dried samples were coated with matrix using the automatic sprayer SunCollect MALDI spotter (SunChrom, Friedrichsdorf, Germany). Matrix solution contained 5 mg.mL^{-1} α-HCCA dissolved in methanol and Milli-Q water acidified with trifluoroacetic acid (MeOH:-H_2O:TFA 9:0.99:0.01 *v:v:v*). During the spraying procedure, the nozzle was positioned to its lowest setting and its moving speeds in the *X* and *Y* axis were set at medium 10 (1540 mm·min^{-1}). Matrix flow rates started from 5 μL·min^{-1} up to the fourth layer for which flow rates were increased to 10 μL·min^{-1} until the last deposition layer. The number of layers required to obtain a homogeneous coating of roughly 10 nmol·mm^{-2} of matrix was calculated for each spray deposition. The amount of matrix sprayed is confirmed by weight comparison of the ITO glass slide before and after the spray process. Later in this study, the optimized amount of deposited matrix is 5 nmol·mm^{-2}.

2.5. MALDI Mass Spectrometry Imaging. Mass spectrometry acquisitions were performed on ESI/MALDI dual-source MALDI FT-ICRs equipped with the ParaCell (solariX XR 9.4T and scimaX 2XR 7T, Bruker Daltonics, Bremen, Germany) operating in MALDI positive mode with data point sizes of 2, 4, and 8 or 16 M for the scimaX 2XR in the 300−1200m/z mass range using the Amplitude mode. Other relevant parameters for the solariX XR 9.4T and the scimaX 7T 2XR are listed in Table 1. The shimming of the ICR cells was performed using the recommended procedure by the manufacturer based on the infusion of sodium trifluoroacetic solution in 50% acetonitrile. Before m/z calibration, the tissue to be analyzed, or a seriated tissue test section, was first probed to determine the minimum required laser power and monitor the ion current to set the laser parameters. Then, the m/z calibration of the spectrometer was performed using the odd-numbered clusters of red phosphorus spotted close to the analyzed samples.[31] During calibration, a TIC as close as possible to the value obtained on tissue was targeted with the help of selective accumulation upper and lower cutoff set at maximum ±20% of the probed TIC on the sample. In our case, the laser powers of the solariX XR and scimaX 2XR were adjusted from 10 to 16% depending on the number of ions to be injected into the ICR cell. The laser power could be higher depending on the rate of wear of the laser. Typical high vacuum values of the ICR cells were about 2.5 × 10^{-10} mbar, and the targeted TIC with a data size of 4 M was 5 × 10^8 cps.

Automated acquisitions were performed using the software FlexImaging 5.0 (Bruker Daltonics, Bremen, Germany) with a raster of 50 μm in both (*x*,*y*) axes.

2.6. Data Processing. All datasets were visualized with SCiLS Lab 2016b (SCiLS, Bremen, Germany) after conversion into scilslab format using the SQlite file generated by the instrument. MALDI MSI were generated after total ion count normalization (unless specified otherwise) and automatic hot spot removal (at 99% quantile). Database bulk structure searches were performed using the LIPID MAPS Structure Database (LMSD) tool offered by LIPID MAPS Lipidomics Gateway (lipidmaps.org).[32,33] Queries were submitted on the full database with a 5mDa mass tolerance for $[M + H]^+$, $[M + H\text{-}H_2O]^+$, $[M + Na]^+$, $[M + K]^+$, and $[M + 2Na\text{-}H]^+$ ions. The nomenclature of lipids used in this work is based on the recommended lipids classification by Fahy and co-workers.[34] An in-house script written in R language has been used to calculate the standard deviation for MMA and R.P. for a given m/z window within an MSI dataset converted to imzML format by FlexImaging 5.0.

3. RESULTS AND DISCUSSION

Most of the published work reporting the optimization of MALDI FT-ICR MSI methods was performed on instruments fitted with superconducting magnets of 12T and 15T or above. We propose here to visit or revisit the influence of the instrument parameters to produce MALDI images with the highest possible mass R.P. and MMA that such instruments can offer on most readily available commercial FT-ICR instruments equipped with a 7 or 9.4T magnet.

A higher magnetic field limits the space-charge effects inside the ICR cell and provides improved tolerance in regard to the number of injected ions. During MALDI MSI experiments, the amount of injected ion significantly varies due to the intrinsic heterogeneity of the biological material in terms of molecular composition and the dynamic range of the acquisition method. Consequently, instruments using lower magnetic fields could be substantially affected by impaired performance. The analytical workflow, from matrix deposition to ion optics parameters, was investigated and applied to MALDI FT-ICR instruments using 7T (scimaX 2XR) or 9.4T (solariX XR) superconducting magnets and the 1ω or 2ω detection mode, when available. Additionally, ion source parameters, only poorly explored in the literature, were explored at optimal settings to improve the quality of MALDI images at very high mass R.P.

3.1. Magnetic Field and Charge-Space Effects. The determination of the m/z ratios by FT-ICR is obtained by converting the rotational frequencies of the ions by Fourier Transformation, which depends on the masses and carried charges under the influence of the applied magnetic fields. The space-charge effect limits the performance of an FT-ICR due to the influence of the charge repulsion between ion packets if the ICR cell is loaded with more ions than the magnetic field can constrain. Using a superconductive magnet with higher magnetic fields would limit this influence as discussed elsewhere.[35,36]

In this paper, the main focus is on limiting the ion current fluctuation during the MSI acquisition and restricting the observed mass shift to the extent possible. This corresponds to limiting the fluctuations of the space-charge effect components at a constant magnetic field. Nonetheless, the requirement for the magnetic field to produce MALDI images at extreme mass R.P. was evaluated by comparing experimental results from 9.4T to a 7T operated in 1ω and/or 2ω detection modes.

Considering theory, the best chance to restrain the experimental mass shift in the average MS images under the specification of the FT-ICR is to prevent space-charge effects. This was investigated by experimental work and a literature survey for the different steps of the production of MALDI images, from sample preparation (matrix deposition protocols) to instrumental parameters (ion optics transmission, ICR ion optics). Furthermore, optimization works for laser adjustments are still scarce in the literature and were also investigated.

The monitored outputs of experimental parameters during measurement at very high mass R.P. (500,000 and above in the lipid mass range) were mainly the stability of the total ion current (TIC) fluctuation, the mass shift (i.e., pixel-to-pixel variation of the m/z peak apexes), and the mass R.P. (R.P. expressed as full width at half-maximum, FWHM) for the

https://doi.org/10.1021/acs.analchem.2c00754
Anal. Chem. 2022, 94, 9316−9326

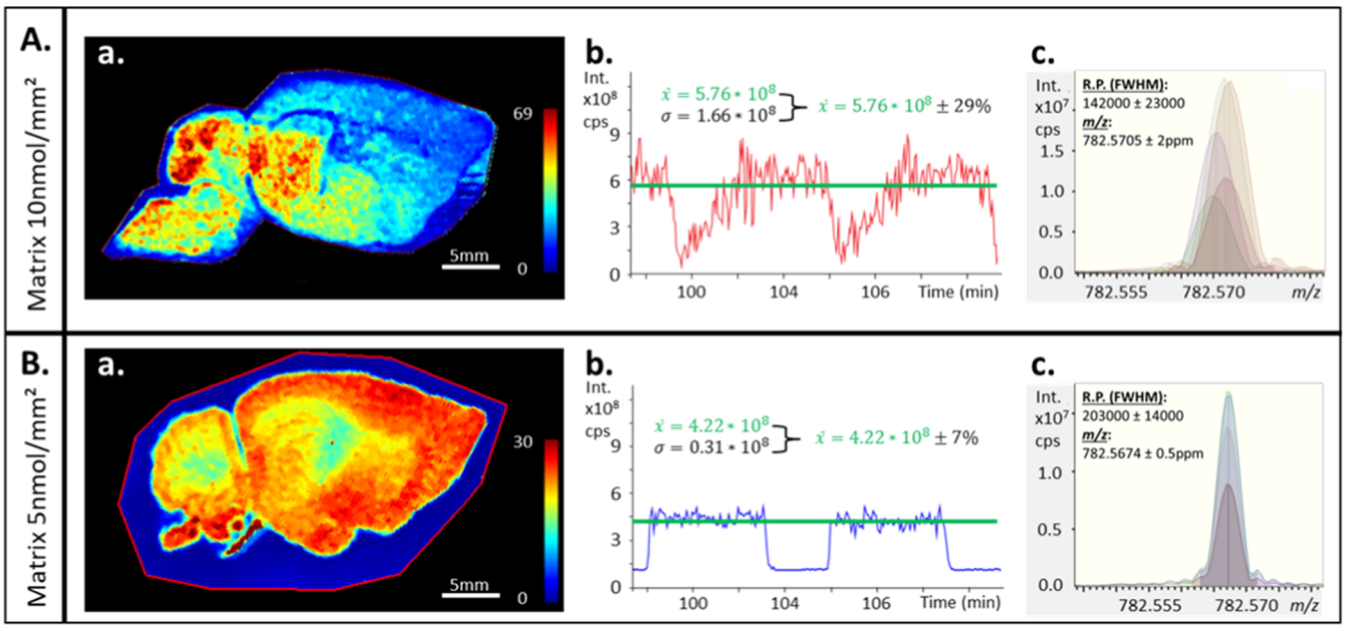


Figure 1. Comparison of dynamically harmonized MALDI FT-ICR MSI acquired on a solariX XR 9.4T at 4M with a manufacturer recommendation-based method (A) and a 6 laser shot-based method (B); see Section 3.2 for details. Reconstructed heat maps of the non-normalized total ion count of MS images (a). Portions of the TIC over time of the MSI acquisition and the computed mean intensities with standard deviation (b). Observable gaps on the TIC are values from pixels outside of the tissue section and were excluded to compute the standard deviation. The TIC presented in the top panel was obtained when 10 nmol·mm^{-2} α-HCCA matrix was deposited using 98% laser focus and from 5 nmol·mm^{-2} α-HCCA matrix with 80% laser focus (bottom panel). Multipixel mass spectra overlay of m/z 782.5674 shows a notable improvement in terms of mass R.P. and mass accuracy measurements during the MALDI images between the unstable (top panel) and stable (bottom panel) total ion count (c).

individual pixel and in the profile average mass spectrum of the image. The apparent intrascan dynamic range between major and minor peaks of lipids was also monitored and reported during the laser parameter optimization.

The effect of significant TIC fluctuations on the m/z shift was evaluated by monitoring the signal produced by a standard Splash LipidomiX solution of deuterated lipids spotted with α-HCCA matrix on an ITO glass slide. After the acquisition of a single MALDI MS scan involving 10 laser shots, the instrument was post-calibrated using the signals from [PC (15:0/18:1(d7)) + H]$^+$ m/z 753.613, [LysoPC (18:1(d7)) + H]$^+$ m/z 529.399, and [SM (d18:1/18:1(d9)) + H]$^+$ m/z 738.647. The acquired signal was observed with an R.P. above 200,000 (FWHM) at their respective m/z and MMA better than 0.5 ppm (MMA after post-calibration). When the number of injected ions was increased (i.e., using 400 laser shots), it resulted in a 10× higher total ion count injected into the ICR cell and a mass shift for all experimentally observed m/z values resulting in an MMA between 2 and 8 ppm. The larger amount of ions introduced into the cell in regard to the calibration procedure severely impaired the MMA. An abrupt modification of the TIC intentionally generated by suddenly increasing the number of laser shots was correlated with the observed mass shift. An example is provided in Figure S1 for illustration based on the signal obtained for [PC (15:0/18:1(d7)) + H]$^+$.

From an MSI perspective, such TIC variations commonly appear when inhomogeneous matrix deposition creates hot spots, and/or when samples, such as tissue sections, have intrinsically heterogeneous regions in terms of molecular compositions and/or desorbed/ionized efficiencies.

3.2. Influence of Sample Preparation and the Amount of Sprayed Matrix on the Total Ion Current Stabilization. Avoiding the formation of hot spots due to inhomogeneous deposition of the matrix is important to produce MS images of high quality. The KPMP Consortium (Veličković et al.) and Tressler et al. improved MALDI MSI data after factorial design optimization of the deposited matrix using an automatic sprayer on mice's kidney tissue sections.[24,37] In the presented work, the amount of deposited MALDI matrix was investigated in terms of signal suppression for the analytes of interest when varying the number of laser shots per pixel (see Figure S2). For this purpose, serial sagittal mouse brain slices were prepared with varying amounts of sprayed α-HCCA matrix of 10 and 5 nmol per mm^2, respectively. MS images were acquired either using several hundreds of laser shots at 1000 Hz (Figure 1A) or using 6 laser shots at 60 Hz (Figure 1B). In the latter, the laser focus of the solariX XR 9.4T Smartbeam II laser was adjusted from 98 to 80% to ablate and desorb an equivalent amount of material per pixel. This allowed the generation of similar TIC values between the MALDI MSI methods. Figure 1A shows an unstable TIC for the method employing suboptimal parameters while the new set of optimized settings showed a drastically improved TIC stability through the entire acquisition (Figure 1B). This resulted in improved alignment of the m/z peaks keeping the mass shifts below 0.5 ppm while improving effective mass R.P. in the mean spectrum to around 400,000 at m/z 400 for all detected ions while keeping the same spatial resolution (Figure 1A,B panel c). An in-house script has been used to monitor the m/z channels for every pixel and compute the distribution of their measured values (apex of the m/z peaks) and the R.P. using FWHM. Figure S3 provides a graphical representation of the mass and R.P. distribution for m/z 782.5674 ([PE 39:4 + H]$^+$ or [PC 36:4 + H]$^+$ (according to the LIPID MAPS database) for both

https://doi.org/10.1021/acs.analchem.2c00754
Anal. Chem. 2022, 94, 9316−9326

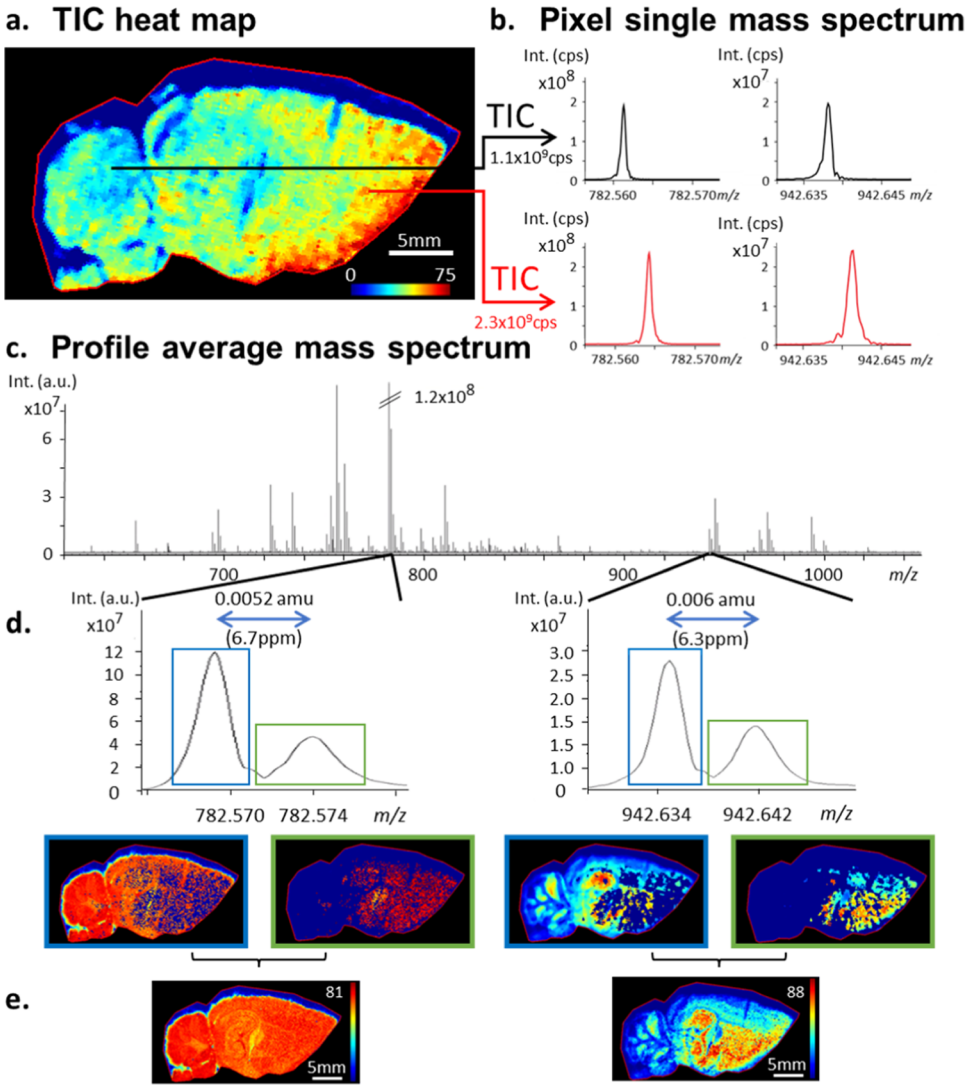


Figure 2. Heat map of the non-normalized TIC of a mouse brain section analyzed by high-resolution MALDI FT-ICR MSI on a solariX XR 9.4T using nonoptimized MSI method (a). Extracted mass spectra from single pixels located in regions with significant differences in total ion current (b). Average mass spectrum (mean spectrum) of the whole MALDI image (c). Zoomed profile average spectrum focused on m/z 782.57 and 942.64 showing artifacts of split peaks and their complementary distributions due to inconsistent mass measurement accuracy during acquisition (d). Obtained localizations with a window selection encompassing both m/z peaks shown in the vicinity of m/z 782.57 and 942.64, respectively (e).

methods which were drastically improved when the TIC fluctuation was restrained.

3.3. ICR Mass Analyzer Optimization and the Influence of Ion Optic Voltages on TIC Stability. First, the method employed was based on values recommended by the manufacturer for MALDI MSI. Minimal modifications were the use of the broadband mode in the 300−1200 amu mass range working at an estimated R.P. above 400.000 at m/z 800 with 4 M data point. The solariX XR 9.4T instrument was operated following a method optimized by Ferey et al.[15] They optimized the MALDI FT-ICR MSI parameters using experimental designs from a 12T magnet instrument fitted with the Paracell. However, our MSI acquisitions performed on our 9.4T magnet suffered from severe mass shifts as shown in Figure 2. For individual pixels of the image, the experimental R.P. was slightly above the one estimated by the FT-ICR control software (FTMS control). Nevertheless, the centroids of the m/z peaks shifted from pixel-to-pixel resulting in an MMA below the specification of the instrument as observed in the profile average mass spectrum of the image. The MSI profile average spectrum showed peak broadening due to the combination of pixels mass spectra, where a significant pixel-to-pixel mass shift of the measured m/z occurred. Extreme cases were observed, where the m/z peaks were splitting by a few milli amus (i.e., several ppm), as shown in Figure 2d. The reconstructed MS images of m/z 782.57 (assumed to be $[PC\ 34{:}1 + Na]^+$) and 942.64 (assumed to be $[CL\ 36{:}4 + NH_4]^+$), both selected with a mass tolerance of ±0.004, result in biased and incomplete ion distributions unless the targeted ions and their shifted counterparts m/z peak were selected together by extending the mass tolerance to ±0.01 for image reconstruction (Figure 2e). A comparison of the extracted spectra on a per-pixel basis (Figure 2a) showed that peak splitting could again be linked to the regions of interest submitted to large TIC variation despite the ion optics optimization adapted from Ferey et al. for our 9.4T instrument.

Ion optic voltages of the ICR mass analyzer were investigated as options to stabilize the TIC signal (i.e.,

https://doi.org/10.1021/acs.analchem.2c00754
Anal. Chem. 2022, 94, 9316−9326

charge-space effects) in our 9.4T FT-ICR at 4M. Thus, investigations were focused on ICR parameters, especially analyzer entrance, front and back trapping, sidekick, and excitation sweep voltages. Out of those parameters, the sidekick was the only parameter that had a slight influence on the TIC stability. The sidekick offset optimization also showed a limited effect on the mass shift, while the front and back trapping, as well as the excitation sweep, affected the peak shapes (as expected) but not the TIC stability. In contrast, the voltage applied to the analyzer entrance had an effect. Increasing the analyzer entrance voltage was followed by a gradual decrease until the absence of an MS signal. In the end, optimizing the ion optics does not significantly improve the TIC stability during MSI experiments.

3.4. Monitoring MALDI Processes and the Influence of the Number of Laser Shots. Laser-based ion sources in the commercially available MALDI MS instruments are typically using pulsed UV laser beams until kilohertz frequency. The main laser parameters accessible to the average user are the laser frequency, number of laser shots, laser power (usually in percent), and laser focus (how large is the area illuminated by the laser). The laser focus is usually constrained to keep the good lateral resolution of the MS images (i.e., density of pixels), and the laser power is set to the lowest value allowing the production of ions in any region of the imaged sample. Laser power also depends on how the laser is aging. Excluding the duration of the measurement of the transient ion frequencies, the duration of a scan to produce the mass spectrum of one pixel of the MS image is depending on the laser frequency and the number of laser shots set by the user. Because acquisitions with a single laser shot resulted in the absence of signals from both, the matrix and the tissue, 2 laser shots per scan were employed for the minimum number of laser shots per scan. Under this experimental condition, the laser frequency had to be lowered below 300 Hz once again due to an absence of signal above this threshold. Below the laser frequency of 300 Hz, no significant variations of the MS signal were observed (data not shown). To investigate the influence of the number of laser shots to produce an adequate number of ions resulting in high-quality mass spectra for MSI, the laser shots to laser frequency ratio was kept at 1:10, resulting in a constant laser shots step duration. This 1:10 ratio is effective in the 2−200 laser shots range because 2000 Hz is the operational limit of the SmartBeam II laser of the solariX XR and scimaX 2XR MALDI sources. By fixing the shooting duration and the time between the laser shots, we should avoid most of the kinetic relaxation influences and balance the potential biases due to the ion extraction from the MALDI plume by the ion optics. Beyond 200 laser shots, the maximal laser frequency would be used at the cost of the constant shooting step duration.

Figure 3a shows the TIC accumulation (summed TIC) of 2 laser shots at 20 Hz until the ions produced from 600 laser shots are collected, and Figure 3b shows the TIC value for each acquired individual scan. Figure 3a indicates that most of the accumulated TIC signal (more than one-half) was obtained from the first 100 laser shots and that initial laser shots produced a rather linear increase in TIC, followed by a smaller number of ions produced by subsequent laser shots. Then, a further linear increase is observed after about 120 laser shots due to the accumulation of mainly noise peaks. Figure 3b points out a notable instability of the TIC during the first 50 laser shots. The very first shots showed the highest signal

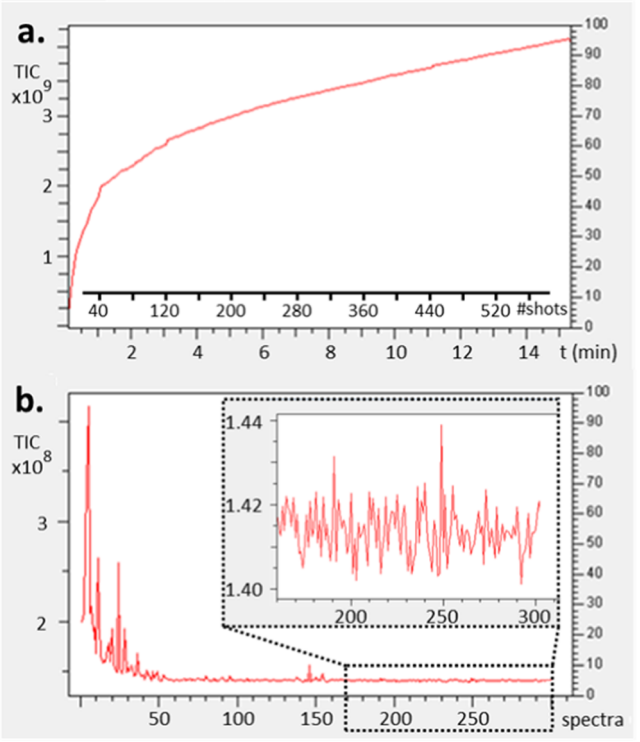


Figure 3. Accumulated total ion current over time (a) and total ion current per scan (b) for a 300 scans acquisition with a setting of 2 laser shots at 20 Hz per scan performed on a dynamically harmonized MALDI FT-ICR solariX XR 9.4T.

abundance in the mass spectrum with a relatively high relative abundance (>10% relative intensity), which aligns with the so-called "first-shot phenomenon" first described by the team of Hillenkamp.[38] The following laser shots, still ablating the same (x,y) position, only poorly contributed to good signal-to-noise ratios for interesting m/z values and the less abundant m/z peaks vanished first. These results suggest that a smaller amount of laser shots is beneficial for the detection of ions with an appropriate signal-to-noise ratio unless the targeted ions require a significantly larger amount of laser energy to be detected. Thus, using fewer laser shots, TIC fluctuations will be minimized to only a small percentage ensuring a more controlled number of ions to be injected into the ICR cell. This leads to constant space-charge effects resulting in the production of ultrahigh mass R.P. MALDI images. The contribution of the laser shots and the desorption/ionization steps of each pixel being imaged in terms of duration is typically less than 1 s from 2 to 10 laser shots when operating the laser shots to laser frequency at a 1:10 constant ratio.

3.5. Influence of the Number of Laser Shots on the Apparent Dynamic Range of Imaged Lipids. New sets of mice brain images were produced by MALDI MSI. The number of laser shots ranged from 10 to 600 at a fixed laser-shooting time duration (i.e., laser shots to laser frequency ratio). As expected, lowering the number of laser shots (from 600 to 10) reduced the overall signal intensities in the mass spectra (TIC) although the detected ions for both methods were comparable. The loss of m/z signals in the method using the lower amount of laser shots was mainly concerning the isotope contributions and peaks that were already close to the 3 × S/N (signal over noise) as computed by the software. Interestingly, the absolute intensities of the minor ions were

https://doi.org/10.1021/acs.analchem.2c00754
Anal. Chem. 2022, 94, 9316−9326

Table 2. Intensities and Ratios of Detected and Identified Lipids in a Mouse Brain Tissue Section Acquired with the MALDI FT-ICR MS (solariX XR 9.4T) Instrument for 10 and 100 Laser Shots

		10 shots				100 shots			
target m/z	identification	intensity (c.p.s)	mass accuracy (ppm)	ratio m/z 772.53 over m/z target	ratio m/z 798.54 over m/z target	intensity (c.p.s)	mass accuracy (ppm)	ratio m/z 772.53 over m/z target	ratio m/z 798.54 over m/z target
770.50975	$[PA\ 36:2+K]^+$	4.8×10^5	−0.09	10.6	11.3	1.3×10^6	−0.12	33.1	40.0
848.55643	$[PC\ 38:4+K]^+$	8.9×10^5	−0.25	5.73	6.07	6.5×10^6	−0.24	6.61	8.00
772.52519	$[PC\ 32:0+K]^+$	5.1×10^6	−0.13		1.06	4.3×10^7	−0.17		1.21
798.54079	$[PC\ 34:1+K]^+$	5.4×10^6	−0.26	0.94		5.2×10^7	−0.26	0.83	

Table 3. Intensities and Ratio of Detected and Identified Lipids in a Mouse Brain Tissue Section Acquired with the MALDI-ToF MS (rapifleX) Instrument (External Calibration, Enhanced Cubic Regression) for 10 and 1000 Laser Shots with the Single Focus Option and Without Beamscan

		10 shots				1000 shots			
target m/z	identification	intensity (c.p.s)	mass accuracy (ppm)	ratio m/z 782.57 over m/z target	ratio m/z 798.54 over m/z target	Intensity (c.p.s)	Mass accuracy (ppm)	Ratio m/z 782.57 over m/z target	Ratio m/z 798.54 over m/z target
782.567	$[PC\ 36:4 + H]^+$	6.8×10^3	+9.71		0.79	1.8×10^4	+7.1		0.81
798.541	$[PC\ 34:1 + K]^+$	5.3×10^3	+11.2	1.27		1.5×10^4	+10.0	1.23	
806.567	$[PS\ 37:0 + K]^+$	4.0×10^3	−12.3	1.68	1.33	9.7×10^3	−11.1	1.90	1.54
844.546	$[PC\ 36:3 + Na]^+$	4.6×10^3	+3.8	1.47	1.16	1.5×10^4	6.2	1.22	0.99

almost not affected compared to the most abundant ones when using fewer laser shots and 5 nmol·mm^{-2} of deposited matrix on mice's brain tissue sections. Table 2 reports the absolute intensity, mass accuracy, and intensity ratio between high and low abundant lipids detected in the MS images of mice's brain tissue section when using 10 or 100 laser shots. The lipids were identified according to the LIPID MAPS database peak annotation. Ions at m/z 772.53 and 798.54 were the most intense signals observed, while m/z 770.51 and 848.56 are among the least intense ions. When comparing 10−100 laser shots, the intensities of minor ions were roughly halved while major ions' intensities decreased by an order of magnitude. By reducing the number of laser shots per scan, the relative intensities of the most intense ions tend to decrease to a larger extent in regard to the less intense ions. Any combination of intense/less intense ion ratios leads to the same observations. Besides, the intensity ratio between ions of comparable intensities (e.g., m/z 772.53 vs 798.54) was almost unaffected by the number of laser shots per scan.

The cause of the disparity in the ion intensity ratio when varying the number of laser shots was further investigated. It could indeed be related to either the ionization process itself or the efficiency of ion transmission by the ion optics and/or the MS analyzer (ICR cell). Similar experiments to determine the influence of the ionization process were conducted on a MALDI-ToF instrument (rapifleX, Bruker, Germany) despite its differences in terms of ion extraction mechanism, source vacuum, and laser compared to the solariX XR and scimaX 2XR. To be somehow comparable with the Smartbeam II, the beamscan option of the Smartbeam 3D was not used, which avoids the laser energy being swept at the surface of the sample (i.e., matrix blaster). No variation of the ion intensity ratio was observed with the MALDI-ToF, as shown in Table 3, despite we used a maximum of 1000 laser shots accumulation instead of 100, regardless of the major or minor ions considered. The ablated surface of the sample of only 25 μm^2 with only 10 laser shots allowed the less abundant ions still to be detected, and only matrix signal intensities were strongly affected. The variation of the ion ratio observed with the MALDI FT-ICR was then related to the ion optics and/or the ICR mass analyzer. It is worth reminding that higher magnetic fields improve the dynamic range of the number of trapped ions inside the ICR in the absence of noticeable space-charge effects.

3.6. Robustness of the MALDI MSI at Ultrahigh Mass R.P. Using the MALDI FT-ICR solariX XR 9.4T. When applying the optimized method for extreme-resolution MSI with an estimated R.P. over 500,000 at m/z 800 (>1.000.000 at m/z 400), a much less pronounced mass shift was observed and most importantly is fully within the instrument specifications (±0.5 ppm) even for profile and centroided average MALDI image spectra. FWHM resolutions for the individual spectra are now similar to the ones observed in the profile average spectrum. The acquisition of such mass spectra qualities during an MSI experiment of rat brain samples was previously achieved using a custom prototype of a hybrid linear ion trap coupled to a 21T supra conducting magnet fitted with the Paracell by the NHMFL group at Tallahassee in Florida.[39] In this work, comparable results were obtained in terms of MSI mass R.P. and MMA with a superconducting magnet of 9.4T using the same ICR cell (see Table S1). By comparing the average spectrum of centroided MSI data performed on seriated brain (Figure 4a) and whole-body zebrafish sections (Figure 4b), the difference in image quality is clearly evident. The new method with a controlled ion injection in the cell (TIC stabilized) resulted in narrow m/z peaks due to a significantly reduced mass shift. The image of the ion distribution in the tissue section is also less noisy whether or not TIC- or RMS-normalized (RMS not shown). To demonstrate the robustness of the controlled TIC method, replicates of serial brain sections (roughly 12,000 pixels) and zebrafish whole-body sections (roughly 20,000 pixels) were acquired using our optimized method (see Figure S5). In all cases, the experimental mass R.P. expected by the acquisition software was surpassed. The imaging method was tested for images with R.P.$_{\cdot FWHM}$ beyond 1,000,000 at m/z 800 for the brain region of the zebrafish sample (roughly 1500 pixels). Figure S6 shows observable isotopic fine structures for

https://doi.org/10.1021/acs.analchem.2c00754
Anal. Chem. 2022, 94, 9316−9326

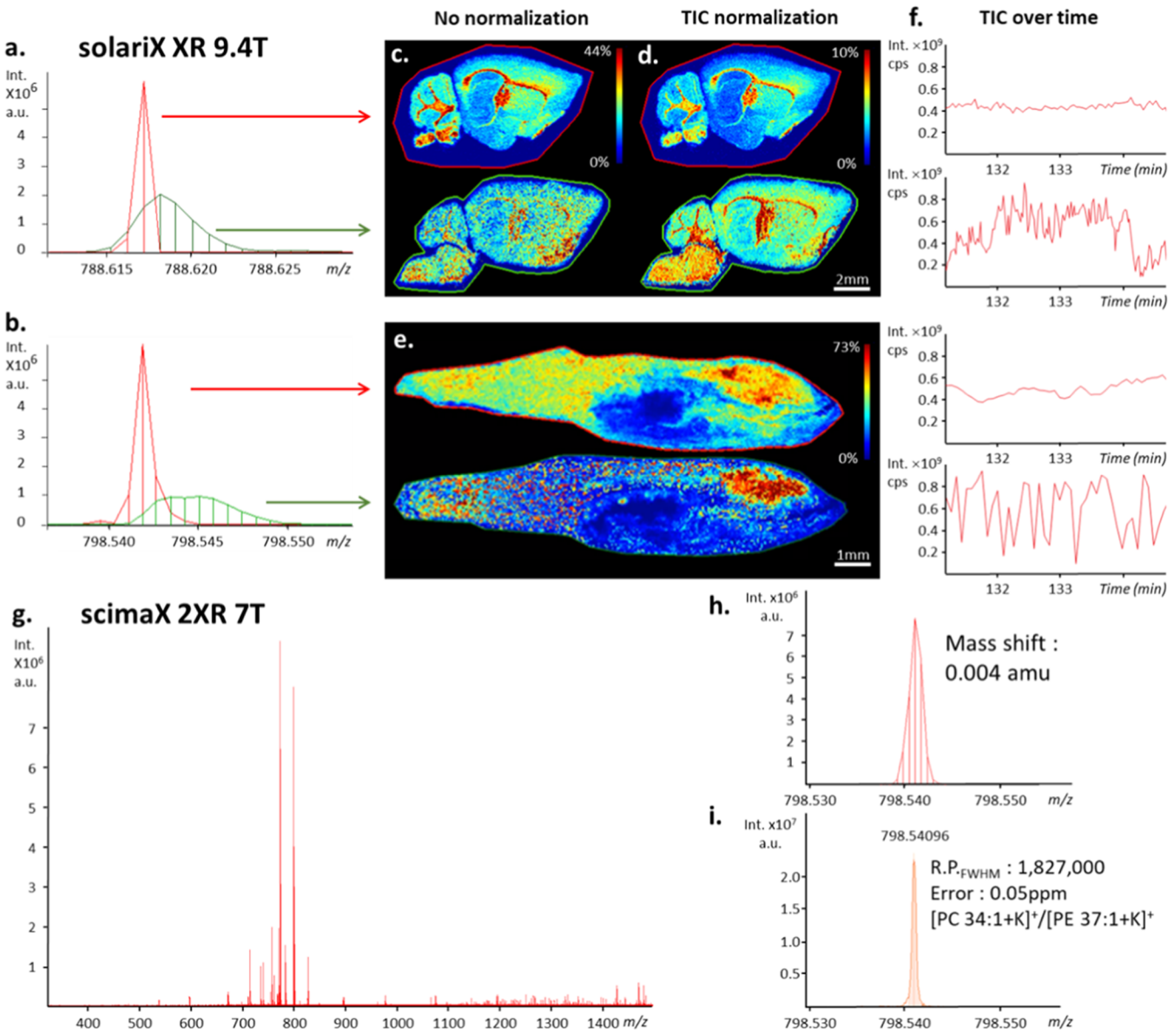


Figure 4. Centroided average MSI mass spectra of mouse brain (a) and zebrafish (b) using a MALDI FT-ICR (solariX XR 9.4T, fitted with ParaCell) instrument, zoomed in on m/z 788.62 and 798.54, respectively, showing peak width differences due to MMA obtained with (red) and without (green) TIC stabilization by optimization of the laser shot number (6 laser shots at 60 Hz). Reconstructed brain slices MS images without (c) and with (d) TIC normalization. Reconstructed zebrafish whole-body slices MS images with TIC normalization (e, TIC stabilization method (up) and no TIC stabilization (down), respectively). Section of the TIC over time to show the TIC stability of the presented MSI (f). Centroided average MSI mass spectrum of a mouse brain tissue section acquired on the scimaX 2XR using the 2ω detection mode with TIC stabilization (g). Zoom in to m/z 798.54 showing peak width difference due to MMA (h). Zoom in on m/z 798.54 in an extracted pixel spectrum showing the obtained R.P. FWHM and MMA (i). See text for details.

abundant ions also observable in the profile average spectrum further increasing the confidence of the identification process of these ions. Note that a slight loss in R.P. is still observed in the profile average spectrum compared to individual pixels spectra even with a contained mass shift below 0.5 ppm (i.e., in agreement with the instrument specification). At such high R.P., the contribution of a mass shift of 0.5 ppm at m/z 800 (i.e., 0.4 mamu) is still impacting negatively the R.P. In complement, a peak realignment strategy by software postprocessing coupled with our proposed MSI method was developed in our group to restore the isotopic fine structure also in the average mass spectrum of MS images having lower mass R.P..[40,41]

3.7. Influence of Magnetic Field Strength. FT-ICRs operating at a high magnetic field (≥12T) are typically used for petroleomics analysis by direct infusion electrospray ionization analysis of diluted raw crude oils. Direct infusion allows for a stable TIC signal and fills the ICR cell with a constant number of ions at each scan. Higher magnetic fields allow the storage of a larger number of ions and also a stable ion motion during long transient signal acquisition enabling very high mass R.P. (≥1,000,000 at m/z 400). Recently, Ge et al.[42] demonstrated for oil samples introduced by direct infusion the capability of FT-ICR mass analyzers operating at 7T and 2ω detection to closely match the performance of a 15T instrument. MSI also takes advantage of greater magnetic

https://doi.org/10.1021/acs.analchem.2c00754
Anal. Chem. 2022, 94, 9316−9326

Table 4. Intensities and Ratio of Detected and Identified Lipids in a Mouse Brain Tissue Section Acquired with the MALDI FT-ICR-MS (scimaX 2XR 7T) Instrument for 6 and 400 Laser Shots[a,b]

		6 laser shots				400 laser shots			
target m/z	identification	intensity (c.p.s)	mass accuracy (ppm)	ratio m/z (a) over m/z target	ratio m/z (b) over m/z target	intensity (c.p.s)	mass accuracy (ppm)	ratio m/z (a) over m/z target	ratio m/z (b) over m/z target
scimaX 2XR 7T, 1ω detection mode									
713.45181	[PA 34:1+K]+	8.0×10^5	+0.40	6.63	30.0	2.7×10^6	+0.61	7.04	18.5
844.52531	[PC 38:6+K]+	2.4×10^5	+0.53	22.1	100	not detected	+0.58	N.C.	N.C.
772.52519	[PC 32:0+K]+	5.3×10^6	+0.37		4.53	1.9×10^7	+0.56		2.63
798.54079	[PC 34:1+K]+	2.4×10^7	+0.26	0.22		5.0×10^7	+0.44	0.38	
scimaX 2XR 7T, 2ω detection mode									
713.45181	[PA 34:1+K]+	4.3×10^6	+0.23	3.72	4.88	4.9×10^6	+0.40	4.29	10.4
844.52531	[PC 38:6+K]+	3.9×10^6	+0.17	4.10	5.38	4.3×10^6	+0.33	3.13	7.61
772.52519	[PC 32:0+K]+	1.6×10^7	+0.22		1.31	2.1×10^7	+0.37		2.43
798.54079	[PC 34:1+K]+	2.1×10^7	+0.26	0.76		6.7×10^7	+0.31	0.41	

[a] m/z (a) corresponds to 772.53, and m/z (b) corresponds to 798.54. [b] N.C. Not Computed.

fields to produce higher-quality images.[15,39] In this work, the influence of TIC variation was investigated for an FT-ICR instrument equipped with a ParaCell but using lower magnetic field strength, i.e., 7T. The scimaX 2XR 7T instrument also provides the 2ω detection mode, which recycles the excitation plates into detection plates to improve the duty cycle (transient signal) by a factor of 2 compared to the 1ω detection mode. Mice brain images were compared for both 1ω and 2ω detection. Also, for this instrument, TIC control improved the MMA and spectral resolutions (see Figures 4f–h and S7). Interestingly, while no direct influence on ion current stability was observed when the 2ω detection mode was activated (Figure S8a), the mass shift was easier to constrain compared to 1ω mode datasets (Figure S8b). We assumed that the drastic diminution of the transient signal duration prevented peak coalescence as well as the decoherence of the ion packets inside the ICR cell.[8,43] It is worth mentioning that the comparison of MS images performed on serial tissue sections using 1ω and 2ω detection mode showed no tangible differences in terms of co-localization of the observed ions (Figure S8b). As an illustration, an MS image acquired with the solariX 9.4T and the scimaX 2XR 7T using the 2ω detection mode at 16M data points of 1000 pixels of mouse brain tissue section showed an $R.P._{FWHM}$ above 1,500,000 at m/z 800 and an MMA of 0.15 ppm (Figure 4f–h). The typical time to produce an image of 1000 pixels was 195 min using our solariX XR 9.4T (i.e., 1ω at 8M data point) and 205 min using the scimaX 2XR 7T operating at 2ω and 16 M data point.

Table 4 shows the dynamic range obtained for high and low abundant lipid signals detected during the MSI experiment of 2 consecutive brain sections with the scimaX 2XR using 1 or 2ω detection mode. Note that the 7T instrument still required the ICR cell to be loaded with fewer ions than the 9.4T instrument using 1ω or 2ω detection to restrict the experimental mass shift in average mass spectra. Therefore, fewer laser shots were used to produce the data in Table 4 compared to Tables 2 and 3. The ratios obtained for the scimaX 2XR using 1ω detection are somewhat similar to that obtained with the 9.4T solariX XR. The 2ω detection mode seems to be also beneficial because the dynamic range of the lipids detected in the MALDI images was less affected than the 1ω detection mode. The higher power of the magnet is still beneficial for reaching a wider intrascan dynamic range, or if the TIC cannot be efficiently stabilized, even after optimizing the laser parameters. Nevertheless, MSI acquisition at extreme mass R.P. is possible using the 7T superconducting magnet and 2ω detection.

3.8. Improvement of Peak Annotation. Finally, database queries for mass lists obtained with the optimized method showed an improvement in terms of peak annotations: fewer false positives and negatives were observed due to the improved R.P. and MMA. Figure S9 shows some examples of database results as histograms of the matching counts at a given mass accuracy (in ppm) to detect readily any oddities in the dataset. When the TIC was not stabilized (not optimized method), most of the identifications had mass accuracy around −1.5 ppm. These values were not consistent with the specification of a properly calibrated FT-ICR instrument and they do not fit with the requirement for proper annotation of lipids from the LIPID MAPS database. In contrast, the MALDI image acquired with the optimized method and stabilized TIC led to a larger number of identifications, with scores around +0.4 ppm that are well within the nominal performance for the instrument. The number of total matches is drastically improved due to fewer false-negative identifications and, similarly, potentially fewer false-positive results. Of course, the addition of the isotopic fine structure further improved the confidence level of the identified lipids.

4. CONCLUSIONS

In this work, we successfully limited the space-charge effects and limited the resulting mass shift to improve the mass accuracy for MALDI MS images of mouse brains and Zebrafish tissue sections by introducing a controlled TIC injection method in the ICR cell. The method was successfully applied on the solariX XR 9.4T and the scimaX 2XR 7T, two commercially available dual-source ESI/MALDI instruments fitted with the Paracell. Under optimal instrumental settings, this was achieved primarily by optimizing laser parameters and the concentration of deposited/sprayed matrix. MSI with a resolving mass power beyond 1,000,000 at m/z 800 was successfully achieved within around 200 min for 1000 imaged pixels (transient duration of 11.7 s at 8M data points in an operated mass range between m/z 300 and 1200 using the common Amplitude mode for the solariX XR 9.4T, and a transient of 12.3 s for the scimaX 2XR 7T in 2ω detection mode at 16 M data points) with no mass shift beyond 1 ppm

https://doi.org/10.1021/acs.analchem.2c00754
Anal. Chem. 2022, 94, 9316−9326

(typical mass shift <0.5 ppm), which correlates to approximately 0.5 mamu in the lipid mass range. The resulting images appeared sharper and showed improved contrast at constant lateral resolution and matrix deposition method. Extreme-resolution MS images obtained with relatively limited power of magnetic fields (<12T) require a stabilized TIC throughout the acquisition to retain the instrument specifications. The intrascan dynamic range obtained during this work using the commercially available 9.4T and 7T instruments seemed to be around 100, while Bowman et al.[39] reported a dynamic range of around 500 using a custom 21T MALDI FT-ICR instrument. Using 2ω detection on higher-magnetic-field instruments will speed up the scan time by a factor of 2, allowing more samples to be measured at constant mass R.P. in the same time frame. Peak annotations using the LIPID MAPS database correspond to identification scores better than 0.4 ppm, limiting misidentification of lipids, especially for measurement generating isotopic fine structures. Revisiting the laser parameters improved method reproducibility from pixel to pixel and also from sample to sample, which improved the robustness of our method by successfully performing similar MALDI images of consecutive tissue sections in replicates.

It was found that the entrance voltage to the analyzer affects the number of ions introduced into the ICR cell in an interesting way that could potentially be used to limit the overflow of ions to be injected into the ICR cell, acting as an ion injection control device. The idea would be to limit ion current fluctuations in real time for samples with high-concentration heterogeneity of target compounds. This would require further investigation as it is currently considered a double-edged sword, as the signal can be easily lost if this voltage value is not properly set.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.2c00754.

Trace of the TIC obtained from a deuterated lipid using two different laser parameters; comparison of the profile average MSI spectra at different laser settings; distribution of an ion mass-to-charge ratio measurement and FWHM resolution distributions at different laser settings; co-localization of one of the lipid during MSI acquisitions on same tissue but different laser settings; additional MSI replicate on brain tissues section and Zebrafish tissue sections; lipid identification using the isotope fine structure of one of the ions detected during a MALDI mass spectrometry image at mass resolving power better than 1 million; traces of the TIC and influence of TIC stability for a scimaX 7T operating in 1ω detection mode; performance of the scimaX 7T using 1 and 2ω detection mode before and after optimization; distribution of the matching exact masses with the LIPID MAPS structure database before and after method optimization; MSI performances of FT-ICR fitted with dynamically harmonized cells and using different superconducting magnets of 7T, 9.4T, and 21T instruments with 1 or 2ω detection mode; and plot representing the constant laser shot-to-laser frequency ratio with the graphical representation of the limitation of the operational laser settings (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Johann Far − *Mass Spectrometry Laboratory, MolSys Research Unit, University of Liège, 4000 Liège, Belgium;* orcid.org/0000-0003-1208-6262; Email: johann.far@uliege.be

Authors

Mathieu Tiquet − *Mass Spectrometry Laboratory, MolSys Research Unit, University of Liège, 4000 Liège, Belgium;* orcid.org/0000-0002-9873-0065

Raphaël La Rocca − *Mass Spectrometry Laboratory, MolSys Research Unit, University of Liège, 4000 Liège, Belgium;* orcid.org/0000-0003-3388-3041

Stefan Kirnbauer − *Institute of Chemical Technologies and Analytics, TU Wien (Vienna University of Technology), 1060 Vienna, Austria; Austrian Cluster for Tissue Regeneration, TU Wien (Vienna University of Technology), 1060 Vienna, Austria*

Samuele Zoratto − *Institute of Chemical Technologies and Analytics, TU Wien (Vienna University of Technology), 1060 Vienna, Austria; Austrian Cluster for Tissue Regeneration and Christian Doppler Laboratory for Skin Multimodal Imaging of Aging and Senescence, TU Wien (Vienna University of Technology), 1060 Vienna, Austria*

Daan Van Kruining − *Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University, 6229ER Maastricht, the Netherlands*

Loïc Quinton − *Mass Spectrometry Laboratory, MolSys Research Unit, University of Liège, 4000 Liège, Belgium;* orcid.org/0000-0001-8153-9590

Gauthier Eppe − *Mass Spectrometry Laboratory, MolSys Research Unit, University of Liège, 4000 Liège, Belgium;* orcid.org/0000-0002-4821-3115

Pilar Martinez-Martinez − *Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University, 6229ER Maastricht, the Netherlands*

Martina Marchetti-Deschmann − *Institute of Chemical Technologies and Analytics, TU Wien (Vienna University of Technology), 1060 Vienna, Austria; Austrian Cluster for Tissue Regeneration and Christian Doppler Laboratory for Skin Multimodal Imaging of Aging and Senescence, TU Wien (Vienna University of Technology), 1060 Vienna, Austria;* orcid.org/0000-0002-8060-7851

Edwin De Pauw − *Mass Spectrometry Laboratory, MolSys Research Unit, University of Liège, 4000 Liège, Belgium;* orcid.org/0000-0003-3475-1315

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.2c00754

Funding

Open Access is funded by the Austrian Science Fund (FWF).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Eurlipids Interreg V-A Euregio Meuse-Rhine with support from the European Fund for Regional Development of the European Union (www.eurlipids.com). This work was also a contribution to the EU Horizon 2020 research and Innovation program under grant

agreement no. 731077 and the European Project EU_FT-ICR_MS (H2020 INFRAIA-02-2017). The MALDI-ToF rapifleX and the MALDI FT-ICR solariX XR were co-funded by FEDER BIOMED HUB Technology Support (number 2.2.1/996). This work was also a contribution to ZonMw Memorabel program (number: 733050105), Cost action under grant CA16122 − BIONECA and Hersenstichting (number: DR-2018-00274). The scimaX 7T 2XR was funded by TU Wien and the Federal Ministery Republic of Austria for Education, Science and Research (HRSM2016). This work was also a contribution to the Christian Doppler Laboratory for Skin Multimodal Imaging of Aging and Senescence (SkinMAGINE) and the Austrian Science Foundation project ChemTalk (P32179-B). All experiments were performed with permission from the Committee on Animal Welfare of Maastricht University and the Committee on Animal Welfare of Liege University, according to Dutch or Belgian governmental legislation, respectively.

REFERENCES

(1) Le Rhun, E.; Duhamel, M.; Wisztorski, M.; Gimeno, J.-P.; Zairi, F.; Escande, F.; Reyns, N.; Kobeissy, F.; Maurage, C.-A.; Salzet, M.; Fournier, I. *Biochim. Biophys. Acta, Proteins Proteomics* **2017**, *1865*, 875−890.
(2) Ellis, S. R.; Cappell, J.; Potočnik, N. O.; Balluff, B.; Hamaide, J.; Van der Linden, A.; Heeren, R. M. A. *Analyst* **2016**, *141*, 3832−3841.
(3) Lamont, L.; Eijkel, G. B.; Jones, E. A.; Flinders, B.; Ellis, S. R.; Porta Siegel, T.; Heeren, R. M. A.; Vreeken, R. J. *Anal. Chem.* **2018**, *90*, 13229−13235.
(4) Wang, J.; Wang, C.; Han, X. *Anal. Chim. Acta* **2019**, *1061*, 28−41.
(5) Marshall, A. G.; Hendrickson, C. L. *Annual Rev. Anal. Chem.* **2008**, *1*, 579−599.
(6) Kostyukevich, Y. I.; Vladimirov, G. N.; Nikolaev, E. N. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 2198−2207.
(7) Nikolaev, E. N.; Kostyukevich, Y. I.; Vladimirov, G. N. *Mass Spectrom. Rev.* **2016**, *35*, 219−258.
(8) Boldin, I. A.; Nikolaev, E. N. *Rapid Commun. Mass Spectrom.* **2011**, *25*, 122−126.
(9) Jertz, R.; Friedrich, J.; Kriete, C.; Nikolaev, E. N.; Baykut, G. *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 1349−1366.
(10) Popov, I. A.; Nagornov, K.; N Vladimirov, G.; Kostyukevich, Y. I.; Nikolaev, E. N. *J. Am. Soc. Mass Spectrom.* **2014**, *25*, 790−799.
(11) Nikolaev, E. N.; Gorshkov, M. V.; Mordehai, A. V.; Talrose, V. L. *Rapid Commun. Mass Spectrom.* **1990**, *4*, 144−146.
(12) Pan, Y.; Ridge, D. P.; Rockwood, A. L. *Int. J. Mass Spectrom. Ion Processes* **1988**, *84*, 293−304.
(13) Schweikhard, L. *Int. J. Mass Spectrom. Ion Processes* **1991**, *107*, 281−292.
(14) Pan, Y.; Ridge, D. P.; Wronka, J.; Rockwood, A. L.; Marshall, A. G. *Rapid Commun. Mass Spectrom.* **1987**, *1*, 120−121.
(15) Ferey, J.; Marguet, F.; Laquerrière, A.; Marret, S.; Schmitz-Afonso, I.; Bekri, S.; Afonso, C.; Tebani, A. *Anal. Bioanal. Chem.* **2019**, *411*, 3891−3903.
(16) Longuespée, R.; Kriegsmann, K.; Cremer, M.; Zgorzelski, C.; Casadonte, R.; Kazdal, D.; Kriegsmann, J.; Weichert, W.; Schwamborn, K.; Fresnais, M.; Schirmacher, P.; Kriegsmann, M. *Proteomics: Clin. Appl.* **2019**, *13*, No. 1800074.
(17) Goodwin, R. J. A. *J. Proteomics* **2012**, *75*, 4893−4911.
(18) Shimma, S.; Sugiura, Y. *Mass Spectrom.* **2014**, *3*, S0029.
(19) Morikawa-Ichinose, T.; Fujimura, Y.; Murayama, F.; Yamazaki, Y.; Yamamoto, T.; Wariishi, H.; Miura, D. *J. Am. Soc. Mass Spectrom.* **2019**, *30*, 1512−1520.
(20) Kaletaş, B. K.; van der Wiel, I. M.; Stauber, J.; Lennard, J. D.; Güzel, C.; Kros, J. M.; Luider, T. M.; Heeren, R. M. A. *Proteomics* **2009**, *9*, 2622−2633.
(21) Grassl, J.; Taylor, N. L.; Millar, A. *Plant Methods* **2011**, *7*, 21.
(22) Nishidate, M.; Hayashi, M.; Aikawa, H.; Tanaka, K.; Nakada, N.; Miura, S.; Ryu, S.; Higashi, T.; Ikarashi, Y.; Fujiwara, Y.; Hamada, A. *Drug Metab. Pharmocokinet.* **2019**, *34*, 209−216.
(23) Prideaux, B.; Stoeckli, M. *J. Proteomics* **2012**, *75*, 4999−5013.
(24) Tressler, C.; Tilley, S.; Yang, E.; Donohue, C.; Barton, E.; Creissen, A.; Glunde, K. *J. Am. Soc. Mass Spectrom.* **2021**, *32*, 2728−2737.
(25) Römpp, A.; Guenther, S.; Schober, Y.; Schulz, O.; Takats, Z.; Kummer, W.; Spengler, B. *Angew. Chem., Int. Ed.* **2010**, *49*, 3834−3838.
(26) Barry, J. A.; Robichaud, G.; Muddiman, D. C. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 1137−1145.
(27) Smith, D. F.; Kharchenko, A.; Konijnenburg, M.; Klinkert, I.; Paša-Tolić, L.; Heeren, R. M. A. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 1865−1872.
(28) Youmans, K. L.; Tai, L. M.; Nwabuisi-Heath, E.; Jungbauer, L.; Kanekiyo, T.; Gan, M.; Kim, J.; Eimer, W. A.; Estus, S.; Rebeck, G. W.; Weeber, E. J.; Bu, G.; Yu, C.; LaDu, M. J. *J. Biol. Chem.* **2012**, *287*, 41774−41786.
(29) Oakley, H.; Cole, S. L.; Logan, S.; Maus, E.; Shao, P.; Craft, J.; Guillozet-Bongaarts, A.; Ohno, M.; Disterhoft, J.; Van Eldik, L.; Berry, R.; Vassar, R. *J. Neurosci.* **2006**, *26*, 10129−10140.
(30) Goodwin, R. J. A.; Pennington, S. R.; Pitt, A. R. *Proteomics* **2008**, *8*, 3785−3800.
(31) Sládková, K.; Houška, J.; Havel, J. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 3114−3118.
(32) Fahy, E.; Sud, M.; Cotter, D.; Subramaniam, S. *Nucleic Acids Res.* **2007**, *35*, W606−W612.
(33) Sud, M.; Fahy, E.; Cotter, D.; Brown, A.; Dennis, E. A.; Glass, C. K.; Merrill, A. H.; Murphy, R. C.; Raetz, C. R. H.; Russell, D. W.; Subramaniam, S. *Nucleic Acids Res.* **2007**, *35*, D527−D532.
(34) Fahy, E.; Subramaniam, S.; Brown, H. A.; Glass, C. K.; Merrill, A. H.; Murphy, R. C.; Raetz, C. R. H.; Russell, D. W.; Seyama, Y.; Shaw, W.; Shimizu, T.; Spener, F.; van Meer, G.; VanNieuwenhze, M. S.; White, S. H.; Witztum, J. L.; Dennis, E. A. *J. Lipid Res.* **2005**, *46*, 839−861.
(35) Francl, T. J.; Sherman, M. G.; Hunter, R. L.; Locke, M. J.; Bowers, W. D.; McIver, R. T. *Int. J. Mass Spectrom. Ion Processes* **1983**, *54*, 189−199.
(36) Easterling, M. L.; Mize, T. H.; Amster, I. J. *Anal. Chem.* **1999**, *71*, 624−632.
(37) Veličković, D.; Zhang, G.; Bezbradica, D.; Bhattacharjee, A.; Paša-Tolić, L.; Sharma, K.; Alexandrov, T.; Anderton, C. R. *J. Am. Soc. Mass Spectrom.* **2020**, *31*, 508−516.
(38) Horneffer, V.; Strupat, K.; Hillenkamp, F. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 1599−1604.
(39) Bowman, A. P.; Blakney, G. T.; Hendrickson, C. L.; Ellis, S. R.; Heeren, R. M. A.; Smith, D. F. *Anal. Chem.* **2020**, *92*, 3133−3142.
(40) La Rocca, R.; Kune, C.; Tiquet, M.; Stuart, L.; Eppe, G.; Alexandrov, T.; De Pauw, E.; Quinton, L. *Anal. Chem.* **2021**, *93*, 4066−4074.
(41) McCann, A.; Rappe, S.; La Rocca, R.; Tiquet, M.; Quinton, L.; Eppe, G.; Far, J.; De Pauw, E.; Kune, C. *Anal. Bioanal. Chem.* **2021**, *413*, 2831−2844.
(42) Ge, J.; Ma, C.; Qi, Y.; Wang, X.; Wang, W.; Hu, M.; Hu, Q.; Yi, Y.; Shi, D.; Yue, F.; Li, S.; Volmer, D. A. *Anal. Sci. Adv.* **2021**, *2*, 272−278.
(43) van Agthoven, M. A.; Lam, Y. P. Y.; O'Connor, P. B.; Rolando, C.; Delsuc, M.-A. *Eur. Biophys. J.* **2019**, *48*, 213−229.

https://doi.org/10.1021/acs.analchem.2c00754
Anal. Chem. 2022, 94, 9316−9326

Associated papers of Chapter 4

A4- Original publication of “Rapid Visualization of Chemically Related Compounds Using Kendrick Mass Defect As a Filter in Mass Spectrometry Imaging.”

analytical chemistry

Article

Cite This: *Anal. Chem.* 2019, 91, 13112–13118

pubs.acs.org/ac

Rapid Visualization of Chemically Related Compounds Using Kendrick Mass Defect As a Filter in Mass Spectrometry Imaging

Christopher Kune,[†,∥] Andréa McCann,[†,∥] La Rocca Raphaël,[†] Anthony Arguelles Arias,[‡] Mathieu Tiquet,[†] Daan Van Kruining,[§] Pilar Martinez Martinez,[§] Marc Ongena,[‡] Gauthier Eppe,[†] Loïc Quinton,[†] Johann Far,[†] and Edwin De Pauw*[,†]

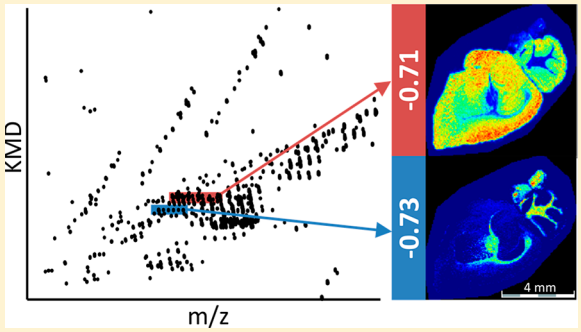
[†]Mass Spectrometry Laboratory, MolSys Research Unit, University of Liège, Liège, Belgium

[‡]Microbial Processes and Interactions, Gembloux Agro-Bio Tech, TERRA Teaching and Research Centre, University of Liege, Gembloux, Belgium

[§]Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University, Maastricht, The Netherlands

Supporting Information

ABSTRACT: Kendrick mass defect (KMD) analysis is widely used for helping the detection and identification of chemically related compounds based on exact mass measurements. We report here the use of KMD as a criterion for filtering complex mass spectrometry data set. The method allow automated, easy and efficient data processing, enabling the reconstruction of 2D distributions of families of homologous compounds from MSI images. We show that KMD filtering, based on in-house software, is suitable and robust for high resolution (full width at half-maximum, fwhm, at m/z 410 of 20 000) and very high-resolution (fwhm, at m/z 410 of 160 000) MSI data. This method has been successfully applied to two different types of samples, bacteria cocultures, and brain tissue sections.



Downloaded via UNIV DE LIEGE on July 18, 2025 at 14:05:11 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

Identification of analytes in complex mixtures is still a major challenge in analytical chemistry.[1] For that purpose, mass spectrometers are among the most widely used instruments. They are commonly coupled to direct infusion or hyphenated with separation techniques, for example, liquid chromatography (LC), capillary electrophoresis (CE), ion mobility spectrometry (IMS). Mass spectrometry imaging (MSI) is a particular case in which 2D distribution of analytes can be reconstructed after recording full mass spectra at the coordinate of the "pixels" sampled by a LASER beam. The resolving power and the mass accuracy of analyzers greatly improved to reach values allowing the direct determination of molecular formula of low molecular weight analytes. However, due to the massive amount of information generated now in (ultra) high resolution ((U)HR) MS data,[2] the extraction of chemical information from spectra is becoming more and more challenging.

In 1963, Kendrick[3] introduced a method based on the two-dimensional projection of the atomic composition space enabling a rapid identification of families of compounds. The Kendrick method is based on the transformation of the mass spectra (mass-to charge ratio or m/z in the X-axis and intensity in the Y-axis) to Kendrick plots where the X-axis corresponds to the Kendrick mass (KM) and the Y-axis corresponds to the Kendrick mass defect (KMD). The KM is defined by converting the exact masses of a given group of atoms (the Kendrick base) to the nearest integer value in atomic mass unit (amu). Kendrick's method changes the IUPAC amu reference (i.e., $^{1}/_{12}$ of the ^{12}C mass) to another mass unit (e.g., $^{1}/_{14}$ of the $^{12}C^{1}H_2$ radical mass). KM can be calculated from m/z value by the eq 1 where the nominal-$M_{\text{Kendrick reference}}$ and $M_{\text{Kendrick reference}}$ are respectively the nominal mass and the exact mass of the Kendrick reference used for the atomic mass unit definition.

$$\text{KM} = m/z \times \frac{\text{nominal} - M_{\text{Kendrick reference}}}{M_{\text{Kendrick reference}}} \quad (1)$$

The KMD corresponds to the difference between the rounded KM values and the KM (eq 2).

$$\text{KMD} = \text{round}(\text{KM}) - \text{KM} \quad (2)$$

All the compounds differing only by the number of repeating unit of the Kendrick reference will have the same KMD for different KM. Consequently, chemically related compounds will be aligned in the KMD axis on the Kendrick plot, making their identification easier. Recently higher order Kendrick transformations have also been applied when oblique correlations suggest the presence of further repeating units.[4]

Received: July 23, 2019
Accepted: September 11, 2019
Published: September 11, 2019

ACS Publications © 2019 American Chemical Society

DOI: 10.1021/acs.analchem.9b03333
Anal. Chem. 2019, 91, 13112–13118

Over the years, KMD analysis has been extended to the analysis of complex samples composed of chemically related structures, such as petrochemicals, but also polymers or lipids.[5−11]

In this paper, we show that the KMD, which is specific to chemically related homologous series of compounds, can be used to replace the m/z axis of the mean spectra classically used for the analysis of images in MSI. The KMD can be adjusted to filter data in a nontargeted, semitargeted and targeted mode. This approach enables an automated, faster and efficient data processing, allowing the rapid identification and localization of families of compounds in the complex chemical space of the image. This method has been successfully applied to two different types of samples, bacteria cocultures and brain tissue sections.

EXPERIMENTAL SECTION

Materials. The MALDI matrices 2,3-diaminonaphthalene (DAN) and α-cyano-4-hydroxycinnamic acid (HCCA) were purchased from Sigma-Aldrich (Belgium). Acetonitrile and Methanol were LC-MS grade from Biosolve (Belgium). Poly(ethylene oxide) monomethyl ether (CH_3O-PEO-H) having an average nominal mass of 750 g/mol, tetraalkylammonium bromides (reagent grade >98%), trifluoroacetic acid (TFA), formic acid, bovine serum albumin (BSA), methyl-*tert*-butyl ether (MtBE), acetone HPLC grade and sodium chloride were purchased from Sigma-Aldrich (Belgium). Pr. Philippe Jacques (Terra Teaching and Research Center, Microbial Processes and Interactions, University of Liège, Belgium) and Lipofabrik (Lille, France) kindly provided standards of lipopeptides.

Sample Preparation for MALDI-FT-ICR MS Imaging Proof of Concept. DAN matrix solution was prepared at saturation (>30 mg/mL) in ACN/water 70:30 vol/vol spiked with 0.2% TFA. Tetraalkylammonium bromide stock solution was dissolved in ACN to reach the final concentration of 10 μM. Trypsin digestion of BSA using MS grade porcine trypsin (purchased from PIERCE, Thermofisher, Belgium) was prepared at a final concentration of 15 μM in 50% methanol spiked with 0.1% formic acid following the procedure described in the Supporting Information (SI). Lipids were extracted from cultured eukaryotic cells based on Matysah protocol[12] and resuspended in MeOH. Lipopeptides were dissolved in MeOH spiked with 0.1% TFA at final concentration of 10 μM. CH_3O-PEO-H was dissolved in ACN 10 μM NaCl to reach a concentration of 10 μM. Each sample was mixed separately with DAN matrix at a ratio of 1:1, and were spotted randomly on a MALDI AnchorChip target plate (Bruker, Bremen, Germany) in duplicate for high and low resolution MS imaging. In addition, 1 μL of each sample was mixed and DAN matrix was added at a 1:1 ratio and spotted onto the MALDI target plate. Finally, a blank spot of DAN matrix was added.

Bacterial Strains, Medium, and Culture Conditions. The strains used for this study were *Bacillus velezensis* GA1 and *Pseudomonas* sp. CMR12a. Both strains were inoculated on a semisolid agar-based PDA medium (Potato Dextrose Agar) and incubated overnight at 30 °C. At the end of the culture, Petri dishes were sealed with Parafilm M and stored at 4 °C prior to analysis.

Animals, Tissue Sampling, and Sectioning. Mouse brain samples were provided by Prof. Martinez (School of Mental Health and Neuroscience at Maastricht University, Netherlands), and bred in-house as described elsewhere.[13] Animals were sacrificed by CO_2 inhalation and brain tissues were extracted. After extraction, the brains were cut across the sagittal midline and immediately fresh-frozen in liquid nitrogen. The brain parts were subsequently stored at −80 °C. On the day of transport, samples were placed on dry ice and transferred to the University of Liege. All procedures were approved by the Animal Welfare Committee of Maastricht University and were performed according to Dutch federal regulations for animal protection. Frozen brain tissues were sectioned into 14 μm thick sagittal slices using a CryoStar NX70 (Thermo Fisher Scientific, Waltham, MA) at −20 °C and thaw-mounted onto indium−tin oxide (ITO) conductive glass slides (Bruker Daltonics, Bremen, Germany).

Sample Preparation for MALDI-FT-ICR MS Imaging. Microbial colonies on agar and region of interest were cut directly from the Petri dish and transferred to the target ITO plate (Bruker, Bremen, Germany), previously covered with double sided conductive carbon tape (StructureProbe INC, West Chester, PA). This assembly was then placed in a vacuum desiccator until complete drying (overnight). A HCCA matrix solution was prepared at the concentration of 5 mg/mL in ACN/water 70:30 vol/vol spiked with 0.2% TFA, based on the previous work of Debois et al.[14] Application of the matrix solution on the dried bacterial sample was performed using the SunCollect (SunChrom, Friedrichsdorf, Germany) spraying system. In total, 60 layers of HCCA matrix were sprayed on the sample. The first three layers were sprayed at 5 μL/min, and the other at 10 μL/min. Mouse brain tissue slides deposed on ITO slides were first put into a vacuum desiccator for approximatively 20 min. A solution of HCCA at the concentration of 5 mg/mL in MeOH/water 90:10 vol/vol spiked with 0.2% TFA was prepared, and sprayed onto the dried sample by the SunCollect. The number of spraying layers were set to 30, at a spraying speed of 10 μL/min, except for the three first layers sprayed at 5 μL/min.

FT-ICR Mass Spectrometry. Mass spectrometry images were obtained using a FT-ICR mass spectrometer (SolariX XR 9.4T, Bruker Daltonics, Bremen, Germany). The mass spectrometer was systematically mass calibrated from 200 m/z to 2,300 m/z before each analysis with a red phosphorus solution in pure acetone spotted directly onto the ITO Glass slide or onto the MALDI plate to reach a mass accuracy better than 0.5 ppm. FlexImaging 5.0 (Bruker Daltonics, Bremen, Germany) software was used for MALDI MS imaging acquisition, with a pixel step size for the surface raster set to 150 μm for Kendrick proof of concept imaging and to 100 μm for bacteria and brain tissue imaging. For each mass spectrum, one scan of 20 LASER shots was performed at a repetition rate of 200 Hz. The LASER power was set to 50% and the beam focus was set to "small". The mass spectra for high and low resolution mass images were recorded by setting the number of data points in the transient at respectively 1 000 000 (1M) and 256 000 (256 K).

Kendrick Filtering for Data Reduction and Clustering. MSconvert (Proteowizard)[15] and Fleximaging 5.0 (Bruker) have been respectively used to convert MS raw files (obtained from, for example, direct infusion MS, liquid chromatograph MS, ion mobility MS···) to mzML and MSI raw files (obtained from imaging MS) to imzML files. These files are processed under Python in-house software (MSKendrickFilter, available upon request from c.kune@uliege.be).

DOI: 10.1021/acs.analchem.9b03333
Anal. Chem. 2019, 91, 13112−13118

After a conversion to binary format files (.ICK), specifically designed for our software, data are filtered based on the KMD. For a MS analysis, the .ICK file contains the mass-to-charge and intensity information for each detected ion of each recorded scan (or pixel in case of MSI application). A noise reduction, based on a minimum intensity threshold, can be applied to limit the.ICK file size and the CPU and RAM usage of the software.

The Kendrick filtering method workflow for MSI data or MS data (e.g., direct infusion, chromatogram, mobilogram, electrophoregram···) is depicted in Figure 1. Once the .ICK file has been generated, an ion list containing all the m/z present in the mean spectrum (sum of each MS spectra in MS or MSI data) is created. The software calculates a Kendrick mass defect (KMD) for each m/z value according to eq 1 and eq 2 in the user-defined Kendrick reference (i.e., $-CH_2-$ in this work). Different rounding functions are available in our software, that is, floor, ceil, and round. Floor option has been used for the results reported here. All m/z ratio are then associated with a KMD. The second step consists in filtering the total ion list by keeping only the ions whose KMD is included in the user-defined KMD range. A m/z ratio selection range can also be applied as a criterion for data filtering. The resulting ion list contains chemically related compounds. In the case of different compound families sharing a similar KMD, an additional algorithm is added to cluster the saved ion by compound families. This clustering step (reported as mass difference clustering algorithm in Figure 1) groups ions according to their mass difference corresponding to a multiple of the chosen Kendrick reference mass. The software allows a user-defined m/z tolerance value (e.g., here we used 5 and 2 ppm for m/z tolerances with resolution of, 20 000 and 160 000 full width at half-maximum, fwhm, at 410 m/z, respectively). In this version of the software, isotope compounds are considered as different families due to Kendrick mass defect deviations. This algorithm generates an ion list for each putative compound family. Finally, a chromatogram (for liquid chromatography or gas chromatography MS), a mobilogram (for ion mobility MS), an electrophoregram (for capillary electrophoresis MS) or an image (for MSI) can be generated from these ion lists. The entire process can be repeated with different KMD and m/z ranges to generate additional MS data of different compound families.

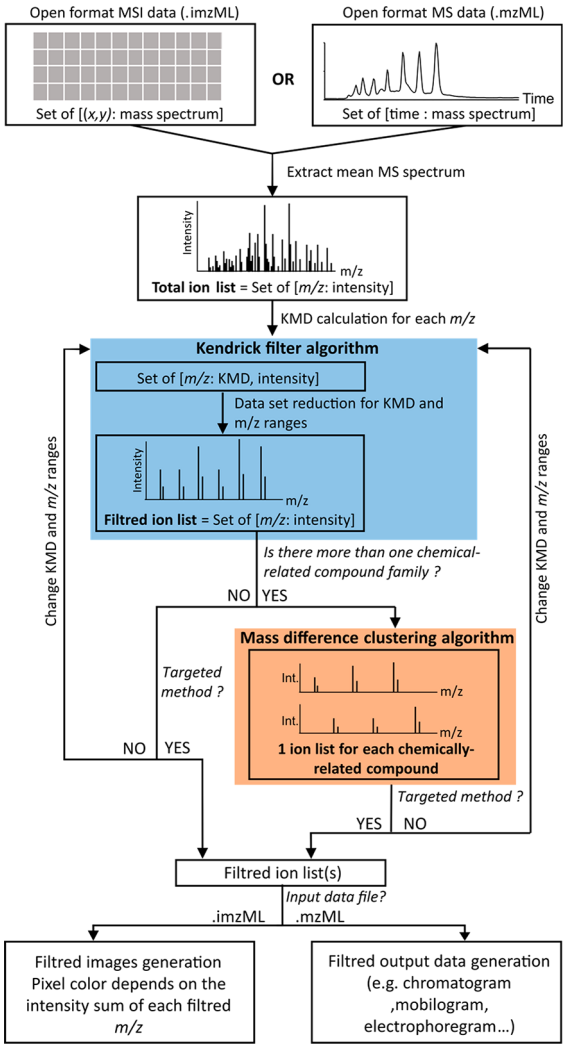


Figure 1. Kendrick filtering method workflow applied on MSI data and MS data (such as chromatography, ion mobility, capillary electrophoresis···).

In this work, we only focused the Kendrick filtering application on MSI data as a rapid visualization and localization tool of chemically related compounds. An intensity threshold value of 200.000 counts per scan and a normalization to the total ion current (TIC) for each pixel was applied for all the reported images.

RESULTS AND DISCUSSION

KMD Filtering Concept Applied to MS Data. The illustration of the KMD filtering concept is shown in Figure 2 for a MALDI FT-ICR MS data. The MALDI spots were prepared with a mix of different compound families including lipids, lipopeptide, polymers, tetraalkylammonium (TAA) and peptides of BSA tryptic digest. The resulting mass spectra (Figure 2B) is rather complex and the assignment of MS peak associated with compounds families is challenging. The Kendrick plot (Figure 2A), generated using $-CH_2-$ as Kendrick reference significantly reduces the complexity by aligning the compounds bearing a repeating unit of $-CH_2-$. Tetraalkylammonium, lipids, and lipopeptides families are then easily detected (respectively highlighted in blue, green, and orange), and can be quickly selected in a rectangular area in the KMD plot. Polymers families, which are in the same mass range of lipids families, are revealed by changing the Kendrick reference to the repeating monomer unit $-C_2H_4O-$ (see SI Figure S1). Thus, the goal of the proposed Kendrick filtering method relies on its capacity to extract the relevant peaks associated with a given family of compounds from a raw mass spectrum and to generate a new filtered spectrum, chromatogram, mobilogram, or image. Depending on the mass spectrometer resolution, setting a KMD range as well as a MS range might not be sufficient to isolate a single family of chemically related compounds. Indeed, it can be explained by the presence of other compounds exhibiting similar KMD and m/z. In these cases, the mass difference clustering algorithm (see the Materials section) can be applied to separate overlapping families of compounds. Thus, a sequential strategy consisting in first filtering by KMD and then by the mass difference clustering gives clean filtered spectra of tetraalkylammonium (Figure 2.C), surfactins (lipopeptides family, Figure 2.D) and phosphatidylcholines (lipid family, Figure 2.E) from a MALDI spot.

DOI: 10.1021/acs.analchem.9b03333
Anal. Chem. 2019, 91, 13112−13118

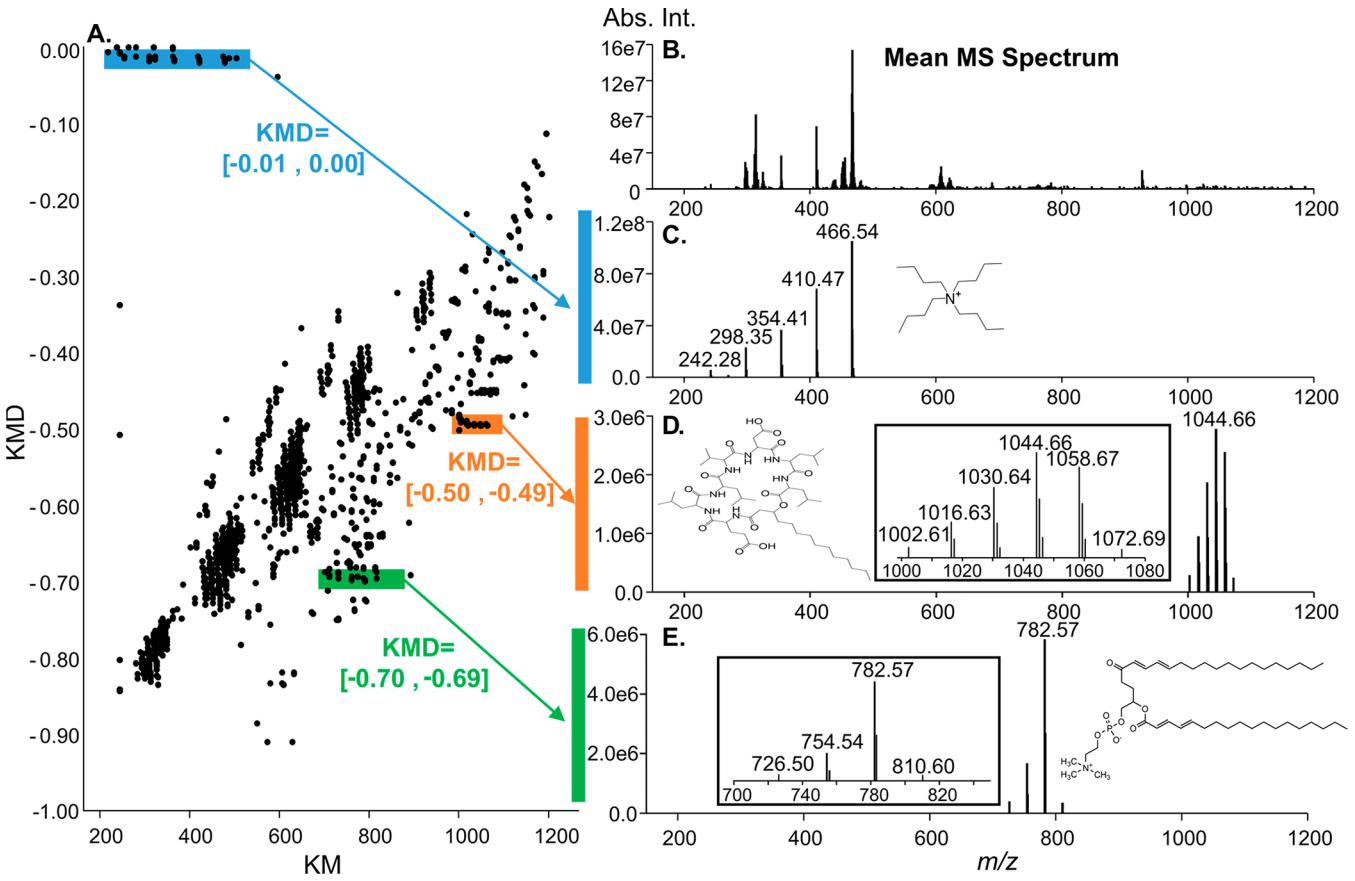


Figure 2. Illustration of the KMD filtering method. (A) Kendrick mass defect plot of the spectra obtained by FT-ICR MS. (B) Full MS spectrum when no filtration is applied. By selecting a specific KMD range and eventually a specific mass range, it is possible to filter the MS Spectra and show only the chemically related molecules such as TAA (C), lipopeptides: surfactins (D) or lipids: phosphatidylcholines (E).

KMD Filtering Concept Applied to MSI Data. As a proof-of-concept, but also to assess the effectiveness of KMD filtering process on MSI data, we randomly spotted a MALDI AnchorChip plate with several droplets of lipids, tetraalkylammonium salts, lipopeptides and polymers (Figure 3.A). The proposed KMD filtering algorithm developed here was applied on the imaging data set, using directly the KMD values previously obtained in the MS analysis. The MSI data were acquired with two different MS resolution calculated at the *m/z* 410 (i.e., 20 000 and 160 000, respectively). A mixture spot (spot 8, containing all the investigated compound families) and a blank matrix spot (Spot 3) were added as a positive and a negative control, respectively.

Lipids, lipopeptides and TAA were isolated using the $-CH_2-$ reference, polymers were isolated using the $-C_2H_4O-$ reference (see SI Figure S2). The KMD filtering process allowed to determine the precise localization of each chemically related compounds for both MS resolution acquired data sets. The similarities of the images produced at 20 K and 160 K resolution is in fact ensured by the mass difference clustering algorithm. In addition, one should note that none of the targeted compounds were found in the negative control spot 3 at both selected MS resolution.

These results outline the power of KMD filtering approach as a rapid visualization tool on MSI data generated from (U)HR MS analyzer with different MS resolutions.

MALDI Imaging of Bacterial Culture. Mass spectrometry imaging of bacterial culture has gained growing attention.[14,16−19] By combining the chemical identification of a compound with its localization, MSI could notably provide more insights into the bacterial interspecies molecular interactions. The objective would be to detect and localize compounds of interest, such as lipopeptides, produced by the bacteria in different environments or at different incubation times. Since most of lipopeptides contain a variable hydrophobic lipid chain-length,[20] applying a KMD plot with a $-CH_2-$ repeating unit, seems to be a fit-for-purpose approach to rapidly screen the detected compounds in the region of interest. The results were analyzed according to a nontargeted approach, by screening along the KMD axis different families of compounds (Figure 4). These compounds were then identified based on their exact mass. Three different families of lipopeptides were detected and identified as iturins (Figure 4B), surfactins (Figure 4C) and sessilins (Figure 4F). As expected, the distribution of lipopeptides differs according to the different families. Surfactins and iturins are secreted by *Bacillus* and readily diffuse in the agar medium, suggesting an interplay between *Bacillus* and *Pseuomonas*.[21] By contrast, the sessilins formed by *Pseudomonas*, remained associated with the cells and are not detected outside the bacterial colony.

In the KMD range −0.65 to −0.66, at least two different lipid families were detected (Figure 4D and E). Thanks to the mass difference clustering algorithm, it is possible to discriminate these two groups. The first group of lipids (Figure 4.D) contained the *m/z* values: 710.444, 724.459, and 738.477, belonging to the sub class 1-(1Z-alkenyl)2-acylglycerophosphoethanolamines of the glycerophosphoethanolamines lipid class. Based on its localization on the ITO slide, this lipid

DOI: 10.1021/acs.analchem.9b03333
Anal. Chem. 2019, 91, 13112−13118

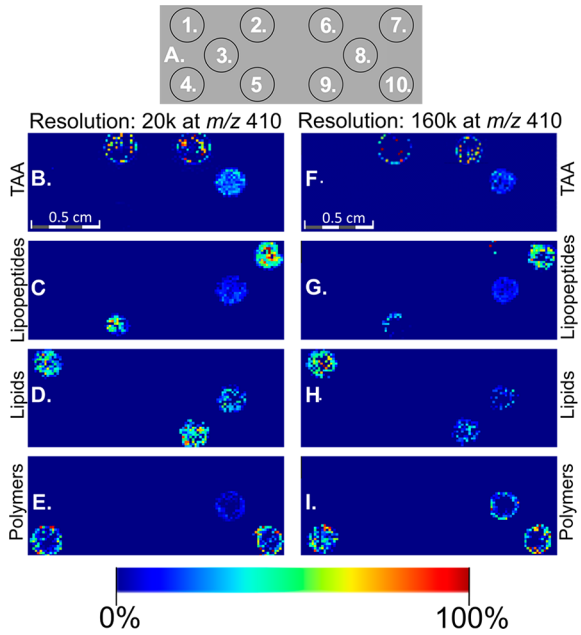


Figure 3. KMD filtering applied on MALDI mass spectrometry imaging data where different chemically related compounds were randomly spotted on a MALDI AnchorChip target plate. A. Scheme of the distribution of chemically related compounds (spots 1. and 9.: lipids, spots 2. and 6.: tetraalkylammonium, spot 3.: matrix, spot 4. and 10.: polymers, spots 5. and 7.: lipopeptides, spot 8.: Mix). B to E are generated image after KMD filtering for, respectively, TAA, lipopeptides, lipids and polymers from MSI data with a MS resolution of fwhm 20 000 for m/z 410. F to I are generated image after KMD filtering for, respectively, TAA, lipopeptides, lipids and polymers from MSI data with a MS resolution of fwhm 160 000 for m/z 410. The color scale represents a normalized intensity of the ions.

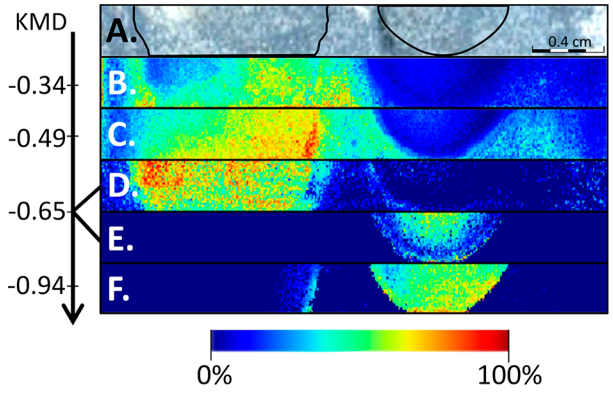


Figure 4. KMD filtering applied on MALDI Imaging of bacterial culture. (A) Optical image of the sample covered with matrix, *Bacillus* is circled on the left side, and *Pseudomonas* is circled on the right side. The identified families are presented according to their KMD value. (B) Iturins A (sodium adducts) at the KMD range −0.33 to −0.35 (C) Surfactins (sodium adducts) at the KMD range −0.49 to −0.51. (D) and (E) identification of two different lipid classes at the KMD range −0.65 to −0.66, (F) Sessilins (sodium adducts) at the KMD range −0.93 to −0.95. The color scale represents the normalized intensity of the ions.

group was specifically related to *Bacillus velezensis* GA1. The second group of lipids (Figure 4.E) composed of the following m/z values; 734.472 and 748.487 was identified as diacylglycerophosphoethanolamines or diacylglycerophosphocholines. Their spatial distribution was specific to *Pseudomonas* sp. CMR12a.

MALDI Imaging of a Mouse Brain Section. The last example treated in this paper belongs to MSI applied on mice tissue. MSI enables the localization of a set of diverse molecules in a specific tissue, such as brain,[22] liver,[23] or kidney,[24] providing local molecular information for a better understanding of illnesses and deeper insights into physiological mechanisms. Among the detected compounds, lipids represent a high percentage of ions since they are present in every cell by constituting the membranes and energy storage vesicles.[25,26] Their implication in various biological functions and their capacity to reflect the physiological and environmental conditions makes their study essential for understanding diseases or biomarker discovery.[27−30] The results obtained by MSI of a mouse brain section (Figure 5) shows

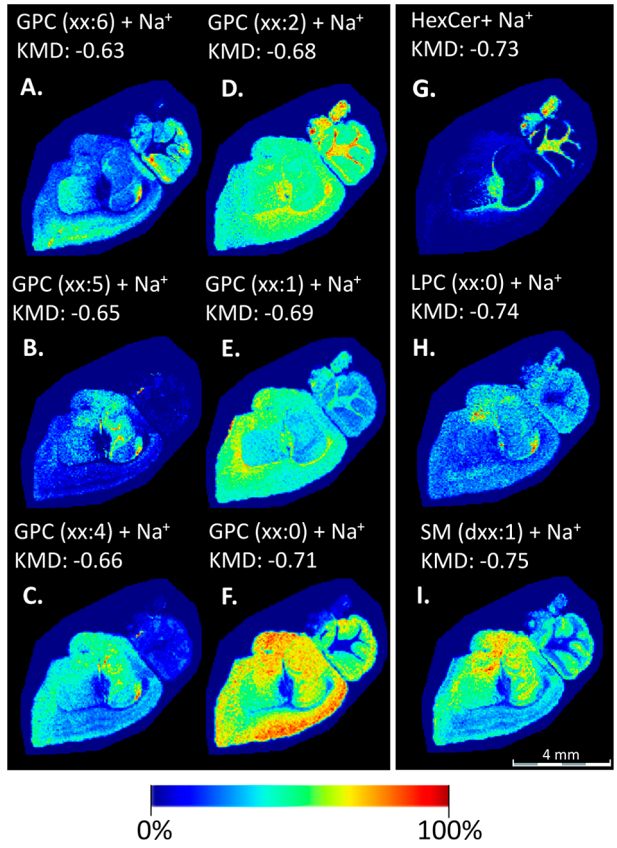


Figure 5. KMD filtering applied on MALDI imaging of a sagittal slice of mouse brain. (A) to (F) Image of sodiated glycerophosphocholines,$[GPCs+Na]^+$ with 6, 5, 4, 2, 1, and 0 insaturations, respectively. (G) Image of sodiated hexosylceramides, $[HexCers+Na]^+$. (H) Image of sodiated lysophoshocholins, $[LPCs+Na]^+$. (I) Image of sodiated sphingomyelins, $[SMs+Na]^+$. The color scale represents the normalized intensity of the ions.

different patterns of lipid distribution belonging to the families of glycerophosphocholines (GPCs), hexosylceramides (HexCers), lysophoshocholins (LPCs), and sphingomyelins (SMs) classes, as described in the literature on brain tissue analysis.[31,32] The software shows the list of peaks used to build the distribution of each family. From that list, individual distributions can be obtained. All the detected families of compounds were tentatively assigned according to their class (GPCs, SMs, LPCs) and according to their unsaturation

DOI: 10.1021/acs.analchem.9b03333
Anal. Chem. 2019, 91, 13112−13118

(xx:n), where "xx" represents the number of carbon and "n" corresponds to the number of unsaturation. The attribution of a specific class was done based on high-resolution MS data, using LIPIDMAPS database (The LIPID MAPS Lipidomics Gateway, http://www.lipidmaps.org/). The database search was performed only on even-chain lipids, with a mass tolerance of ±0.005 m/z. The list of detected compounds and tentatively assigned compounds is available in SI Figure S3. An additional image on an adjacent section of the mouse brain tissue was recorded in the same experimental conditions to confirm the reproducibility of the spatial distribution after KMD filtering (see SI Figure S4). Brain regions (see SI Figure S5) were identified according to the Allen mouse brain atlas (https://mouse.brain-map.org/).

Interestingly, after applying our Kendrick filtering software, some of the detected GPCs from mouse brain tissue sections tend to be differentially colocalized, depending on their unsaturation degree. At six unsaturations on the lipid chain (Figure 5A. and SI Figure S6), GPCs are mainly detected in the isocortex, the cerebral nuclei and cerebellar cortex, while GPCs with 5 unsaturations are detected mainly in the hippocampal formation (Figure 5B. and SI Figure S6), and the GPCs with four unsaturations are detected in the hippocampal formation and the olfactory areas (Figure 5C). When the GPCs lipid chain contains one or two unsaturations, they are mainly localized in the fiber tracts (Figure 5D and E.), but when the GPCs are fully unsaturated (Figure 5F), they will be detected only in the cerebrum and cerebellar cortex. This variation in the localization of GPCs conforms to previous imaging data of lipids in brain tissue.[33] Finally, additional relevant lipid distribution can also be detected, such as the specific localization of HexCers in the fiber tracts (Figure 5G.), of LPCs mainly in the hippocampal formation (Figure 5H.), and the SMs in the hippocampal formation, the cerebral nuclei, the olfactory area and the cerebellar cortex (Figure 5I.).

By applying the KMD MSI data filtering approach proposed here, any type of biological modification occurring on lipids such as shorter or longer lipid chain length or lipid oxidation could be easily detected on the KMD plot. Image reconstruction based on the KMD plot provides, for a given KMD range, a specific lipid fingerprint that can further be used to compare different images.

CONCLUSION

In this work, we have used KMD as a filtering tool to visualize the chemically related compounds in complex mass spectra, as described in the literature. The mass resolving power and mass accuracy of modern instruments ensure an accurate KMD filtering. For overlapping families, a second filtering process using the mass distance between peaks has been introduced. Based on those selection criteria, we were able to group compounds presenting the same KMD in a single spectrum and to apply the KMD analysis to molecular images obtained by MALDI MSI where it reveals all its power. Replacing the m/z axis of the reconstructed mean mass spectrum by a KMD axis brings a real benefit for rapid and automatic image reconstruction not only for specific organic molecules but also for families, speeding up the identification process and facilitating data analysis. The use of the KMD over normal mass[34] defect enables to choose the reference unit adapted ($-CH_2-$ for lipids, $-C_2H_4O-$ for polymers), and simplifies the area selection on the KMD plot, since all the chemically related families would be aligned. Finally, future developments on higher-order KMD will be soon implemented in our software. The developed approach has been successfully applied through two practical examples, that is, bacterial cocultures and histological sections of mouse brain. The KMD filtering method applied to bacterial culture image data has permitted the rapid detection and localization of different groups of compounds, such as lipopeptides or lipids. The application of the KMD filtering method to histological tissue sections of a mouse brain highlighted the localization of lipids according to their family and their number of unsaturation.

In light of these results, we argue that the methodology outlined in this paper can be used as a nontargeted screening tool, reconstructing the image by scanning the KMD scale according to the workflow proposed in Figure 1. Nontargeted, semi targeted or targeted modes are different options developed in the software, allowing to rapidly detect and localize metabolites containing the repetitive unit of the substrate molecule or a specific family of compounds, through their known KMD. The same approach can be extended to any type of data sets in mass spectrometry imaging with different ionization process (e.g., MALDI MSI or DESI MSI) but also when MS is hyphenated with separation methods, no matter the separation method employed (e.g., ion mobility or liquid chromatography) or the mass analyzers used.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.9b03333.

Full detailed experimental section, KMD analysis applied on polymers (S1), KMD plots for KMD filtering applied on MALDI mass spectrometry imaging data (S2), List of identified compounds after KMD filtering applied on MALDI imaging of a sagittal mouse brain (S3), KMD filtering applied on MALDI imaging of a sagittal slice of mouse brain tissue (S4), annotated sagittal mouse brain slice (S5). and individual lipid distribution on sagittal slice of mouse brain tissue (S6) (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: e.depauw@uliege.be.

ORCID

Christopher Kune: 0000-0002-3010-8173
Andréa McCann: 0000-0002-1093-6414
Loïc Quinton: 0000-0001-8153-9590
Edwin De Pauw: 0000-0003-3475-1315

Author Contributions

‖C.K. and A.M. contributed equally.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Prof. Philippe Jacques, from the Microbial Processes and Interactions (MiPI) research unit of Gembloux Agro-Bio Tech (Belgium) and Lipofabrik (Lille, France) for providing lipopeptides samples. We acknowledge financial support from the Excellence Of Science Program of the FNRS F.R.S (Rhizoclip EOS 18000802), from the Interreg EMR project: EURLIPIDS (R-8598) and from the European Union's

DOI: 10.1021/acs.analchem.9b03333
Anal. Chem. 2019, 91, 13112−13118

Horizon 2020 research and innovation program under grant agreement No. 731077. Finally, we acknowledge the European Union and Wallonia program FEDER for the founding of the instruments (SolariX XR 9.4T). All experiments were done with permission from the Committee on Animal Welfare of Maastricht University, according to Dutch governmental rules.

REFERENCES

(1) Schmitt-Kopplin, P.; Hertkorn, N. *Anal. Bioanal. Chem.* **2007**, *389* (5), 1309−1310.
(2) Cho, Y.; Ahmed, A.; Islam, A.; Kim, S. *Mass Spectrom. Rev.* **2015**, *34* (2), 248−263.
(3) Kendrick, E. *Anal. Chem.* **1963**, *35* (13), 2146−2154.
(4) Roach, P. J.; Laskin, J.; Laskin, A. *Anal. Chem.* **2011**, *83* (12), 4924−4929.
(5) Hughey, C. A.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G.; Qian, K. *Anal. Chem.* **2001**, *73* (19), 4676−4681.
(6) Marshall, A. G.; Rodgers, R. P. *Acc. Chem. Res.* **2004**, *37* (1), 53−59.
(7) Morgan, T. E.; Ellacott, S. H.; Wootton, C. A.; Barrow, M. P.; Bristow, A. W. T.; Perrier, S.; O'Connor, P. B. *Anal. Chem.* **2018**, *90* (19), 11710−11715.
(8) Lerno, L. A.; German, J. B.; Lebrilla, C. B. *Anal. Chem.* **2010**, *82* (10), 4236−4245.
(9) Nakamura, S.; Cody, R. B.; Sato, H.; Fouquet, T. *Anal. Chem.* **2019**, *91* (3), 2004−2012.
(10) Zheng, Q.; Morimoto, M.; Sato, H.; Fouquet, T. *Fuel* **2019**, *235*, 944−953.
(11) Roullier-Gall, C.; Witting, M.; Gougeon, R. D.; Schmitt-Kopplin, P. *Front. Chem.* **2014**, *2*, 102.
(12) Matyash, V.; Liebisch, G.; Kurzchalia, T. V.; Shevchenko, A.; Schwudke, D. *J. Lipid Res.* **2008**, *49* (5), 1137−46.
(13) Youmans, K. L.; Tai, L. M.; Nwabuisi-Heath, E.; Jungbauer, L.; Kanekiyo, T.; Gan, M.; Kim, J.; Eimer, W. A.; Estus, S.; Rebeck, G. W.; Weeber, E. J.; Bu, G.; Yu, C.; Ladu, M. J. *J. Biol. Chem.* **2012**, *287* (50), 41774−86.
(14) Debois, D.; Ongena, M.; Cawoy, H.; De Pauw, E. *J. Am. Soc. Mass Spectrom.* **2013**, *24* (8), 1202−13.
(15) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. *Nat. Biotechnol.* **2012**, *30* (10), 918−20.
(16) Yang, J. Y.; Phelan, V. V.; Simkovsky, R.; Watrous, J. D.; Trial, R. M.; Fleming, T. C.; Wenter, R.; Moore, B. S.; Golden, S. S.; Pogliano, K.; Dorrestein, P. C. *J. Bacteriol.* **2012**, *194* (22), 6023−6028.
(17) Debois, D.; Jourdan, E.; Smargiasso, N.; Thonart, P.; De Pauw, E.; Ongena, M. *Anal. Chem.* **2014**, *86* (9), 4431−4438.
(18) Boya, P. C. A.; Fernández-Marín, H.; Mejía, L. C.; Spadafora, C.; Dorrestein, P. C.; Gutiérrez, M. *Sci. Rep.* **2017**, *7* (1), 5604.
(19) Dunham, S. J. B.; Ellis, J. F.; Li, B.; Sweedler, J. V. *Acc. Chem. Res.* **2017**, *50* (1), 96−104.
(20) Ongena, M.; Jacques, P. *Trends Microbiol.* **2008**, *16* (3), 115−125.
(21) Molina-Santiago, C.; Pearson, J. R.; Navarro, Y.; Berlanga-Clavero, M. V.; Caraballo-Rodriguez, A. M.; Petras, D.; Garcia-Martin, M. L.; Lamon, G.; Haberstein, B.; Cazorla, F. M.; de Vicente, A.; Loquet, A.; Dorrestein, P. C.; Romero, D. *Nat. Commun.* **2019**, *10* (1), 1919.
(22) Stoeckli, M.; Staab, D.; Staufenbiel, M.; Wiederhold, K.-H.; Signor, L. *Anal. Biochem.* **2002**, *311* (1), 33−39.
(23) Shimma, S.; Sugiura, Y.; Hayasaka, T.; Hoshikawa, Y.; Noda, T.; Setou, M. *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* **2007**, *855* (1), 98−103.
(24) Zoriy, M.; Matusch, A.; Spruss, T.; Becker, J. S. *Int. J. Mass Spectrom.* **2007**, *260* (2), 102−106.
(25) Murphy, R. C.; Hankin, J. A.; Barkley, R. M. *J. Lipid Res.* **2009**, *50* (Suppl), S317−22.
(26) Burnum, K. E.; Cornett, D. S.; Puolitaival, S. M.; Milne, S. B.; Myers, D. S.; Tranguch, S.; Brown, H. A.; Dey, S. K.; Caprioli, R. M. *J. Lipid Res.* **2009**, *50* (11), 2290−2298.
(27) Wenk, M. R. *Nat. Rev. Drug Discovery* **2005**, *4* (7), 594−610.
(28) Orešič, M.; Hänninen, V. A.; Vidal-Puig, A. *Trends Biotechnol.* **2008**, *26* (12), 647−652.
(29) Yang, K.; Han, X. *Trends Biochem. Sci.* **2016**, *41* (11), 954−969.
(30) Lam, S. M.; Shui, G. *J. Genet. Genomics* **2013**, *40* (8), 375−390.
(31) Sparvero, L. J.; Amoscato, A. A.; Dixon, C. E.; Long, J. B.; Kochanek, P. M.; Pitt, B. R.; Bayır, H.; Kagan, V. E. *Chem. Phys. Lipids* **2012**, *165* (5), 545−562.
(32) Ellis, S. R.; Paine, M. R. L.; Eijkel, G. B.; Pauling, J. K.; Husen, P.; Jervelund, M. W.; Hermansson, M.; Ejsing, C. S.; Heeren, R. M. A. *Nat. Methods* **2018**, *15* (7), 515−518.
(33) Carter, C. L.; McLeod, C. W.; Bunch, J. *J. Am. Soc. Mass Spectrom.* **2011**, *22* (11), 1991.
(34) McDonnell, L. A.; van Remoortere, A.; de Velde, N.; van Zeijl, R. J. M.; Deeldera, A. M. *J. Am. Soc. Mass Spectrom.* **2010**, *21* (12), 1969−1978.

DOI: 10.1021/acs.analchem.9b03333
Anal. Chem. 2019, 91, 13112−13118

Associated papers of Chapter 4

A5- Original publication of “Organ-resolved lipid mapping in Steatoda nobilis spider model using high-resolution mass spectrometry imaging and Kendrick mass defect analysis.”

frontiers | Frontiers in Chemistry

TYPE Original Research
PUBLISHED 03 September 2025
DOI 10.3389/fchem.2025.1658546

Check for updates

OPEN ACCESS

EDITED BY
Stefan A. Pieczonka,
Technical University of Munich, Germany

REVIEWED BY
James Marshall,
Mars Petcare, United Kingdom
Laimdota Zizmare,
Eberhard Karls University, Germany
Stefan Ruchti,
Technical University of Munich, Germany

*CORRESPONDENCE
Loïc Quinton,
loic.quinton@uliege.be

†These authors have contributed equally to this work

RECEIVED 02 July 2025
ACCEPTED 13 August 2025
PUBLISHED 03 September 2025

CITATION
Redureau D, Dunbar JP, La Rocca R, De Monts De Savasse A, Bastiaens Q, Bertrand V, Kune C, Tiquet M, Far J, De Pauw E, Dugon MM and Quinton L (2025) Organ-resolved lipid mapping in *Steatoda nobilis* spider model using high-resolution mass spectrometry imaging and Kendrick mass defect analysis. *Front. Chem.* 13:1658546. doi: 10.3389/fchem.2025.1658546

COPYRIGHT
© 2025 Redureau, Dunbar, La Rocca, De Monts De Savasse, Bastiaens, Bertrand, Kune, Tiquet, Far, De Pauw, Dugon and Quinton. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Organ-resolved lipid mapping in *Steatoda nobilis* spider model using high-resolution mass spectrometry imaging and Kendrick mass defect analysis

Damien Redureau[1†], J. P. Dunbar[2,3†], Raphaël La Rocca[1], Axel De Monts De Savasse[1], Quentin Bastiaens[1], Virginie Bertrand[1], Christopher Kune[1], Mathieu Tiquet[1], Johann Far[1], Edwin De Pauw[1], Michel M. Dugon[2] and Loïc Quinton[1*]

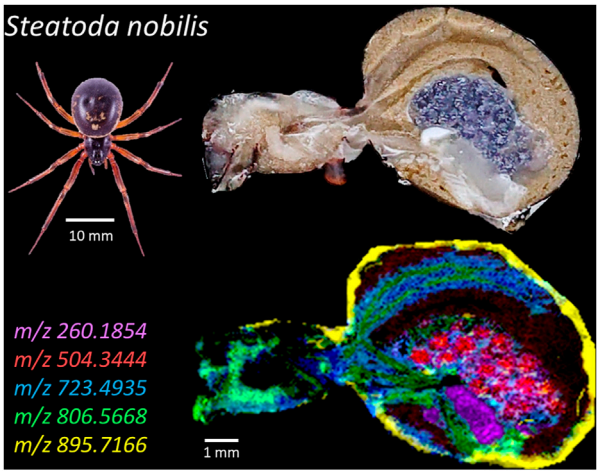
[1]Mass Spectrometry Laboratory, MolSys RU, University of Liège, Liège, Belgium, [2]Venom Systems and Proteomics Lab, School of Natural Sciences, Ryan Institute, University of Galway, Galway, Ireland, [3]Midlands Bug and Reptile Zoo, Longford, Ireland

The noble false widow spider (*Steatoda nobilis*), a rapidly spreading member of the Theridiidae family, has gained attention for its increasing presence near human habitats and its medical significance due to envenomation reports. Recent studies have revealed that its venom contains α-latrotoxins, toxins also found in *Latrodectus* (black widows), responsible for latrodectism symptoms. Despite this growing interest, little is known about the lipidome and metabolome of *S. nobilis*, which could offer insights into its ecological role, dietary metabolism, and chemical communication. In this study, we used Matrix-Assisted Laser Desorption/Ionization Fourier-Transform Ion Cyclotron Resonance (MALDI-FT-ICR) mass spectrometry imaging (MSI) to investigate the whole-body lipid and metabolite distribution in *S. nobilis*. MSI is a powerful tool that couples molecular analysis with spatial information, enabling detailed visualization of biomolecules in tissues. Applying MSI to arachnids offers a novel approach to explore organ-specific metabolic profiles and identify potentially bioactive or adaptive compounds. One of the major challenges was preserving the spider's fragile internal anatomy during sample preparation. We developed a gelatin-based fixation method to obtain intact histological sections suitable for MSI analysis. This allowed us to clearly distinguish organ-specific lipid and metabolite distributions *in situ*, including within the silk glands, ovaries, and nervous tissues. A second challenge was managing the vast data generated by MSI, with each image yielding thousands of molecular peaks. To streamline analysis, we employed Kendrick Mass Defect (KMD) plots to classify ions into structural families. This approach enabled us to link specific ions to molecular families and localize them within the spider's body, enhancing our anatomical understanding at the molecular level. This work not only provides foundational insights into *S. nobilis* biochemistry but also demonstrates the potential of MSI for advancing arachnid lipidomics and uncovering molecules of ecological or

biomedical interest. It opens the gates for broader applications of spatial lipidomics in other small biosystems and animals, particularly those previously inaccessible to detailed biochemical analysis.

KEYWORDS

lipidomics, lipids/chemistry, mass spectrometry imaging, Kendrick mass defects, whole-body imaging, arachnid



GRAPHICAL ABSTRACT

Introduction

Mass spectrometry has been applied to elucidate the lipidome and metabolome of various biological systems, but the world of arachnids has been mostly unexplored. Arachnids, a class of chelicerate arthropods characterized by eight legs and a body divided into two tagmata, represent a fascinating group of organisms with unique physiological adaptations and ecological roles (Stork, 2017). This class encompasses a wide variety of organisms, including spiders, scorpions, ticks, and mites. With approximately 120,000 species described, arachnids occupy a broad range of ecological niches, displaying remarkable adaptations to different environments and lifestyles. As predatory arthropods, arachnids play essential roles in controlling insect populations, maintaining ecosystem balance, and serving as indicators of environmental health (Marc et al., 1999). Within the Class Arachnida, spiders (Order Araneae) produce venoms mainly used for predation and defense. Venom composition is highly diverse and contains a complex mixture of proteins, peptides, and small molecules, such as metabolites (Langenegger et al., 2019). Most of the current studies target the peptide and protein content of venoms, looking for novel pharmacological tools and potential future drugs. In this context, spider venom lipidome and metabolome have been poorly considered (Chen et al., 2023; Klupczynska et al., 2018), even if these could provide insights into the envenoming process. Spiders are also renowned for their silk production, which serves multiple purposes, including prey capture and the construction of shelters, brooding chambers and egg sacs (Römer and Scheibel, 2008). While extensive research has focused on the composition (Lu et al., 2024) and gene expression (Connor et al., 2024) of silk proteins, relatively little attention has been given to the underlying metabolic pathways that support silk biosynthesis. Lastly, spiders often experience long periods of fasting and have therefore evolved metabolic mechanisms to store and mobilize energy reserves very efficiently, which can be investigated again through lipidomics and, at a lower level, metabolomics.

Steatoda nobilis, our model for whole-body MSI in lipidomics

We chose the noble false widow spider *Steatoda nobilis,* native to the Canary Islands and Madeira, as a model organism. The species is now well-established and invasive in parts of the Americas and Western Europe, particularly the United Kingdom, Ireland, and increasingly throughout western continental Europe due to a range of adaptive traits after spreading using global trade routes (Dugon et al., 2017; Rayner et al., 2022). Its success in colonizing new

Abbreviations: CHCA, α-Cyano-4-hydroxycinnamic acid; DG, Diacylglyceride; FT-ICR, Fourier Transform Ion Cyclotron Resonance; ITO, Indium-Tin Oxide; KMD, Kendrick Mass Defect; LC-MS/MS, Liquid Chromatography-Tandem Mass Spectrometry; LPC, Lysophosphatidylcholine; LPE, Lysophosphatidylethanolamine; MG, Monoacylglyceride; MS, Mass Spectrometry; MSI, Mass Spectrometry Imaging; OCT, Optimal Cutting Temperature; PA, Phosphatidic Acid; PC, Phosphatidylcholine; PE, Phosphatidylethanolamine; ppm, Parts Per Million; ROI, Region of Interest; ST, Sterol; TG, Triacylglyceride; TIC, Total Ion Current.

environments has raised ecological and public health concerns (Dunbar et al., 2018; Dunbar et al., 2022b). Morphologically, *S. nobilis* is a medium-sized spider, with females reaching up to 15 mm in body length and over 30 mm in leg span (Supplementary Figure S1). *S. nobilis* presents a dark brown to black bulbous abdomen with a distinctive pale cream or ivory-colored pattern that may resemble a skull or hexagonal geometric motif. Like all other spiders, *S. nobilis* possesses a cephalothorax and an abdomen, connected by a narrow pedicel, eight jointed legs protruding from the cephalothorax, prominent spinnerets on the posterior aspect of the abdomen, and anterior chelicerae from which venom is secreted. Internally, spiders feature a simple digestive system, book lungs, silk glands, a pair of large venom glands extending within the cephalothorax and connected to the chelicerae, a prominent curved heart and a nervous system adapted for efficient prey capture and environmental sensing. In addition, females possess an ovary sac, an oviduct and a sperm receptacle used for fecundation (Supplementary Figure S2). While the detailed internal anatomy of *S. nobilis* has never been investigated, main morphological characters are generally conserved across spiders, with minor variations occurring (e.g., number of eyes, number and types of silk glands). A notable exception is the occurrence of complex, highly variable and species-specific genitalia of both males and females. Considering their close phylogenetic relatedness, it is highly likely that the internal anatomy of *S. nobilis* is almost identical to the one of Black widows of the genus Latrodectus, which have been studied into more details (Smith et al., 1969; Whitehead and Rempel, 1959).

The diet of *S. nobilis* includes insects, other spiders, and occasionally small vertebrates, which raises concerns over potential ecological impacts on native fauna (Dugon et al., 2023; Dunbar J. P. et al., 2018; Dunbar et al., 2022a). *S. nobilis* venom has been studied at the proteotranscriptomic level, demonstrating the presence of α-latrotoxins and, all of which are also present in the venom of black widows. (Dunbar et al., 2020a). While the immediate medical threat posed by *S. nobilis* remains limited, public perception, its invasive status (Rayner et al., 2022) and its potential to disseminate antibiotic resistant bacteria (Dunbar et al., 2020b), justify further biological investigations. In that sense, studying the lipidome and the metabolome of *S. nobilis* may offer promising insights. As an emerging species of medical and ecological relevance, understanding its cuticular chemistry and organ-specific lipids could provide further insight into its adaptability, diet, and interactions with its environment. It could also help distinguish *S. nobilis* from native species at the molecular level, and reveal biochemical pathways linked to its success as an urban-adapted predator within its new geographical range.

Metabolomics or lipidomics?

Metabolomics is a rapidly evolving field in the realm of systems biology that focuses on the comprehensive analysis of small bioorganic molecules, known as metabolites. Metabolites encompass a wide array of compounds, such as amino acids, lipids, sugars, and organic acids, which play critical roles in various biochemical pathways. Sometimes described as a subfield of metabolomics (Belhaj et al., 2021), lipidomics is increasingly recognized as a standalone field due to the complexity and functional diversity of lipids (Shevchenko and Simons, 2010). Lipidomics is defined as the large-scale study of pathways and networks of cellular lipids in biological systems including lipid composition, metabolism and function (Wenk, 2005). Together, metabolomics and lipidomics provide valuable insights into the dynamic molecular processes occurring within cells, tissues, and organisms, making it an essential tool for understanding the molecular underpinnings of physiological and pathological conditions (Wishart, 2019; Yang and Han, 2016).

Mass spectrometry and liquid chromatography as gold standard techniques in lipidomics

Mass spectrometry (MS) has emerged as one of the most powerful and versatile analytical techniques for molecular research (Alseekh et al., 2021). MS allows for the accurate determination of the mass-to-charge ratio (m/z) of various ions, providing information about the elemental composition (high-resolution mass spectrometry) and structure of molecules (tandem mass spectrometry or MS/MS). In the context of lipidomics, MS is particularly valuable due to its ability to work in negative and positive polarities, to detect a wide range of lipids, from small one (fatty acids) to larger ones (triglycerides, ceramides . . .), with high sensitivity and selectivity. Tandem mass spectrometry is intensively exploited to elucidate with increased certainty, structures of molecules and even isomers, a real bottleneck in lipidomics.

Liquid chromatography (LC)-MS/MS is widely used in lipidomics, and more globally in metabolomics, because it offers high sensitivity, excellent selectivity, and the ability to separate and identify a broad range of metabolites in complex biological samples, even polar and large ones.(Hajnajafi and Iqbal, 2025). However, a major drawback of LC-MS/MS is that it requires extensive sample preparation and extraction, which destroys the spatial distribution of molecules within each organ. As a result, information about their localization is lost. This limitation is overcome by mass spectrometry imaging (MSI), which allows for the direct, label-free visualization of molecular distributions across tissue sections while preserving spatial information(Abu Sammour et al., 2023).

Insights gained from mass spectrometry imaging in that context

Focusing on arthropods, MSI has been used to analyze metabolites in insect tissues, enabling visualization of the composition, abundance, and spatial distribution of endogenous and exogenous metabolites (Yang et al., 2020). MSI has additionally been employed for the rapid and accurate identification of various arthropod species, including mosquitoes and ticks, which are all vectors of human diseases (Sánchez-Juanes et al., 2022). Furthermore, MSI has been applied to map the architecture of animal toxin systems, generating molecular images at the micrometer scale to spatially map toxin compositions in producing organisms (Damm et al., 2025). Although full body MS imaging has already been proposed for rat pharmacokinetics

(Stoeckli et al., 2007) and for methodological aims on zebrafish (Ma et al., 2022; Stutts et al., 2020), the knowledge about organ-specific lipidome and metabolome of other animals, including spiders, is very limited (Khalil et al., 2015; Khalil et al., 2017). This is mainly because the size of most spiders represents a significant bottleneck, but also because obtaining traditional histological sections from mid-sized spiders such as *Steatoda* presents significant challenges due to their unique anatomy and tissue organization.

Indeed, unlike vertebrates, whose bodies contain large, soft internal organs embedded in tissue-rich matrices, spiders possess a compact and highly specialized body plan dominated by non-cellular structures and fluid-filled cavities (Foelix, 2010; Propistsova et al., 2023). Indeed, the exoskeleton of spiders is made of chitin, a hard and chemically resistant polysaccharide. This rigid outer cuticle encases most of the body and acts as a protective shell, making it difficult for fixatives such as formalin to penetrate the tissues uniformly. As a result, internal structures may not be adequately preserved during the slicing, leading to poor morphology or tissue degradation during processing (Polilov et al., 2021). Secondly, the internal anatomy of spiders is not dominated by solid tissues but rather by the hemocoel, a large cavity filled with hemolymph. Most internal organs are suspended within this cavity and are often fragile or loosely organized (Supplementary Figure S2). Another complicating factor is the small size and dense packing of functionally diverse organs within the cephalothorax and abdomen. Organs are often intertwined or stacked closely together with minimal connective tissue separation, making it difficult to isolate specific systems in sections or to orient the specimen predictably (Propistsova et al., 2023). Finally, spiders have numerous air-filled tracheae and book lungs (2 or 4), which can easily collapse. These structures are difficult to preserve in a way that maintains their three-dimensional architecture, and their presence often leads to inconsistent sectioning and tissue fragmentation. Although previous studies have established methods for fixative infusion through the cuticle of decapod crustacean larvae to preserve internal organ architecture prior to sectioning (Torres et al., 2021), applying such approaches to spiders may prove more challenging due to their more flexible and resistant cuticle. For these reasons, the first part of the study focuses on the sample preparation strategy we employed to get histological slices suitable for mass spectrometry imaging.

From another perspective, a key strength of MSI lies in its ability to analyze organs at the molecular level without micrometer-scale laser microdissection, preserving anatomical context and revealing lipidomes within structures like silk glands or the digestive tract. MSI resolution is crucial for the quality and interpretability of spatial molecular data. For a spider around 1 cm long, the number of pixels depends on the lateral resolution: at 100 µm, ~10,000 pixels provide a general organ-level view; at 10 µm, over 1 million pixels can reveal cellular or glandular features. While higher resolution offers finer molecular maps, it also increases acquisition time, data size, and computational load. Thus, selecting resolution involves balancing spatial detail with data throughput, especially for small, densely packed organisms like spiders.

Moreover, each pixel is associated with a full mass spectrum, made of thousands of peaks, increasing even more the complexity of the datasets that finally require advanced computational tools for interpretation. Data processing appears then as a major bottleneck in MSI studies (Krijnen et al., 2024). Reducing the difficulty of data treatment becomes highly appealing. Strategies such as data compression, region-of-interest (ROI) selection, segmentation or machine learning-based feature extraction help simplify analysis without compromising biological insight (Burguet et al., 2024; La Rocca et al., 2021). Streamlining data workflows not only accelerate interpretation but also makes high-resolution MSI more accessible and scalable, especially when applied to intricate samples like whole-body spider imaging.

Moreover, given the molecular diversity detected across different organs and tissues, traditional peak annotation can be challenging even with the help of exact mass measurements and various databases of metabolites. Kendrick Mass Defect (KMD) approach offers a particularly elegant solution for managing complex MSI data (Kune et al., 2019). This approach is based on a pioneering study made in 1963 by Edward Kendrick who proposed an identification of organic compounds based on the 2D projection of their atomic composition (Kendrick, 1963). KMD analysis enables the grouping of chemically related compounds (e.g., lipid series or metabolites) by aligning them based on repeating units (like CH_2), simplifying spectral interpretation. As a result, KMD scatter plots, constructed by plotting the KMD in function of the *m/z*, becomes a powerful tool to reveal biochemical patterns and tissue-specific lipidomic and metabolomic signatures in spiders, even when reference databases are limited. Applied to MSI, this method allows researchers to identify molecular families and track their spatial distribution throughout the body.

In this study, we explore the use of mass spectrometry imaging to investigate the anatomy of *S. nobilis* based on its lipidome. The unique anatomical features of arachnids, such as their rigid exoskeleton, fluid-filled body cavities, and densely packed, delicate internal organs, pose significant challenges to traditional histological and molecular approaches. MSI, by enabling spatially resolved molecular analysis without the need for dissection or labelling, offers a promising alternative. However, imaging an entire organism, such as a 1 cm-long spider, at high resolution generates vast and complex datasets. Strategies that simplify data treatment, such as optimized resolution choices and dimensionality reduction, become essential. Notably, the integration of Kendrick Mass Defect (KMD) analysis further enhances interpretability by allowing the visualization and classification of molecular families across the body. Altogether, these methodological advances pave the way for a detailed, spatially resolved exploration of the spider's organ-specific metabolome. More globally, the development of full-body mass spectrometry imaging (MSI) represents a significant advancement in the field of spatial lipidomics, enabling comprehensive, untargeted biochemical mapping across entire organisms. This approach opens new avenues for investigating the molecular architecture of small biosystems and animal models, particularly those that have historically been inaccessible to detailed biochemical analysis due to limitations in spatial resolution, sample size, or tissue complexity.

Materials and methods

Spider sourcing

S. nobilis specimens were collected at night on street furniture (e.g., metal fences, stone walls) and tree trunks in the suburban areas

of County Dublin in Ireland, using forceps and 50 mL Falcon tubes. Once collected, specimens were maintained at room temperature and lightly misted with water every 2 days until being processed.

Sample preparation

Spiders do not behave like homogeneous, solid biological tissues during freezing, which makes standard cryosectioning techniques extremely challenging. To overcome these issues, one solution is to perform whole-body cryoembedding followed by careful cryosectioning at low temperatures, which preserves delicate structures better. Traditionally, biological frozen tissues are embedded in tissue Tek OCT (Optimal cutting temperature), a soluble resin matrix often used for the imaging of individual organs. This approach is however not suitable for spiders due to their hard exoskeletons and the lack of internal support structure. It additionally may disrupt the MS spectra by contaminating the samples and adding polymer-like signals that may mask metabolite ions. Ice embedding has also been considered but does not stabilize the spider body enough. Ice is very fragile during sectioning as well and tends to chip with hard structures like cuticles. Instead, we have chosen to work with gelatin embedding because it represents an excellent compromise between structural preservation and compatibility with downstream MS analyses. Live specimens were anesthetized using a CO2 chamber, incorporated into melted gelatin 20% w/v. Gelatin powder (VWR 24350.262, CAS 9000-70-8) was dissolved in ultrapure water by heating the solution to 40 °C under continuous stirring using a magnetic stirrer), flash-frozen in isopentane (Sigma-Aldrich, CAS 78-78-4) pre-cooled with liquid nitrogen. Frozen gelatin blocks were mounted in a cryostat set between −18 °C and −20 °C and sectioned at a thickness of 14 µm. Slicing parameters are the following: blade angle (10°–15°), temperature (−10 °C to −15 °C). The obtained sections were then directly transferred either to classical slides for standard histological staining or to indium tin oxide (ITO)-coated conductive glass slides (Bruker Daltonics), dedicated to MSI analysis. To visualize the tissue topology and assist in the anatomical assignment of mass spectrometry signals, hematoxylin and eosin staining was performed on cryosections serial to those used in MSI analysis. Slides underwent sequential rehydration in ethanol and water: absolute ethanol (4s), 90% ethanol (4s), 70% ethanol (4s), and distilled water (4s). Samples were then stained in Gill III hematoxylin for 2 min, followed by 0.1% HCl dip (2s), rinsing under running tap water (3 min), eosin staining (1 min), distilled water rinse (4s), and dehydration through ascending ethanol series: 70% ethanol (4s), 90% ethanol (4s), and absolute ethanol (4s). For MSI, the MALDI matrix was applied using the SunCollect (SunChrom, Germany) with the Z-axis set to its minimum height to ensure optimal deposition on the spider surface. A solution of α-cyano-4-hydroxycinnamic acid (CHCA) at 5 mg/mL in acetonitrile/water/trifluoroacetic acid (ACN/H_2O/TFA, 70:29.9:0.1, v/v/v) was used. The sprayer operated with an X:Y stage speed of 1540 mm/min. A total of 12 matrix layers were deposited with a progressive flow rate: the first layer at 5 µL/min, the second at 10 µL/min, the third at 15 µL/min, and layers four through twelve at 20 µL/min. This gradual increase in flow rate ensured homogeneous matrix coverage and minimized analyte delocalization.

Mass spectrometry imaging

MSI data were acquired on a ESI/MALDI solariX xR 9.4T instrument (Bruker Daltonics, Bremen, Germany) using a mass range of 200–1400 m/*z* in amplitude mode and a file size of 2M time domain. This FT-ICR mass spectrometer is fitted with a high field supraconductive magnet, dynamically harmonized ICR (ion cyclotron resonance) cell, and 1ω detection capable of extreme mass resolving power (FWMH of minimum 1 million at *m*/*z* 400) and an expected mass accuracy often better than 1 ppm). In MSI experiments, such performances are hardly met due to space charge, even in FT-ICR Mass Spectrometry (McCann et al., 2021), or again the constraint imposed by the file size. In the current study, the targeted conditions were a FWMH ±300,000 @ 400 *m*/*z*, and an accuracy of 3 ppm. For each mass spectrum (pixel) of the image, the MALDI laser beam focus was set to "small" to reach a lateral resolution of the pixel's image around 50 µm. Twenty laser shots were performed at a repetition rate of 200 Hz on each pixel. The laser power was tuned to produce a stable ion current between pixels as described elsewhere (Tiquet et al., 2022). The mass spectrometer was externally mass calibrated from 200 m/*z* to 1400 m/*z* using red phosphorus suspension in pure acetone and spotted directly onto the slides. The laser power was tuned to produce a similar number of ions than those obtained from the sample. FlexImaging v5 (Bruker Daltonics, Bremen, Germany) software was used to design the imaged areas for MALDI-MSI acquisition (outlines of the spider), using a pixel step size for the surface raster set to 50 µm.

Data analysis

Imaging raw data files were first exported as imzML files (i.e., standard open format for MS file (Schramm et al., 2012)). MSI data was recalibrated (in the m/z dimension) using the method described by La Rocca and colleagues (La Rocca et al., 2021) La Rocca et al. The recalibration function was applied with the following parameters: st 0.0003, -tl 0.02, and -lm 0.002. Prior to recalibration, the MSI dataset was exported in centroid format and submitted to METASPACE for annotation using the following databases: HMDB v4, LIPID MAPS (2017-12-12), ChEBI (2018-01), and CoreMetabolome v3. A list of theoretical m/z values was compiled by selecting all annotated features with a false discovery rate (FDR) ≤ 10%, and these values were used as reference peaks for the recalibration process. After recalibration, the MSI data were re-submitted to METASPACE and annotated using the same databases and parameters. For Kendrick mass defect analysis, the mzML file was converted to ICK files and processed with our in-house Python software, MSKendrickFilter (MSKF) developed by Kune et al. (2019). Profile MS data was converted into centroid MS data using an implemented peak picking algorithm. An intensity threshold, a peak prominence, and a peak distance (number of data point between peaks) of, respectively, 200,000 cps, 0 and 5 were considered for this peak picking. Centroided MSI data was realigned to reduce centroid *m*/*z* value variations. Kendrick reference used in MSKF for lipids was -CH_2-, and the Kendrick mass (KM) and Kendrick mass defect (KMD) were computed according to Equations 1, 2, where *m*/*z* is the measured mass-to-charge ratio of the ions using IUPAC scale.

$$KM = \frac{m}{z} \cdot \frac{KR_{nominal\ mass}}{KR_{exact\ mass}} = \frac{m}{z} \cdot \frac{14.00000}{14.01565} \quad (1)$$

$$KMD = round\,(KM) - KM \quad (2)$$

The program allows for different types of rounding for Equation 2: either rounding down to the nearest lower integer (KMD ranging from −1 to 0), rounding up to the nearest higher integer (KMD ranging from 0 to 1), or rounding to the nearest integer (KMD ranging from −0.5 to 0.5). In this work, rounding up was chosen as it provides a better distribution of lipid KMD values near the center of the KMD plot. The KMD filtering was applied to retain only the information linked with lipids of interest, by restricting the dataset to theoretical KMD values between 0.10 and 0.50 and m/z between 200 and 1,400. This range was chosen because it concentrates the majority of relevant lipid signals while minimizing noise and unrelated features. The MSKF software allows the metabolite and lipid signals to be grouped by classes, based on their KMD value (since ions varying only by -CH_2- units have the same KMD value), with a given KMD tolerance. In the case where two or more lipid classes have a similar KMD value, a second algorithm is applied to group the lipid signals based on mass difference. This algorithm groups all ions whose mass difference corresponds to a given formula, here -CH_2- with a given relative mass tolerance (2 ppm). It results in a set of potential lipid series that can be identified by database research. Moreover, MSKF allows for a simplification of the KMD plots by transforming the signal series coming from different isotopes or cation adducts of the same species into the protonated monoisotopic lipid series (Hustin et al., 2022).

Database search and metabolite identification

Filtered data were analyzed by MSKF to isolate the lipid classes and the metabolites present in the spiders. In a second round, the LIPID MAPS structure database (Conroy et al., 2024), METASPACE (Palmer et al., 2017) and Human Metabolome DataBase (Wishart et al., 2022) were interrogated for potential identification. The article strictly follows the lipid nomenclature defined by LIPID MAPS, adopting a structured and standardized classification of lipid categories, classes, and subclasses.

Results

Quality of the slices using gelatin embedment

As discussed in the introduction, spiders pose a challenge for histological slicing due to their rigid chitinous exoskeleton, small size, and internal anatomy composed largely of hemolymph-filled cavities with fragile and loosely organized organs. We have observed that the presence of hair on spider's cuticle could prevent the embedding solution from making uniform contact with the entire surface of the sample, leading to poorly coated areas and potential air pockets, which in turn may cause distortions in the tissue sections.

Instead of traditional methodologies, we have chosen to work with gelatin embedding because it represents an excellent compromise between structural preservation and compatibility with downstream MS analyses. To obtain high-quality anatomical sections of *S. nobilis* suitable for mass spectrometry imaging (MSI), we developed a protocol combining gentle anesthesia, cryoembedding, and controlled sectioning (Figure 1). The procedure begins with the anesthesia of live spiders using carbon dioxide (Figure 1A). Individual spiders were placed in small transparent plastic bags and exposed to a CO_2 flow for a few minutes until complete immobilization. This method provided a quick (and reversible) anesthetic effect, minimizing stress and preserving tissue integrity. Once anesthetized, the spiders were positioned in small metal embedding molds containing a thin layer of melted gelatin. Melted gelatin (10%–20%) was subsequently poured over each specimen to embed their entire body (Figure 1B). The molds were kept on ice to ensure rapid solidification of the gelatin matrix while maintaining the anatomical posture of the specimen (Figure 1C). After the gelatin had sufficiently solidified, the entire mold was rapidly frozen by immersion in isopentane pre-cooled with liquid nitrogen (Figure 1D). This step ensures fast freezing while avoiding excessive ice crystal formation, which could compromise tissue morphology. The resulting frozen gelatin blocks were then mounted on the specimen holder of a cryostat typically set between −18 °C and −20 °C (Figure 1E). The embedded spiders were sectioned at a thickness of 14 µm, a compromise between anatomical visibility and MSI compatibility (Figure 1F). Sectioning was performed with careful adjustment of the blade angle, temperature, and cutting speed to accommodate both the rigid exoskeleton and the soft internal structures stabilized by the gelatin matrix. The blade angle was set between 10° and 15° to ensure optimal alignment with the sample surface. The cryostat chamber temperature was maintained between −10 °C and −15 °C, with the exact setting chosen based on the friability of the specimen. Cutting speed was adjusted manually during sectioning, slowing the feed when increased resistance was encountered (exoskeleton) to prevent tearing or compression of the internal organs and structures. The obtained sections, mostly preserving both the external morphology and internal organization of the spider body (Figure 1G), were then directly transferred either to classical slides for standard histological staining or to indium tin oxide (ITO)-coated conductive glass slides for downstream MSI acquisition.

The results of the histological staining performed on sagittal sections of both male and female *S. nobilis* specimens are presented in Figure 2. In female spiders, histological contrast was particularly effective in revealing the internal organization of the body (Figure 2B). Several anatomical regions, including the brain, digestive tubules, and reproductive structures such as the ovaries, were readily distinguishable and could be reliably annotated based on tissue morphology. In contrast, the histological interpretation of male spider sections proved more challenging. Owing to the male's significantly smaller size, approximately half that of the female, its internal organs are more compact, making their localization and identification considerably more difficult (Figure 2C). This anatomical constraint limits the resolution and clarity of histological features in male sections. Although classical staining provides valuable anatomical insight, it offers limited information about tissue function. One of the aims of this study is to demonstrate that MSI can not only improve organ localization but also reveal

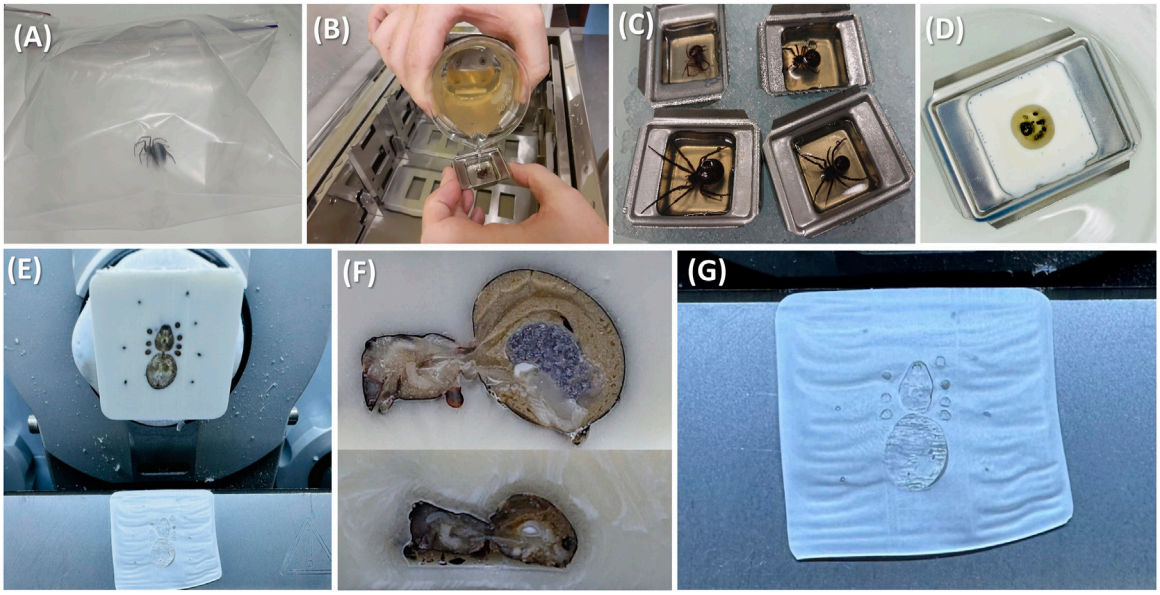


FIGURE 1
Sample preparation workflow allowing the production of histological spider slices suitable for MSI. **(A)** Spiders are anesthetized using carbon dioxide. **(B)** Melted gelatin is poured over the spider, then **(C)** kept on ice. **(D)** The gelatin is solidified using a bath of isopropanol cooled down with liquid nitrogen. **(E)** The embedded spider is cut in thin slices (14 µm) using a cryostat. **(F)** Female (top) and male (bottom) anatomy can already be discovered at the histological level. MSI aims to provide a molecular view of these structures. **(G)** Slices are ready to be covered by the suitable MALDI matrix and introduced in the mass spectrometer.

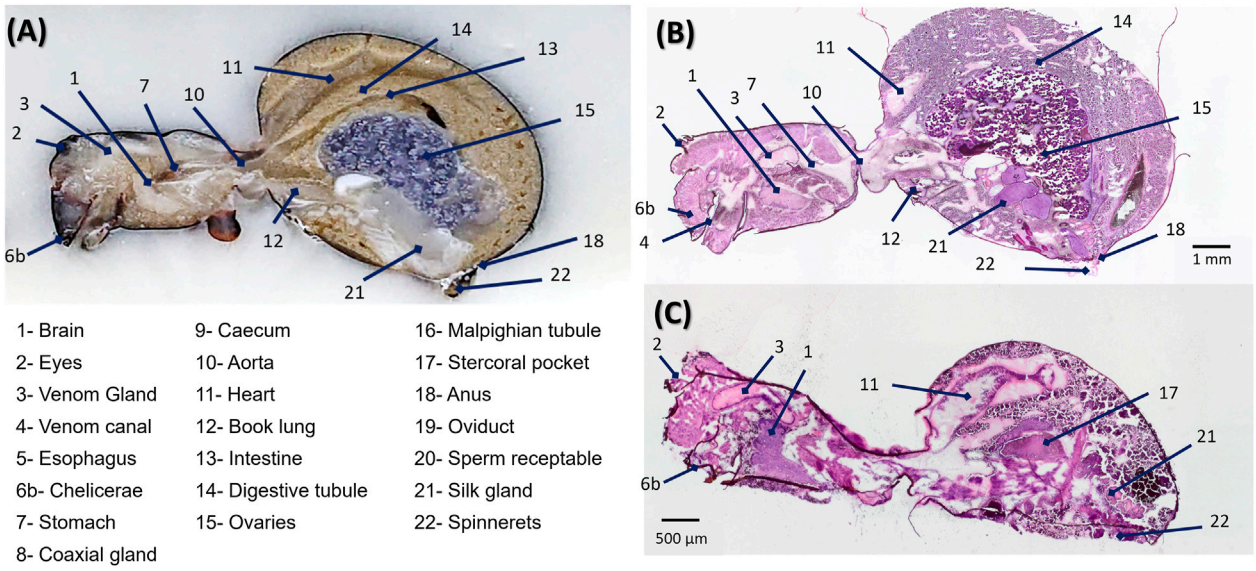


FIGURE 2
Internal anatomy of a female *Steatoda nobilis* spider. **(A)** Annotation of the internal organs within the prosoma and opisthosoma of an adult female *S. nobilis*. Major anatomical features include the central nervous system, digestive tract, silk glands, reproductive system, and circulatory system. By Ryan Wilson and Kimmel (2010). Histological staining of **(B)** female and **(C)** male specimen of *S. nobilis*. Incomplete organ identification can be made based on tissue morphology, especially for the smaller male, where the task is even more challenging.

organ-specific metabolic activity. Given the greater anatomical clarity and the presence of metabolically relevant female-specific organs, such as the ovaries, we chose to focus the remainder of the MSI analysis on female specimens exclusively.

The slides dedicated to mass spectrometry imaging were gently air-dried in a desiccator at room temperature to remove surface moisture while avoiding shrinkage or delocalization of analytes. The classical strategy for acquiring the image was then applied (see

Supplementary Figure S3). To summarize, matrix application was performed with an automated spray-coating system (SunCollect, Sunchrom) to deposit a uniform layer of α-cyano-4-hydroxycinnamic acid (CHCA), a matrix suitable for MALDI analysis of small molecules and phospholipids. After matrix deposition, the slides were introduced into the MALDI-FT-ICR mass spectrometer (SolariX, Bruker) and a raster scan was configured over the tissue surface, with a laser spot size of 50 µm. Each pixel yielded a full mass spectrum, and thousands of spectra were recorded across the tissue surface, allowing for the reconstruction of ion distribution maps. Spectral data were processed using dedicated MSI software (SCiLS©), where ion images are typically generated for selected *m/z* values or processed by segmentation. The resulting molecular maps were overlaid on optical images of the section to allow direct anatomical correlation between chemical signals and internal structures of the spider.

Data treatment by Kendrick mass defects

The high-resolution mass spectrometry imaging data acquired with MALDI-FT-ICR yielded rich spectral datasets containing thousands of *m/z* features distributed across the anatomical sections. Studying all these peaks one by one to extract biologically meaningful information is time consuming and poorly efficient. Dedicated computational tools, such as spectral alignment and normalization to ion image reconstruction, multivariate analysis, and molecular annotation, are then used to assist the researchers in this complex and challenging task.

While classical data analysis approaches such as unsupervised segmentation or principal component analysis (PCA) are effective for highlighting spatial heterogeneity across tissue sections, they often provide limited chemical interpretability, especially in complex, untargeted datasets. These methods cluster signals based on overall variance or spatial patterns, but do not inherently group chemically related species, making it challenging to distinguish molecular families or detect systematic modifications. In contrast, Kendrick Mass Defect (KMD) plots offer a chemically driven strategy that enables the recognition of homologous series and molecular families by normalizing mass values based on repeating units such as CH_2, convenient for lipids. By reprojecting the data in KMD space [[KMD = f(*m/z*)], chemically related ions align along horizontal lines or clusters, even when their spatial localization is diffuse or variable. In this approach, identifying a single molecule within a homologous series enables the annotation of all related compounds aligned along the same horizontal line. This is possible because chemically related species share the same Kendrick mass defect, allowing for rapid and structure-informed propagation of identification across the dataset, assuming that molecules aligned horizontally (e.g., in a Kendrick mass defect plot) belong to the same structural lipid family, an assumption generally valid, though the propagation strategy should be applied with caution.

Each of the acquired MSI datasets was converted from the original ".imzML" format obtained from the acquisition software SCiLS (Bruker) into ".ick" files using the dedicated export tool integrated in the home made MSKendrickFilter (MSKF) software (*see Experimental Methods*). This step enabled visualization of the data in Kendrick Mass Defect (KMD) vs. *m/z* plots. The transformation was performed using CH_2 as the base unit, a common strategy to align structurally related metabolites, particularly lipid series.

The result obtained for a female *S. nobilis* is shown in Figure 3. In these plots, spectral peaks are represented as individual dots, which tend to cluster into distinct "clouds" or lines, reflecting families of chemically similar molecules. The image reconstructed from the total ion current (TIC) does not present any structure or contrast either (Figure 3A). This is mainly because the average spectrum is dominated by matrix-related ions present across large areas of the sample or even in background regions (Figure 3B). These undesired signals, however, do not align horizontally in our Kendrick's plots, as it has been constructed with a CH_2 base unit, which is not repeated in CHCA matrix ions. Therefore, they can be easily recognized due to their tendency to form characteristic and recurrent clusters and excluded from the signals used for image reconstruction (grey dots in the Kendrick plots of Figure 3). In the same way, lipids align very well and define a quite clear KMD-*m/z* space where most of the ions of interest will be grouped. The MS image constructed based on these signals only, already reveals the underlying structures of the spider anatomy (Figure 3C).

KMD plots for the analysis of lipid groups

After a global inspection of the full KMD plot, which offers a comprehensive overview of the molecular diversity within the sample, the analysis was refined by focusing on specific "clouds", distinct clusters of points that represent structurally related compound families. These groupings, often characterized by consistent Kendrick mass defects and coherent spatial distributions, provide a powerful means to investigate biochemical pathways, homologous series, or organ-specific lipid families with greater chemical and anatomical selectivity. Three main clouds are first investigated (Figure 4). The image associated with the compounds belonging to Cloud 1 (Figure 4A) and Cloud 2 (Figure 4B) show that they are localized in the surroundings of the spider, more precisely in the abdomen cuticle, even if anatomical structures appear lightly in the case of the Cloud 2.

Lipid identification was performed only for the most intense signals of each cloud, as the identification of the remaining compounds present in the clouds can be made at a second time, by exploiting the properties of Kendrick's mass defects plots (series and alignments). Lipid identification was done by matching the exact masses obtained with our high-resolution FT-ICR mass spectrometer to known lipid compositions, provided by LIPIDMAPS LMSD, HMDB or Metaboscape. The exceptional mass accuracy and resolving power of FT-ICR enabled precise annotation of molecular formulas, even within complex isobaric regions. When multiple elemental compositions matched the same *m/z* value, the selection of the best candidate formula was guided by the following criteria. First, the ΔM (difference between the experimental and the theoretical masses) expressed in ppm was calculated and the composition with the lowest mass error was prioritized. Second, the polarity of ionization was used to give priority to lipids that ionize well in positive mode. Third, the

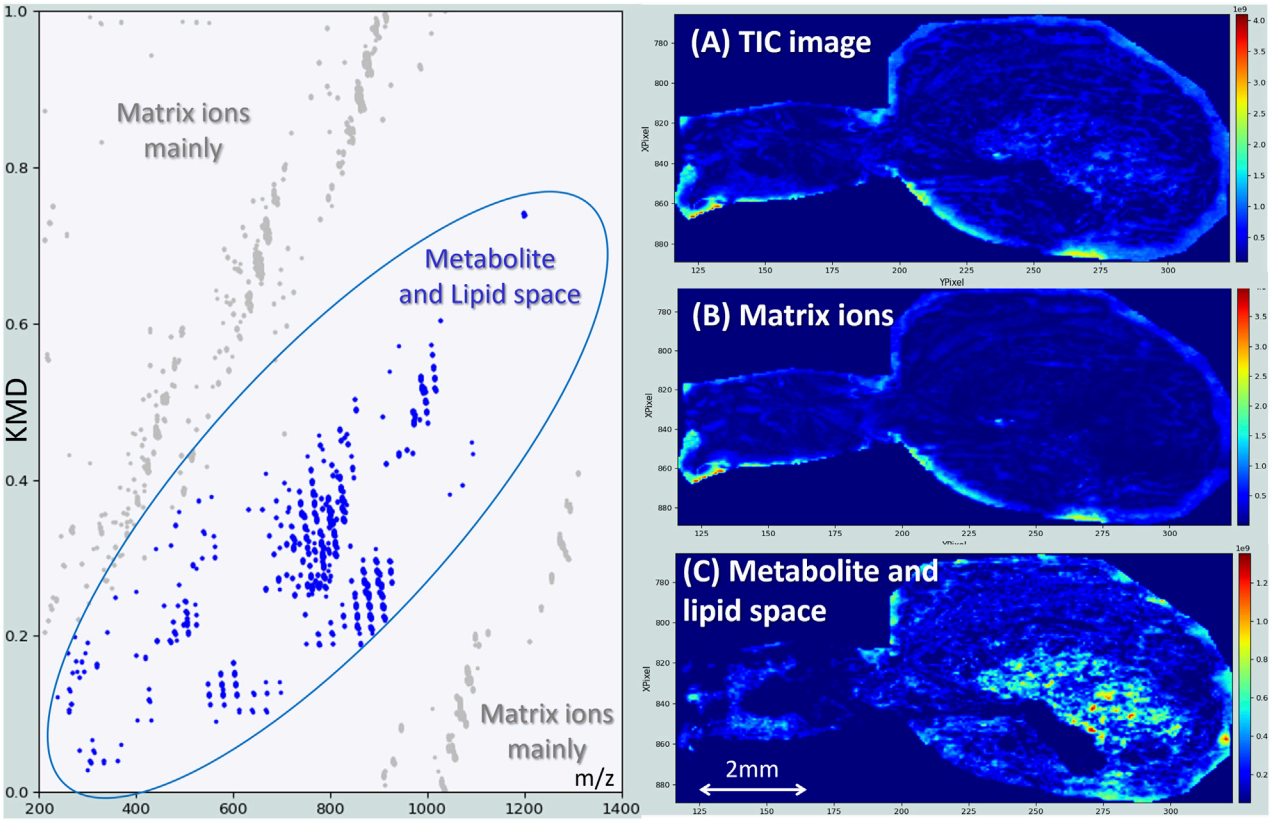


FIGURE 3
Kendrick's mass defects plot obtained for the MSI of a female *S. nobilis*. When the contribution of the signals associated with matrix clusters [greys dots, image **(B)**] is removed from the Total Ion Current [greys and blue dots, image **(A)**], fine structures of spider anatomy clearly appear [image **(C)**]. The signals responsible for these structures are represented as blue dots.

molecules with high oxidation state, *i.e.*, compositions with a high number of double bonds or oxygen atoms (indicative of highly oxidized or polyunsaturated lipids) were excluded, as they were considered biologically less likely in the context of cuticular lipid composition (Drozdowski and Boguś, 2025; Gołębiowski et al., 2012; Kaczmarek et al., 2020). This strategy ensured both analytical precision and biological plausibility in the final lipid annotations. When necessary, an additional validation by manual isotopic pattern comparison was performed (see SI 4). The annotation of the most intense signals extracted for the two clouds are summarized in Table 1.

As shown in Figure 4, most of the lipids identified in Cloud 1 and 2 are respectively triacylglycerides (TGs). The detection of such lipids in the cuticle of *Steatoda* spiders can be rationalized by their role in waterproofing, barrier function, and possibly chemical communication. Indeed, like insects, spiders must prevent desiccation, especially given their small size and high surface-area-to-volume ratio and their cuticle layer is coated with a lipid-rich surface that minimizes water loss (Trabalon and Garcia, 2021; Wang et al., 2016). TGs are particularly hydrophobic, excellent for waterproofing and may be secreted from specialized epidermal spider glands. It is important to note that monoacylglycerides (MGs) and diacylglycerides (DGs), detected as well, may additionally result from enzymatic hydrolysis of triacylglycerides (TGs) by lipases within the cuticle or hemolymph, in the context of biosynthetic turnover. Analyzing these two groups by horizontal or vertical alignment does not change their localization, suggesting their ubiquitous presence in the cuticle, even if TG 48:2 (*m/z* 825.6943) and TG 48:1 (*m/z* 827.7099) look much intense in the spinneret's region (see Supplementary Figure S5). Horizontal alignments reveal homologous series of compounds differing by repeating units (e.g., CH_2), allowing the identification of lipid families with consistent structural backbones such as for example, TG 50:3, TG 52:3 and TG 54:3. Meanwhile, vertical alignments highlight nearly isobaric species, differing in their elemental composition due to variation in degrees of unsaturation or oxidation like TG 50:1, TG 50:2 and TG 50:3. Interestingly, the presence of lipids on spider cuticle has also been associated to potential antimicrobial properties and communication pheromones that could be exploited to manage, for example, spider colonization (Chajduk et al., 2024).

Concerning Cloud 3, the most populated, the analysis is less straightforward as many compounds are not co-localized. To enhance the structural interpretation of this complex KMD space, the analysis through both vertical groupings (Figure 5A) and horizontally aligned series (Figure 5B) is proposed. This dual approach refines molecular annotation by distinguishing true biochemical series and supports structure filtering, especially in this untargeted lipidomics workflow. The annotations for the most intense signals for each strategy are summarized in Table 2.

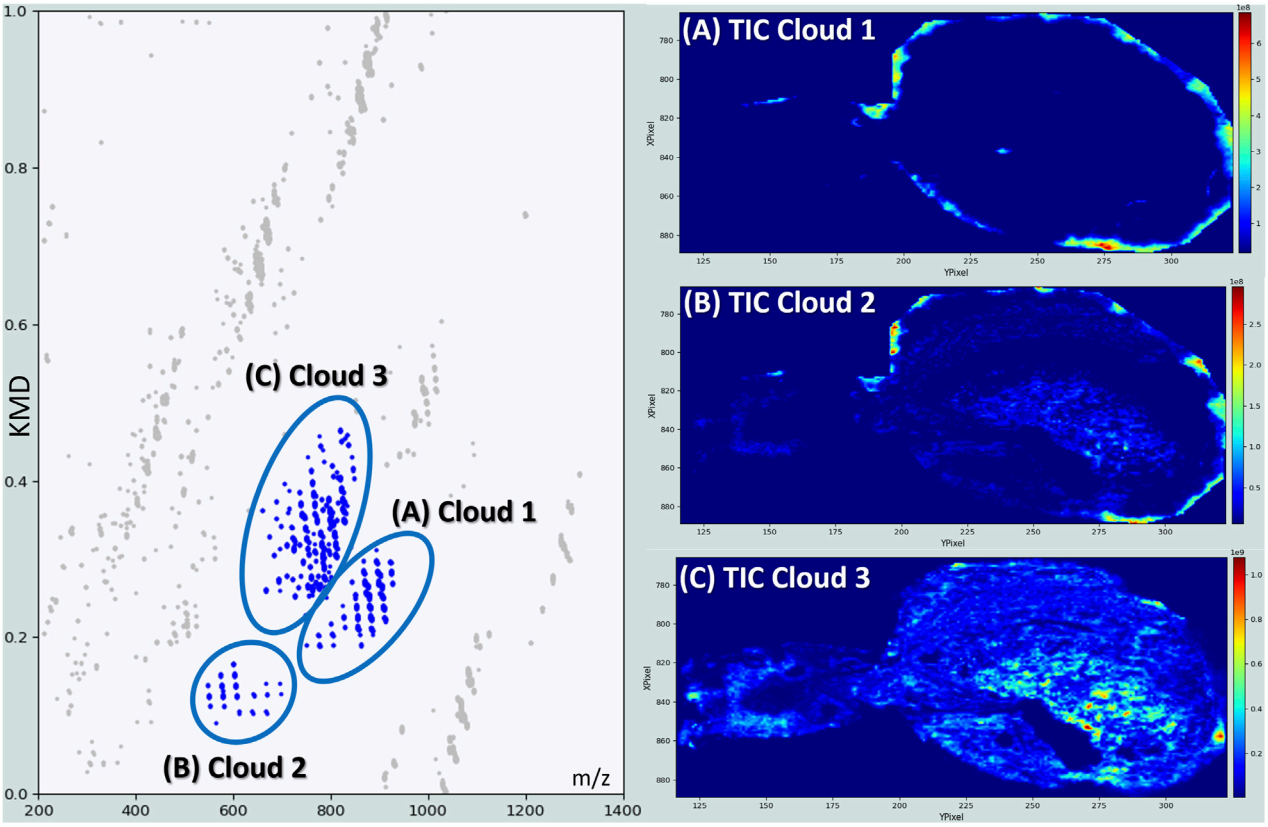


FIGURE 4
Kendrick Mass Defect (KMD) plot and corresponding spatial distributions of ion clusters obtained for the MSI of a female *S. nobilis*. Left: Kendrick Mass Defect (KMD) plot of detected ions from the full-body mass spectrometry imaging experiment. Three distinct clusters of biologically relevant ions (Clouds 1–3) are highlighted in blue. Right: total ion current images showing the spatial distribution of the ion clusters defined in the KMD plot. **(A)** TIC of Cloud 1, **(B)** TIC of Cloud 2, and **(C)** TIC of Cloud 3 and illustrating the accumulation of molecular species across different anatomical regions.

Vertical groups A1, A2 and A3

Vertical groups A1 and A2, are mainly made of phosphatidylcholine PC(34:X) and PC(36:Y), and(/or) their phosphatidylethanolamine isomers PE(37:X) and PE(39:Y). Distinguishing any isomer couple PC(X:Y) and PE(X+3:Y) would necessitate for instance additional LC-MS, as PCs are generally more polar than PEs and therefore elute differently (Triebl et al., 2017). It can also be done using MS/MS experiments as in positive ion mode, PCs are typically identified by a distinctive fragment ion at m/z 184, corresponding to the phosphocholine headgroup, whereas PEs often exhibit a neutral loss of 141 Da, associated with the loss of the phosphoethanolamine moiety (Baker et al., 2014). As MSI does not provide extensive fragmentation data, definitive distinction between PCs and PEs remains difficult at this stage. Lipids from A1 and A2 share globally the same distribution within the spider, present especially in the digestive tubules and the ovaries which reflect their roles as major membrane phospholipids (Ma et al., 2024). Their detection in the rectum/anus area may relate to membrane turnover or lipid-rich excretion pathways.

Vertical groups A4 and A5

Other interesting distributions are the A4 and the A5 where lipids are essentially detected in the ovaries. The signal mainly results from the presence of long-chain phosphatidic acids like PA (36:2), PA (36:3), PA (38:3), and PA (38:4). This makes sense as PAs are the building-block precursors to most phospholipids. In developing ovarian tissue, membrane synthesis and remodeling are very active, which requires the constant synthesis of these complex phospholipids. Moreover, lipids constitute around 30% of the dry weight of insect oocytes, participating in the embryo's development as the main source of energy (Arrese and Soulages, 2010). Intriguingly, PA (36:2) and PA (36:3) are detected on the ovary's shells under their potassium adducts $[M + K]^+$ (A4), whereas the same lipids are detected under their sodium adducts $[M + Na]^+$ in the matrix surrounding the ovaries (A5). This may suggest a different balance between sodium and potassium between the follicular fluid/membrane and the surroundings. Although direct studies on potassium and sodium concentrations in spider ovaries have not been performed yet, a study on *Drosophila* ovarian follicles tends to confirm this idea, affirming the crucial role of potassium transport in follicles during oocyte development (Bohrmann, 1991).

TABLE 1 Representative molecular assignments for ions from Cloud 1 and Cloud 2. List of selected ions detected in Clouds 1 and 2 (as defined in the Kendrick Mass Defect plot), showing their experimental mass-to-charge ratio (m/z_{exp}), signal intensity, best database match, related elemental composition, ion species, theoretical mass-to-charge ratio (m/z_{the}), and the mass error (ΔM) in parts per million (ppm). Cloud 1 is predominantly composed of DGs and TGs, while Cloud 2 is mainly associated with MGs or oxidized DGs. These lipides contribute to the spider protection against desiccation and may participate in chemical communication with her environment.

Cloud	m/z exp	Intensity	Best match	Ion elementary composition	Ion	m/z the	ΔM (ppm)
1	851.7103	5.03E+09	TG 50:3	C53 H96 O6 Na	$[M + Na]^+$	851.7099	0.5
1	853.7261	6.23E+09	TG 50:2	C53 H98 O6 Na	$[M + Na]^+$	853.7256	0.6
1	867.6846	5.37E+09	DG 53:9	C56 H92 O5 Na	$[M + Na]^+$	867.6837	1.0
1	869.7000	7.87E+09	DG 53:8	C56 H94 O5 Na	$[M + Na]^+$	869.6993	0.8
1	871.7158	2.60E+09	TG 50:1	C53 H100 O6 K	$[M + K]^+$	871.7151	0.8
1	879.7418	8.73E+09	TG 52:3	C55 H100 O6 Na	$[M + Na]^+$	879.7412	0.7
1	881.7584	7.84E+09	TG 52:2	C55 H102 O6 Na	$[M + Na]^+$	881.7569	1.7
1	893.6999	2.57E+09	TG 52:4	C55 H98 O6 K	$[M + K]^+$	893.6995	0.4
1	895.7166	7.62E+09	TG 52:3	C55 H100 O6 K	$[M + K]^+$	895.7152	1.6
1	897.7317	6.34E+09	TG 52:2	C55 H102 O6 K	$[M + K]^+$	897.7308	1.0
1	907.7737	2.61E+09	TG 54:3	C57 H104 O6 Na	$[M + Na]^+$	907.7725	1.3
2	549.4883	3.00E+09	MG 30:0 or DG O-30:0	C33 H66 O4 Na	$[M + Na]^+$	549.4853	5.5
2	551.5038	1.52E+09	MG 32:2 or DG O-32:2	C35 H67 O4	$[M + H]^+$	551.5034	0.7
2	573.4880	2.07E+09	MG 32:2 or DG O-32:2	C35 H66 O4 Na	$[M + Na]^+$	573.4853	4.7
2	575.5039	1.53E+09	MG 32:1 or DG O-32:1	C35 H68 O4 Na	$[M + Na]^+$	575.5010	5.0
2	577.5187	2.87E+10	MG 32:0 or DG O-32:0	C35 H70 O4 Na	$[M + Na]^+$	577.5166	3.6
2	601.5189	7.51E+09	MG 34:2 or DG O-34:2	C37 H70 O4 Na	$[M + Na]^+$	601.5166	3.8
2	603.5351	1.54E+10	MG 34:1 or DG O-34:1	C37 H72 O4 Na	$[M + Na]^+$	603.5323	4.6
2	605.5510	3.53E+09	MG 34:0 or DG O-34:0	C37 H74 O4 Na	$[M + Na]^+$	605.5479	5.1

Vertical group A6

Group A6 is associated with molecules predominantly detected in the cephalothorax. While their identities remain unknown, their restricted localization raises the possibility that they may contribute to neural processes within the spider's central nervous system.

Horizontal groups B1 and B3

When looking at the horizontally aligned lipids, a particularly impressive distribution is found in B1 (Figure 5). The image shows an intense detection of lipids in the spinnerets area, in the nervous central system, as the brain and the eyes are clearly highlighted, and in a zone that could be the venom glands and the chelicerae, or more likely the stomach and the mouth area. The contrast is even more marked for the B3 distribution. B1 and B3 distributions are mainly due to PC (34:X) and PC (36:X) and/or PE (37:Y) or PE (39:Y), and PA 38:4. More precisely, PA (38:3) and PA (38:4) are strongly detected in the brain compared to the other species (see Supplementary Figure S6). The prominent presence of these phosphatidic acids in the brain and central nervous system of spiders can be explained by their recognized roles in neural tissues across various species. PA (38:3) has been for example, detected in white matter areas of rat brains (Martínez-Gardeazabal et al., 2017). As already discussed, PAs serve as precursors for the biosynthesis of major membrane phospholipids, contributing to neuronal membrane integrity and, here, synaptic function (Ammar et al., 2014). Additionally, PAs act as a signaling lipid, regulating intracellular pathways pertinent to neurosecretion, membrane trafficking, and cytoskeletal dynamics processes integral to neural activity (Ammar et al., 2014). While no specific studies on spiders can be found to date, the conservation of lipid-mediated mechanisms in neural tissues suggests that PA 38:3 and 38:4 plays similar roles in arachnid neurobiology. Phosphatidic acids appear intense in the spider eyes as well. This is also in line with studies on *Drosophila* which have demonstrated, for instance, the importance of PAs in membrane dynamics essential for proper photoreceptor genesis and function (Raghu et al., 2009).

Horizontal group B4

The distribution B4 is dominated by the signal detected at *m/z* 780.5515 and associated with possibly PC (34:2) and/or PE (37:2). They are intensively detected in the heart and in the eyes of the spider, which is in accordance with the known physiological roles of these phospholipids in various species (Lim et al., 2011; Stark et al.,

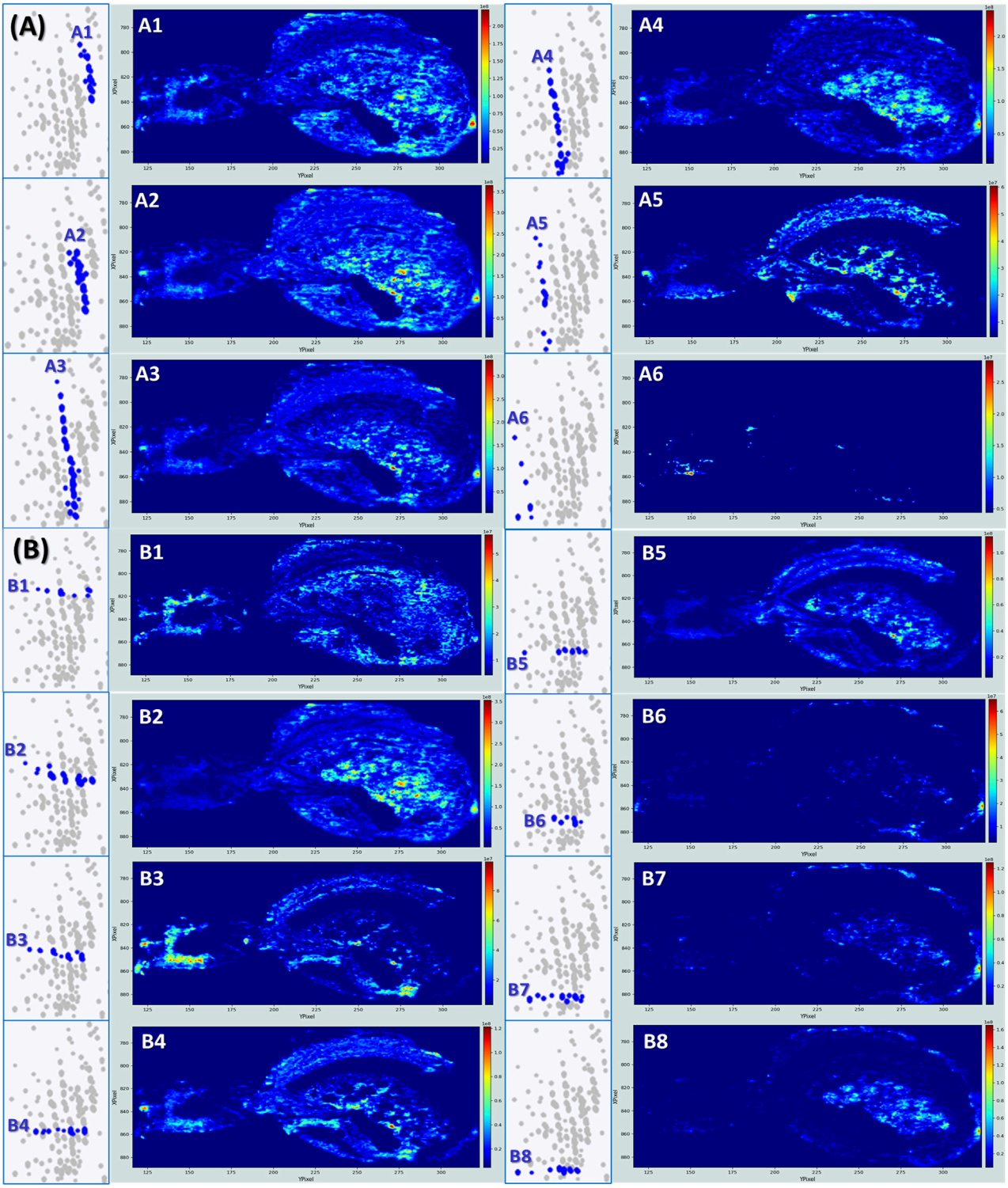


FIGURE 5
Kendrick's mass defect plots analyzed by both vertical groupings and horizontally aligned series to reveal additional distribution of lipids in the spider tissues. **(A)** Series corresponding to the vertical distribution of ions, indicative of different functional groups or degrees of unsaturation within a lipid class. **(B)** Series corresponding to the horizontal distribution of ions, reflecting homologous series with repeating structural units (e.g., CH_2).

1993; Zhao and Wang, 2020). The heart of the spider, even simple, is a highly active muscle metabolically and the importance of PCs in the cardiovascular system has been demonstrated several times. For example, a higher consumption of PCs increases mortality from cardiovascular diseases (Zheng et al., 2016). PCs and PEs additionally play an important role in the eyes, especially in the photoreceptor membrane structure, essential for ocular function. For example, PE forms a Schiff base conjugate with trans-retinal

TABLE 2 Detailed molecular assignments for ion series within Cloud 3. Table showing the experimental m/z values, signal intensities, best tentative matches, associated elemental compositions, detected ion species, theoretical m/z values, and mass errors (ΔM in ppm) for two main ion series (A and B) identified within Cloud 3 from the Kendrick Mass Defect analysis. *Ax* series design the vertical groupings, whereas *By* series indicate the horizontally aligned families.

Cloud	Series	m/z exp	Intensity	Best match	Ion elementary Composition	Ion	m/z the	ΔM (ppm)
3	A1	820.5254	6.23E+09	PC (36:4) or PE (39:4)	C44 H80 N O8 P K	$[M + K]^+$	820.5253	0.1
3	A1	822.5398	3.59E+10	PC (36:3) or PE (39:3)	C44 H82 N O8 P K	$[M + K]^+$	822.5410	1.5
3	A1	824.5559	9.95E+09	PC(36:2) or PE (39:2)	C44 H84 N O8 P K	$[M + K]^+$	824.5566	0.8
3	A1	826.5722	1.01E+10	PC (36:1) or PE (39:1)	C44 H86 N O8 P K	$[M + K]^+$	826.5723	0.1
3	A2	796.5251	5.01E+10	PC (34:2) or PE (37:2)	C42 H80 N O8 P K	$[M + K]^+$	796.5253	0.3
3	A2	798.5416	7.73E+10	PC (34:1) or PE (37:1)	C42 H82 N O8 P K	$[M + K]^+$	798.5410	0.8
3	A2	806.5667	1.53E+10	PC (36:3) or PE(39:3)	C44 H82 N O8 P Na	$[M + Na]^+$	806.5670	0.4
3	A2	808.5825	1.22E+10	PC (36:2) or PE (39:2)	C44 H84 N O8 P Na	$[M + Na]^+$	808.5827	0.2
3	A3	763.4667	1.18E+10	PA (38:4)	C41 H73 O8 P K	$[M + K]^+$	763.4675	1.0
3	A3	765.4824	1.52E+10	PA (38:3)	C41 H75 O8 P K	$[M + K]^+$	765.4831	0.9
3	A3	770.5099	3.45E+09	PC (32:1) or PE (35:1)	C40 H78 N O8 P K	$[M + K]^+$	770.5097	0.3
3	A3	780.5515	2.53E+10	PC (34:2) or PE (37:2)	C42 H80 N O8 P Na	$[M + Na]^+$	780.5514	0.1
3	A3	782.5678	3.83E+10	PC (34:1) or PE(37:1)	C42 H82 N O8 P Na	$[M + Na]^+$	782.5670	1.0
3	A3	786.6008	4.66E+09	PC (36:2) or PE (39:2)	C44 H84 N O8 P	$[M + H]^+$	786.6007	0.1
3	A4	737.4516	1.98E+10	PA (36:3)	C39 H71 O8 P K	$[M + K]^+$	737.4518	0.3
3	A4	739.4676	3.21E+10	PA (36:2)	C39 H73 O8 P K	$[M + K]^+$	739.4675	0.1
3	A4	747.4930	9.13E+09	PA (38:4)	C41 H73 O8 P Na	$[M + Na]^+$	747.4935	0.7
3	A4	749.5087	5.76E+09	PA (38:3)	C41 H75 O8 P Na	$[M + Na]^+$	749.5092	0.7
3	A4	758.5695	7.66E+09	PC (34:2) or PE (37:2)	C42 H80 N O8 P	$[M + H]^+$	758.5694	0.1
3	A4	760.5842	2.45E+10	PC (34:1) or PE (37:1)	C42 H82 N O8 P	$[M + H]^+$	760.5851	1.2
3	A5	721.4773	8.68E+09	PA (36:3)	C39 H71 O8 P Na	$[M + Na]^+$	721.4779	0.8
3	A5	723.4932	1.98E+10	PA (36:2)	C39 H73 O8 P Na	$[M + Na]^+$	723.4935	0.4
3	A5	724.4974	1.42E+09	PA (36:2) - first isotope	12C38 13C H73 O8 P Na	$[M + Na]^+$	724.4970	0.6
3	A6	666.4834	3.28E+08	?	?	-	-	>10
3	A6	682.4575	1.46E+08	?	?	-	-	>19
3	A6	692.4991	5.40E+08	?	?	-	-	> 10
3	B1	763.4667	1.18E+10	PA (38:4)	C41 H73 O8 P K	$[M + K]^+$	763.4675	1.0
3	B1	794.5095	7.04E+09	PC (34:3) or PE (37:3)	C42 H78 N O8 P K	$[M + K]^+$	794.5097	0.3
3	B1	822.5398	3.59E+10	PC (36:3) or PE (39:3)	C44 H82 N O8 P K	$[M + K]^+$	822.5410	1.5
3	B2	739.4676	3.73E+10	PA (36:2)	C39 H73 O8 P K	$[M + K]^+$	739.4675	0.1
3	B2	740.4702	1.06E+10	PA (36:2) - first isotope	12C38 13C H73 O8 P K	$[M + K]^+$	740.4709	0.9
3	B2	767.4999	3.02E+09	PA(38:2)	C41 H77 O8 P K	$[M + K]^+$	767.4988	1.4
3	B2	770.5099	3.45E+09	PC(32:1) or PE 35:1)	C40 H78 N O8 P K	$[M + K]^+$	770.5097	0.3
3	B2	798.5405	7.73E+10	PC (34:1) or PE 37:1)	C42 H82 N O8 P K	$[M + K]^+$	798.5410	0.6
3	B2	804.5505	7.73E+10	PC (36:4) or PE 39:4)	C44 H80 N O8 P Na	$[M + Na]^+$	804.5514	1.1
3	B3	747.4930	9.13E+09	PA (38:4)	C41 H73 O8 P Na	$[M + Na]^+$	747.4935	0.7

(Continued on following page)

TABLE 2 (*Continued*) Detailed molecular assignments for ion series within Cloud 3. Table showing the experimental m/z values, signal intensities, best tentative matches, associated elemental compositions, detected ion species, theoretical m/z values, and mass errors (ΔM in ppm) for two main ion series (A and B) identified within Cloud 3 from the Kendrick Mass Defect analysis. *Ax* series design the vertical groupings, whereas *By* series indicate the horizontally aligned families.

Cloud	Series	m/z exp	Intensity	Best match	Ion elementary Composition	Ion	m/z the	ΔM (ppm)
3	B3	749.5087	5.75E+09	PA (38:3)	C41 H75 O8 P Na	$[M + Na]^+$	749.5092	0.7
3	B3	778.5357	1.76E+10	PC (34:3) or PE (37:3)	C42 H78 N O8 P Na	$[M + Na]^+$	778.5357	0.0
3	B3	806.5668	1.53E+10	PC (36:3) or PE (39:3)	C44 H82 N O8 P Na	$[M + Na]^+$	806.5670	0.2
3	B3	808.5823	8.95E+09	PC (36:2) or PE (39:2)	C44 H84 N O8 P Na	$[M + Na]^+$	808.5827	0.5
3	B4	780.5515	2.53E+10	PC (34:2)or PE (37:2)	C42 H82 N O8 P Na	$[M + Na]^+$	780.5514	0.1
3	B4	808.5825	1.22E+10	PC (36:2) or PE (39:2)	C44 H84 N O8 P Na	$[M + Na]^+$	808.5827	0.2
3	B5	767.5464	5.83E+08	SM (36:2; O2)	C41 H81 N2 O6 P K	$[M + K]^+$	767.5464	0.0
3	B5	782.5663	3.83E+10	PC (34:1) or PE (37:1)	C42 H82 N O8 P Na	$[M + Na]^+$	782.5670	0.9
3	B5	810.5988	1.77E+09	PC (36:1) or PE (39:1)	C44 H86 N O8 P Na	$[M + Na]^+$	810.5983	0.6
3	B6	756.5548	1.77E+08	PC (34:3) or PE (37:3)	C42 H78 N O8 P	$[M + H]^+$	756.5538	1.3
3	B6	784.5856	4.24E+09	PC (36:3) or PE (39:3)	C44 H82 N O8 P	$[M + H]^+$	784.5851	0.6
3	B7	692.4991	5.40E+08	?	?	-	-	>> 10
3	B7	758.5696	7.66E+09	PC (34:2) or PE (37:2)	C42 H80 N O8 P	$[M + H]^+$	758.5694	0.3
3	B7	786.6008	4.47E+09	PC (36:2) or PE (39:2)	C44 H84 N O8 P	$[M + H]^+$	786.6007	0.1
3	B8	760.5842	2.45E+10	PC (34:1) or PE (37:1)	C42 H82 N O8 P	$[M + H]^+$	760.5851	1.2
3	B8	774.6004	2.33E+09	PC (35:1) or PE (38:1)	C43 H84 N O8 P	$[M + H]^+$	774.6007	0.4
3	B8	788.6159	7.59E+08	PC (36:1) or PE (39:1)	C44 H86 N O8 P	$[M + H]^+$	788.6164	0.6

(Vitamin A) to form the N-RET-PE molecule, crucial for the proper functioning of the visual process and pigment photogeneration(Xu et al., 2023).

Another Cloud of dots reveals additional information about the *S. nobilis* lipidome (Figure 6). These ions are visibly mainly localized in the ovaries/eggs of the spider. However, looking in much detail, they can be classified according to two groups. The first one (Figure 6–C1) concerns lipids strongly detected in the ovaries. These molecules correspond to lysophosphatidylcholines (LPCs) and lysophosphatidylethanolamines (LPEs), phospholipids missing one of their fatty acid chains usually removed by a lipase (Table 3).

The detection of LPCs and LPEs in the ovaries and eggs of spiders can be attributed to their crucial roles in lipid metabolism and membrane dynamics during oocyte development and embryogenesis (Steinhauer et al., 2009). Although LPEs are less studied, LPCs have been identified as main lipid components of lipovitellins isolated from the eggs of the wolf spider *Schizocosa malitiosa* (Laino et al., 2013). Lipovitellins are major yolk proteins that serve as energy sources during embryonic development in oviparous species. The detection of LPCs in these lipovitellins indicates their significant role in the reproductive biology of spiders, likely contributing to membrane dynamics and lipid metabolism during oogenesis and embryogenesis.

The second series is composed of molecules mostly localized in the chelicerae area (Figure 6–C2). Chelicerae are kind of cuticular extensions involved in feeding and mechanical delivery of the venom into the prey. The signals are strongly dominated by molecules belonging much probably to the sterol family, never described in this spider organ. Considering however, that insect cuticles contain sterols (Wrońska et al., 2023) and that the mandibular glands of *Apis mellifera* secrete sterols as well (Buttstedt et al., 2023), detecting such molecules in the biting apparatus of *S. nobilis* aligns very well with its cuticular structure and its secretory function.

KMD plots for single metabolite analysis

In addition to global lipid class analysis, KMD plots can also be used to follow individual metabolites, which do not belong to lipid classes. When using a CH_2 base, compounds that are not aligned horizontally do not belong, by definition, to repeating CH_2-based compound classes, making them easily detectable within complex datasets. Figure 7 is dedicated to these compounds and shows atypical distributions for eight of them. Metabolites 1, 2, 3 and 6 display the same distribution and are detected with high intensity in particular organs, identified as the silk glands. One of these metabolites is accurately identified as hexanoylcarnitine (*m/z* 260.1854). Carnitines play a pivotal role in mitochondrial fatty acid transport and β-oxidation, allowing cells to generate ATP from fatty acids (Virmani and Cirulli, 2022). Their detection by MALDI-MSI in the silk glands of *S. nobilis* appears to be biologically coherent, as silk production is a highly energy-demanding process requiring sustained ATP supply (Xu et al., 2022). The presence of

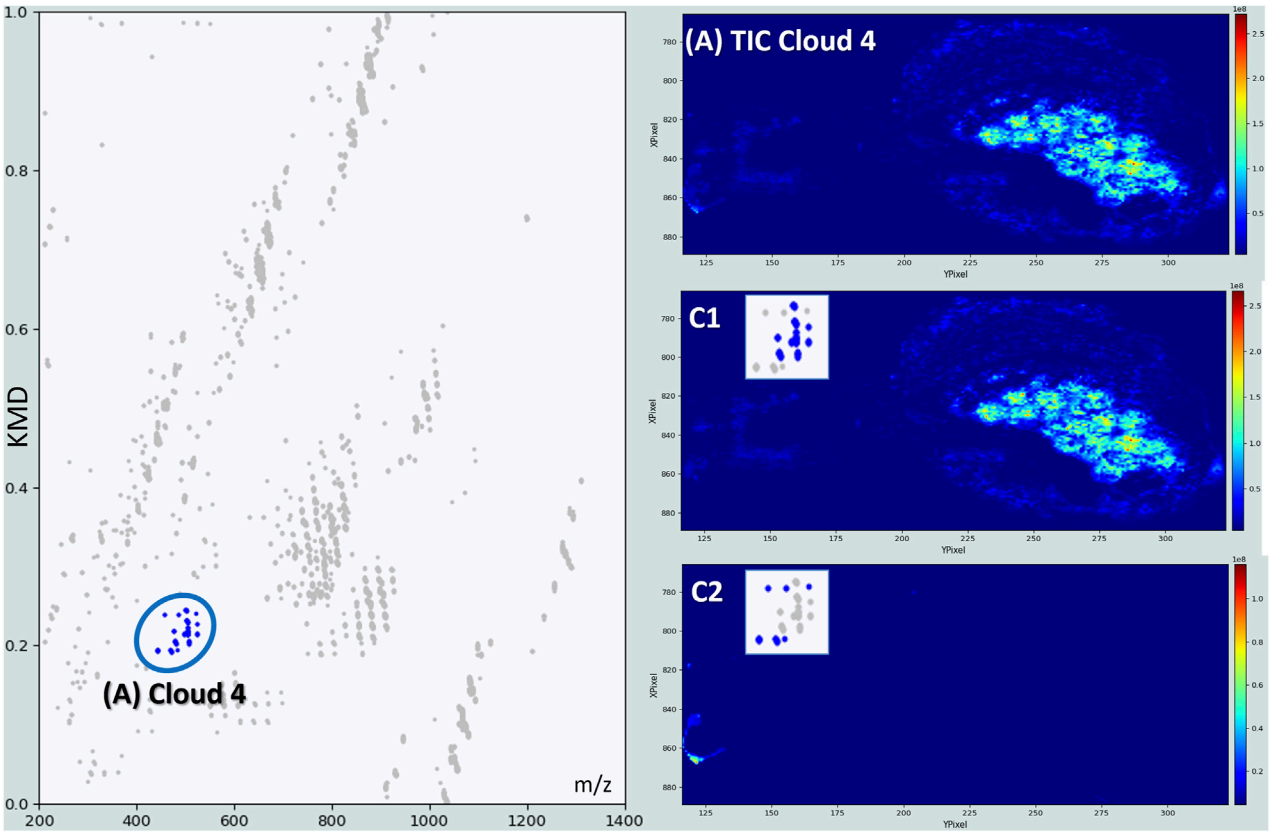


FIGURE 6
Kendrick's mass defects plot obtained for the MSI of a female *S. nobilis*. Displayed images are related to the Total ion current (Top), and to two secondary groups C1, built from ions localized in the ovaries, and C2 related to species detected in the fang area of the spider.

TABLE 3 Detailed molecular assignments for ion series within Cloud 4. Table showing the experimental m/z values, signal intensities, best tentative matches, associated elemental compositions, detected ion species, theoretical m/z values, and mass errors (ΔM in ppm) for the ions identified within Cloud 4 from the Kendrick Mass Defect analysis. Lysophospholipids are precisely located in the spider ovaries, whereas sterol molecules are most intensively detected in the chelicerae.

Cloud	Series	m/z exp	Intensity	Best match	Ion elementary composition	Ion	m/z the	ΔM (ppm)
4	C1	476.3133	6.29E+08	LPC (O-14:0)	C22 H48 N O6 P Na	$[M + Na]^+$	476.3111	4.6
4	C1	478.3289	1.81E+10	LPE (O-19:2)	C24 H48 N O6 P	$[M + H]^+$	478.3292	0.6
4	C1	500.3135	6.19E+07	LPE (O-21:2)	C26 H52 N O6 P	$[M + Na]^+$	500.3111	4.8
4	C1	502.3289	2.93E+10	LPC (O-16:1)	C24 H50 N O6 P Na	$[M + Na]^+$	502.3268	4.2
4	C1	504.3444	8.60E+10	LPC (O-16:0)	C24 H52 N O6 P Na	$[M + Na]^+$	504.3424	4.0
4	C1	506.3604	4.11E+09	LPC (O-18:2)	C26 H52 N O6 P	$[M + H]^+$	506.3605	0.2
4	C2	441.2978	5.40E+08	ST 26:2; O4	C26 H42 O4 Na	$[M + Na]^+$	441.2975	0.7
4	C2	469.3293	7.68E+08	ST 28:2; O4	C28 H46 O4 Na	$[M + Na]^+$	469.3288	1.1

hexanoylcarnitine specifically suggests the mobilization of medium-chain fatty acids, indicating that these are actively oxidized to meet the energetic demands of silk protein synthesis and secretion. While direct studies on carnitine's function in spider silk glands do not exist, the established importance of carnitine in energy-demanding processes in insects implies a similar role in arachnids (Childress et al., 1967).

The metabolite responsible for the signal 4 could match the cyclic dipeptide Leu-Phe (c-Leu-Phe), as the mass corresponds very well (*m/z* 261.1596, ΔM = 0.4 ppm). This metabolite might be endogenous, as a product of peptide and protein degradation, or exogenous, originating from microbial or environmental sources. It is located in a very specific zone, just between the ovaries/eggs and the silk glands. The identification of this organ is not straightforward

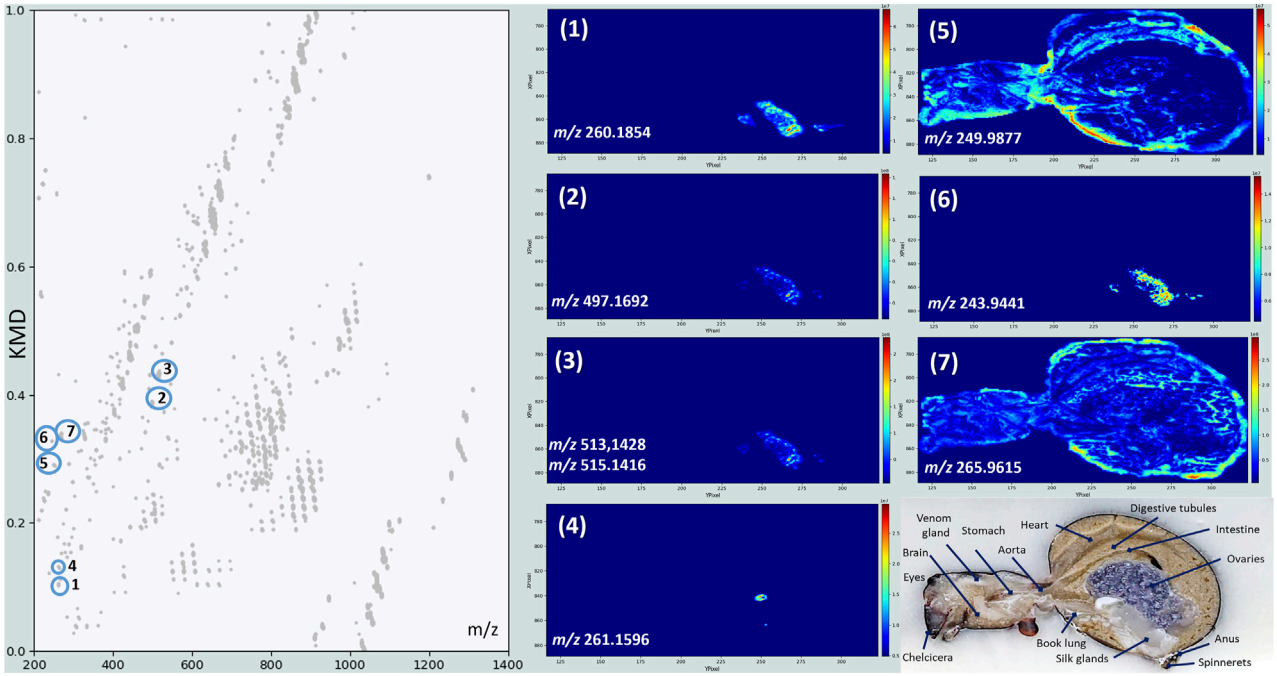


FIGURE 7
Kendrick's mass defects plot obtained for the MSI of a female *S. nobilis*. Displayed images correspond to the individual metabolites labelled in the plot. The silk glands (1–3 and 6), the oviduct (4), the pericardium (5) and the digestive tubules (7) are very well detected.

as only a few histological slices of spiders are documented. Our hypothesis is that this organ may be the oviduct of the spider, from where the eggs are expulsed before being wrapped in a silk cocoon. Metabolite 5 is intensely detected in the close surrounding of the heart, the pericardium. As the spider heart supports hemolymph circulation, pericardial cells need readily available energy (Meyer et al., 2024). The nature of the metabolite would probably support the local energy requirements near the heart or reflect the metabolic buffering role of pericardial cells in processing circulating lipid and lipid derivatives.

Metabolite 7 (*m/z* 265.9615) displays a very interesting distribution. Its signal is more intense in the interstices between the digestive tubules, giving birth to additional histological structures. Its nature is, however, still unidentified, not only because spider metabolome information is scarce, but also because this area is involved in exchange between the digestive epithelium and hemolymph (allowing for instance the nutrient uptake), and a large variety of molecules, possessing various physico-chemical properties, can therefore be present in this place.

Discussion

This study delivers the first organ-specific, spatially resolved lipidomic and metabolomic map of a spider, here the noble false widow spider, *S. nobilis*. It has been achieved through an integrative workflow combining high-resolution Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry Imaging (MALDI-MSI) with Kendrick Mass Defect (KMD) analysis. This powerful analytical framework enabled us to visualize and classify molecular species with suborgan selectivity, offering new insights into the internal chemical anatomy of an arachnid, a group of animals for which molecular-level spatial data remains virtually poorly if not nonexistent. KMD plots were powerful not only for lipid class separation but also for the detection of unusual or outlier metabolites. These plots allowed us to group chemically homologous series and to detect structural analogs even in the absence of prior annotation. The dual capability of KMD, both as a classifier and a screening tool, was crucial in this exploratory context, where spectral libraries are sparse and manual validation is limited. Importantly, the combination of KMD with FT-ICR's sub-ppm mass accuracy allowed for high-confidence elemental composition predictions, a major advantage for metobolomic studies in non-model species. Globally, given the significant biological and ecological roles that lipids and small metabolites play in invertebrate physiology, this work addresses a critical knowledge gap and introduces a versatile, transferable method for studying the functional biochemistry of non-model organisms.

The preservation of anatomical integrity through optimized cryosectioning and gelatin embedding protocols allowed us to retain spatial context during mass spectrometry analysis. This was essential for mapping hundreds of compounds across the complex internal landscape of the spider body, where organs are compact and highly specialized. The resulting molecular maps revealed organ-specific lipid signatures that correspond to physiological functions known or hypothesized from related taxa. In the outer body structures, the cuticle and spinnerets as well, triacylglycerides were dominant. These hydrophobic compounds are likely involved in maintaining water balance and structural integrity under desiccating terrestrial conditions. Their presence near silk-

producing spinnerets also suggests a potential role in facilitating silk extrusion through lubrication or structural stabilization of the silk proteins. Mono- and diacylglycerides, detected broadly in these same regions, may represent lipid turnover intermediates, possibly derived from energy mobilization or membrane remodeling in response to environmental or mechanical stimuli. The internal tissues, particularly those associated with reproduction, revealed a markedly different lipid profile. Ovarian regions showed high concentrations of phosphatidic acids (PAs), which are key intermediates in the biosynthesis of complex phospholipids. These molecules are essential for rapid membrane formation during oogenesis, and their presence supports the hypothesis of active lipid remodeling and trafficking within the developing oocytes. The spatial detection of different adduct forms of these molecules, such as $[M + Na]^+$ and $[M + K]^+$, in distinct microregions of the ovaries suggests that ionic microenvironments may influence lipid dynamics and compartmentalization, an aspect rarely captured in conventional metabolomic studies. Additional lipid species such as lysophosphatidylcholines (LPCs) and lysophosphatidylethanolamines (LPEs) were also localized in the reproductive system, including egg masses. These lysolipids are known to result from phospholipase-mediated cleavage and are implicated in signaling pathways and membrane restructuring. Their detection in reproductive tissues aligns with their previously proposed function in nutrient provisioning and yolk formation in spider embryos, as supported by earlier findings, in *Schizocosa malitiosa* for instance. In the central nervous system, comprising the brain and optic centers, phosphatidylcholines (PCs) and phosphatidylethanolamines (PEs) were intensively detected. These lipids are foundational components of synaptic and axonal membranes, and their abundance here is consistent with high turnover and maintenance demands required by sensory processing and neural communication. Moreover, the co-localization of phosphatidic acids in these regions reinforces their putative involvement in synaptic vesicle formation, signaling, and other neurophysiological roles. One of the most novel findings of this study was the accumulation of hexanoylcarnitine in the silk glands. Carnitines are involved in transporting fatty acids into mitochondria for β-oxidation, and their enrichment in silk-producing tissues strongly suggests a high energetic demand associated with silk biosynthesis. While the mechanical and biochemical cost of silk production in spiders is well known, this is the first time that direct molecular evidence of energy metabolism has been spatially mapped to the silk glands themselves. Another distinctive pattern emerged in the chelicerae and fang region, where sterol-like compounds were highly concentrated. Sterols are critical for membrane fluidity and cuticle structure, and their presence in these mechanically active venom-associated regions suggests a dual role in structural reinforcement and possibly in venom solubilization or delivery. Comparable roles have been observed in insect mandibular glands, hinting at a conserved lipid function across arthropod lineages. While many metabolites could be assigned tentative identities based on exact mass and localization, several compounds remained unannotated yet showed highly specific spatial distributions. For instance, a feature consistent with the cyclic dipeptide Leu-Phe was localized to the oviduct area, possibly reflecting proteolytic activity or microbiota interactions. Other unidentified metabolites clustered around the heart and pericardial cavity may be involved in lipid-derived signaling or energy buffering in the hemolymph system, while small molecules detected in digestive tubules likely participate in nutrient processing and absorption.

Despite these achievements, the lack of tandem MS (MS/MS) data for many ions remains a limiting factor for unambiguous structural identification. Moreover, while the instrument used in this study is capable of delivering mass accuracies typically better than 1 ppm, we observed that some detected features exhibited higher mass errors, occasionally exceeding 4 ppm. These deviations can be due to space-charge effects, which are known to impact mass accuracy in high-abundance or spatially complex samples, such as full-body sections or to low abundances of some molecules. Furthermore, due to the limited coverage of arachnid lipidomes and metabolomes in current spectral databases, the proposed molecular annotations should be considered tentative, particularly when the associated mass error is elevated. In summary, cases where ΔM exceeds 4 ppm, annotations must be considered as hypothetical and should be interpreted with caution.

Future integration of laser microdissection followed by LC-MS/MS, or coupling to ion mobility spectrometry, would provide orthogonal validation and enhance compound resolution. Nevertheless, the anatomical selectivity and mass accuracy afforded by MALDI-MSI and KMD together form a powerful platform for generating hypothesis-driven insights even in data-poor systems like spiders.

Conclusion

In conclusion, this study provides the first molecular atlas of the noble false widow spider, *S. nobilis*, capturing the spatial organization of lipids and metabolites across its internal organs. By combining high-resolution mass spectrometry imaging (MSI) with Kendrick mass defect (KMD) analysis, we demonstrate a robust and versatile approach for deciphering the molecular architecture of small, structurally complex organisms. This methodological framework enables the mapping of hundreds of molecular species in anatomical context, offering novel insights into spider physiology and organ-specific biochemistry.

Our findings represent a major advancement in arachnid lipidomics, revealing both conserved and unique lipid features in a terrestrial invertebrate. They fill a significant gap in comparative lipidomics and provide a valuable reference point for future investigations. Beyond spiders, this work sets the stage for broader applications of spatial lipidomics in other invertebrate systems, particularly those previously inaccessible to detailed biochemical analysis. By establishing a functional and anatomical link between molecular profiles and tissue identity, this study opens new avenues for exploring developmental biology, ecological adaptation, and evolutionary physiology across arthropods and beyond.

Data availability statement

The data presented in the study are deposited in the METASPACE repository, accession number 2025-08-19_22h18m37s under the name Spider_Steatoda_nob_female_CD_1.

Ethics statement

The manuscript presents research on animals that do not require ethical approval for their study.

Author contributions

DR: Data curation, Visualization, Formal Analysis, Methodology, Writing – review and editing. JD: Methodology, Conceptualization, Data curation, Writing – review and editing, Formal Analysis, Writing – original draft, Visualization. RL: Software, Formal Analysis, Visualization, Methodology, Data curation, Investigation, Writing – review and editing. AD: Formal Analysis, Writing – review and editing, Methodology, Investigation. QB: Formal Analysis, Data curation, Methodology, Investigation, Writing – review and editing. VB: Writing – review and editing, Methodology, Supervision, Visualization, Resources, Conceptualization. CK: Formal Analysis, Data curation, Visualization, Validation, Methodology, Software, Writing – review and editing. MT: Conceptualization, Validation, Methodology, Writing – review and editing, Formal Analysis. JF: Supervision, Visualization, Software, Writing – review and editing, Formal Analysis, Methodology. ED: Supervision, Writing – review and editing, Validation, Investigation, Formal Analysis, Software, Funding acquisition, Resources. MM: Methodology, Conceptualization, Formal Analysis, Investigation, Writing – review and editing, Writing – original draft, Resources, Supervision, Funding acquisition. LQ: Investigation, Writing – original draft, Software, Resources, Visualization, Funding acquisition, Validation, Data curation, Formal Analysis, Supervision, Conceptualization, Writing – review and editing, Project administration, Methodology.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement no. 731077 (EU FT-ICR MS project, INFRAIA-02 2017) from the European Union and Wallonia program FEDER BIOMED HUB Technology Support (no. 2.2.1/996) for the funding of the SolariX XR 9.4T, support from the Excellence Of Science Program of the FWO/FNRS F.R.S (Rhizoclip–EOS ID 30650620) and from the Interreg EMR project: EURLIPIDS (R8598).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2025.1658546/full#supplementary-material

References

Abu Sammour, D., Cairns, J. L., Boskamp, T., Marsching, C., Kessler, T., Ramallo Guevara, C., et al. (2023). Spatial probabilistic mapping of metabolite ensembles in mass spectrometry imaging. *Nat. Commun.* 14, 1823. doi:10.1038/s41467-023-37394-z

Alseekh, S., Aharoni, A., Brotman, Y., Contrepois, K., D'Auria, J., Ewald, J., et al. (2021). Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nat. Methods* 18, 747–756. doi:10.1038/s41592-021-01197-1

Ammar, M. R., Kassas, N., Bader, M. F., and Vitale, N. (2014). Phosphatidic acid in neuronal development: a node for membrane and cytoskeleton rearrangements. *Biochimie* 107, 51–57. doi:10.1016/j.biochi.2014.07.026

Arrese, E. L., and Soulages, J. L. (2010). Insect fat body: energy, metabolism, and regulation. *Annu. Rev. Entomol.* 55, 207–225. doi:10.1146/annurev-ento-112408-085356

Baker, P. R. S., Armando, A. M., Campbell, J. L., Quehenberger, O., and Dennis, E. A. (2014). Three-dimensional enhanced lipidomics analysis combining UPLC, differential ion mobility spectrometry, and mass spectrometric separation strategies. *J. Lipid Res.* 55, 2432–2442. doi:10.1194/jlr.D051581

Belhaj, M. R., Lawler, N. G., and Hoffman, N. J. (2021). Metabolomics and lipidomics: expanding the molecular landscape of exercise biology. *Metabolites* 11, 151. doi:10.3390/metabo11030151

Bohrmann, J. (1991). "*In vitro* culture of drosophila ovarian follicles: the influence of different media on development," in *RNA synthesis, protein synthesis and potassium uptake*.

Burguet, P., La Rocca, R., Kune, C., Tellatin, D., Stulanovic, N., Rigolet, A., et al. (2024). Exploiting differential Signal filtering (DSF) and image structure filtering (ISF) methods for untargeted mass spectrometry imaging of bacterial metabolites. *J. Am. Soc. Mass Spectrom.* 35, 1743–1755. doi:10.1021/jasms.4c00129

Buttstedt, A., Pirk, C. W. W., and Yusuf, A. A. (2023). Mandibular glands secrete 24-methylenecholesterol into honey bee (Apis mellifera) food jelly. *Insect Biochem. Mol. Biol.* 161, 104011. doi:10.1016/j.ibmb.2023.104011

Chajduk, M., Tkaczuk, C., and Gołębiowski, M. (2024). Review of the cuticular lipids of spiders (araneae). *Eur. J. Entomol.* 121, 73–82. doi:10.14411/eje.2024.011

Chen, L. jun, zhi, Z., Zhou, X. wei, Xing, X. yi, and Lv, B. (2023). Integrated transcriptome and metabolome analysis reveals molecular responses of spider to

single and combined high temperature and drought stress. *Environ. Pollut.* 317, 120763. doi:10.1016/j.envpol.2022.120763

Childress, C. C., Sacktor, B., and Traynor, D. R. (1967). Function of carnitine in the fatty acid oxidase-deficient insect flight muscle. *J. Biol. Chem.* 242, 754–760. doi:10.1016/s0021-9258(18)96269-1

Connor, A., Zha, R. H., and Koffas, M. (2024). Production and secretion of recombinant spider silk in Bacillus megaterium. *Microb. Cell Fact.* 23, 35. doi:10.1186/s12934-024-02304-5

Conroy, M. J., Andrews, R. M., Andrews, S., Cockayne, L., Dennis, E. A., Fahy, E., et al. (2024). LIPID MAPS: update to databases and tools for the lipidomics community. *Nucleic Acids Res.* 52, D1677–D1682. doi:10.1093/nar/gkad896

Damm, M., Vilcinskas, A., and Lüddecke, T. (2025). Mapping the architecture of animal toxin systems by mass spectrometry imaging. *Biotechnol. Adv.* 81, 108548. doi:10.1016/j.biotechadv.2025.108548

Drozdowski, M., and Boguś, M. I. (2025). Compartmentalization of free fatty acids in blood-feeding Tabanus bovinus females. *Insects* 16, 696. doi:10.3390/insects16070696

Dugon, M. M., Dunbar, J. P., Afoullouss, S., Schulte, J., McEvoy, A., English, M. J., et al. (2017). Occurrence, reproductive rate and identification of the non-native noble false widow spider Steatoda nobilis (Thorell, 1875) in Ireland. *Biol. Environ.* 117B, 77–89. doi:10.1353/bae.2017.0005

Dugon, M. M., Lawton, C., Sturgess, D., and Dunbar, J. P. (2023). Predation on a pygmy shrew, Sorex minutus, by the noble false widow spider, Steatoda nobilis. *Ecosphere* 14, e4422. doi:10.1002/ecs2.4422

Dunbar, J. P., Afoullouss, S., Sulpice, R., and Dugon, M. M. (2018a). Envenomation by the noble false widow spider *Steatoda nobilis* (thorell, 1875) – five new cases of steatodism from Ireland and Great Britain. *Clin. Toxicol.* 56, 433–435. doi:10.1080/15563650.2017.1393084

Dunbar, J. P., Ennis, C., Gandola, R., and Dugon, M. M. (2018b). Biting off more than one can chew: first record of the non-native noble false widow spider Steatoda nobilis (thorell, 1875) feeding on the native viviparous lizard Zootoca vivipara (lichtenstein, 1823) in Ireland. *Biol. Environ. Proc. R. Ir. Acad.* 118B, 45. doi:10.3318/bioe.2018.05

Dunbar, J. P., Fort, A., Redureau, D., Sulpice, R., Dugon, M. M., and Quinton, L. (2020a). Venomics approach reveals a high proportion of lactrodectus-like toxins in the venom of the noble falsewidow spider Steatoda nobilis. *Toxins (Basel)* 12, 402. doi:10.3390/toxins12060402

Dunbar, J. P., Khan, N. A., Abberton, C. L., Brosnan, P., Murphy, J., Afoullouss, S., et al. (2020b). Synanthropic spiders, including the global invasive noble false widow Steatoda nobilis, are reservoirs for medically important and antibiotic resistant bacteria. *Sci. Rep.* 10, 20916. doi:10.1038/s41598-020-77839-9

Dunbar, J. P., Vitkauskaite, A., Lawton, C., Waddams, B., and Dugon, M. M. (2022a). Webslinger vs. dark knight first record of a false widow spider Steatoda nobilis preying on a pipistrelle bat in britain. *Ecosphere* 13, e3959. doi:10.1002/ecs2.3959

Dunbar, J. P., Vitkauskaite, A., O'Keeffe, D. T., Fort, A., Sulpice, R., and Dugon, M. M. (2022b). Bites by the noble false widow spider *Steatoda nobilis* can induce *latrodectus* -like symptoms and vector-borne bacterial infections with implications for public health: a case series. *Clin. Toxicol.* 60, 59–70. doi:10.1080/15563650.2021.1928165

Foelix, R. (2010). *Biology of spiders*. ed. 2010. Oxford University Press.

Gołębiowski, M., Boguś, M. I., Paszkiewicz, M., Wieloch, W., Włóka, E., and Stepnowski, P. (2012). The composition of the cuticular and internal free fatty acids and alcohols from Lucilia sericata males and females. *Lipids* 47, 613–622. doi:10.1007/s11745-012-3662-5

Hajnajafi, K., and Iqbal, M. A. (2025). Mass-spectrometry based metabolomics: an overview of workflows, strategies, data analysis and applications. *Proteome Sci.* 23, 5. doi:10.1186/s12953-025-00241-8

Hustin, J., Kune, C., Far, J., Eppe, G., Debois, D., Quinton, L., et al. (2022). Differential kendrick's plots as an innovative tool for lipidomics in complex samples: comparison of liquid chromatography and infusion-based methods to sample differential Study. *J. Am. Soc. Mass Spectrom.* 33, 2273–2282. doi:10.1021/jasms.2c00232

Kaczmarek, A., Wrońska, A. K., Kazek, M., and Boguś, M. I. (2020). Metamorphosis-related changes in the free fatty acid profiles of sarcophaga (liopygia) argyrostoma (Robineau-Desvoidy, 1830). *Sci. Rep.* 10, 17337. doi:10.1038/s41598-020-74475-1

Kendrick, E. (1963). A mass scale based on $CH_2 = 14.0000$ for high resolution Mass spectrometry of organic compounds. *Anal. Chem.* 35, 2146–2154. doi:10.1021/ac60206a048

Khalil, S. M., Römpp, A., Pretzel, J., Becker, K., and Spengler, B. (2015). Phospholipid topography of whole-body sections of the *Anopheles stephensi* mosquito, characterized by high-resolution atmospheric-pressure scanning microprobe matrix-assisted laser desorption/ionization mass spectrometry imaging. *Anal. Chem.* 87, 11309–11316. doi:10.1021/acs.analchem.5b02781

Khalil, S. M., Pretzel, J., Becker, K., and Spengler, B. (2017). High-resolution AP-SMALDI mass spectrometry imaging of *Drosophila melanogaster*. *Int. J. Mass Spectrom.* 416, 1–19. doi:10.1016/j.ijms.2017.04.001

Klupczynska, A., Pawlak, M., Kokot, Z. J., and Matysiak, J. (2018). Application of metabolomic tools for studying low molecular-weight fraction of animal venoms and poisons. *Toxins (Basel)* 10, 306. doi:10.3390/toxins10080306

Krijnen, K., Blenkinsopp, P., Heeren, R. M. A., and Anthony, I. G. M. (2024). Processing next-generation mass spectrometry imaging data: principal component analysis at scale. *J. Am. Soc. Mass Spectrom.* 35, 3063–3069. doi:10.1021/jasms.4c00314

Kune, C., McCann, A., Raphaël, L. R., Arias, A. A., Tiquet, M., Van Kruining, D., et al. (2019). Rapid visualization of chemically related compounds using kendrick mass defect as a filter in mass spectrometry imaging. *Anal. Chem.* 91, 13112–13118. doi:10.1021/acs.analchem.9b03333

La Rocca, R., Kune, C., Tiquet, M., Stuart, L., Eppe, G., Alexandrov, T., et al. (2021). Adaptive pixel mass recalibration for mass spectrometry imaging based on locally endogenous biological signals. *Anal. Chem.* 93, 4066–4074. doi:10.1021/acs.analchem.0c05071

Laino, A., Cunningham, M., Costa, F. G., and Garcia, C. F. (2013). Energy sources from the eggs of the wolf spider Schizocosa malitiosa: isolation and characterization of lipovitellins. *Comp. Biochem. Physiology - B Biochem. Mol. Biol.* 165, 172–180. doi:10.1016/j.cbpb.2013.04.004

Langenegger, N., Nentwig, W., and Kuhn-Nentwig, L. (2019). Spider venom: components, modes of action, and novel strategies in transcriptomic and proteomic analyses. *Toxins (Basel)* 11, 611. doi:10.3390/toxins11100611

Lim, H. Y., Wang, W., Wessells, R. J., Ocorr, K., and Bodmer, R. (2011). Phospholipid homeostasis regulates lipid metabolism and cardiac function through SREBP signaling in Drosophila. *Genes Dev.* 25, 189–200. doi:10.1101/gad.1992411

Lu, W., Shi, R., Li, X., Ma, S., Yang, D., Shang, D., et al. (2024). A review on complete silk gene sequencing and *de novo* assembly of artificial silk. *Int. J. Biol. Macromol.* 264, 130444. doi:10.1016/j.ijbiomac.2024.130444

Ma, L., Xie, Q., Du, M., Huang, Y., Chen, Y., Chen, D., et al. (2022). Sample preparation optimization of insects and zebrafish for whole-body mass spectrometry imaging. *Anal. Bioanal. Chem.* 414, 4777–4790. doi:10.1007/s00216-022-04102-7

Ma, Q., Wang, Z., Xu, H., Wei, Y., and Liang, M. (2024). Effects of dietary cholesterol on ovary development and reproductive capacity in Pacific white shrimp broodstock, litopenaeus vannamei. *Aquac. Rep.* 38, 102346. doi:10.1016/j.aqrep.2024.102346

Marc, P., Canard, A., and Ysnel, F. (1999). Spiders (araneae) useful for Pest limitation and bioindication. *Agric. Ecosyst. Environ.* 74, 229–273. doi:10.1016/S0167-8809(99)00038-9

Martínez-Gardeazabal, J., González de San Román, E., Moreno-Rodríguez, M., Llorente-Ovejero, A., Manuel, I., and Rodríguez-Puertas, R. (2017). Lipid mapping of the rat brain for models of disease. *Biochim. Biophys. Acta Biomembr.* 1859, 1548–1557. doi:10.1016/j.bbamem.2017.02.011

McCann, A., Rappe, S., La Rocca, R., Tiquet, M., Quinton, L., Eppe, G., et al. (2021). Mass shift in mass spectrometry imaging: comprehensive analysis and practical corrective workflow. *Anal. Bioanal. Chem.* 413, 2831–2844. doi:10.1007/s00216-021-03174-1

Meyer, H., Bossen, J., Janz, M., Müller, X., Künzel, S., Roeder, T., et al. (2024). Combined transcriptome and proteome profiling reveal cell-type-specific functions of drosophila Garland and pericardial nephrocytes. *Commun. Biol.* 7, 1424. doi:10.1038/s42003-024-07062-z

Palmer, A., Phapale, P., Chernyavsky, I., Lavigne, R., Fay, D., Tarasov, A., et al. (2017). FDR-Controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nat. Methods* 14, 57–60. doi:10.1038/nmeth.4072

Polilov, A. A., Makarova, A. A., Pang, S., Shan Xu, C., and Hess, H. (2021). Protocol for preparation of heterogeneous biological samples for 3D electron microscopy: a case study for insects. *Sci. Rep.* 11, 4717. doi:10.1038/s41598-021-83936-0

Propistsova, E. A., Makarova, A. A., Eskov, K. Y., and Polilov, A. A. (2023). Miniaturization does not change conserved spider anatomy, a case study on spider rayforstia (araneae: anapidae). *Sci. Rep.* 13, 17219. doi:10.1038/s41598-023-44230-3

Raghu, P., Coessens, E., Manifava, M., Georgiev, P., Pettitt, T., Wood, E., et al. (2009). Rhabdomere biogenesis in drosophila photoreceptors is acutely sensitive to phosphatidic acid levels. *J. Cell Biol.* 185, 129–145. doi:10.1083/jcb.200807027

Rayner, S., Vitkauskaite, A., Healy, K., Lyons, K., McSharry, L., Leonard, D., et al. (2022). Worldwide web: high venom potency and ability to optimize venom usage make the globally invasive noble false widow spider Steatoda nobilis (thorell, 1875) (theridiidae) highly competitive against native European spiders sharing the same habitats. *Toxins (Basel)* 14, 587. doi:10.3390/toxins14090587

Römer, L., and Scheibel, T. (2008). The elaborate structure of spider silk: structure and function of a natural high performance fiber. *Prion* 2, 154–161. doi:10.4161/pri.2.4.7490

Ryan Wilson (Vector), B. Kimmel (Numbered version), and CC BY 3.0 (2010). Spider figure adapted from the spider book (1912, 1920) by john henry comstock and from biology of spiders (1996) by rainer F. Foelix. Available online at: https://upload.wikimedia.org/wikipedia/commons/3/39/Spider_internal_anatomy-numb.svg.

Sánchez-Juanes, F., Calvo Sánchez, N., Belhassen García, M., Vieira Lista, C., Román, R. M., Álamo Sanz, R., et al. (2022). Applications of MALDI-TOF mass spectrometry to the identification of parasites and arthropod vectors of human diseases. *Microorganisms* 10, 2300. doi:10.3390/microorganisms10112300

Schramm, T., Hester, A., Klinkert, I., Both, J. P., Heeren, R. M. A., Brunelle, A., et al. (2012). ImzML - a common data format for the flexible exchange and processing of mass spectrometry imaging data. *J. Proteomics* 75, 5106–5110. doi:10.1016/j.jprot.2012.07.026

Shevchenko, A., and Simons, K. (2010). Lipidomics: coming to grips with lipid diversity. *Nat. Rev. Mol. Cell Biol.* 11, 593–598. doi:10.1038/nrm2934

Smith, D. S., Järlfors, U., and Russell, F. E. (1969). The fine structure of muscle attachments in a spider (Latrodectus mactans, fabr.). *Tissue Cell* 1, 673–687. doi:10.1016/S0040-8166(69)80040-6

Stark, W. S., Lin, T., Brackhahn, D., Christianson, J. S., and Sun, G. Y. (1993). Phospholipids in *drosophila* heads: effects of visual mutants and phototransduction manipulations. *Lipids* 28, 23–28. doi:10.1007/BF02536355

Steinhauer, J., Gijó, M. A., Riekhof, W. R., Voelker, D. R., Murphy, R. C., Treisman, J. E., et al. (2009). Drosophila lysophospholipid acyltransferases are specifically required for germ cell development. *Mol. Biol. Cell* 20, 5224–5235. doi:10.1091/mbc.e09-05-0382

Stoeckli, M., Staab, D., Schweitzer, A., Gardiner, J., and Seebach, D. (2007). Imaging of a beta-peptide distribution in whole-body mice sections by MALDI mass spectrometry. *J.Am.Soc.Mass Spectrom.* 18, 1921–1924. doi:10.1016/j.jasms.2007.08.005

Stork, N. E. (2017). EN63CH03-Stork ARI how many species of insects and other terrestrial arthropods are there on Earth? EN63CH03-Stork ARI. doi:10.1146/annurev-ento-020117

Stutts, W. L., Knuth, M. M., Ekelöf, M., Mahapatra, D., Kullman, S. W., and Muddiman, D. C. (2020). Methods for cryosectioning and mass spectrometry imaging of whole-body zebrafish. *J. Am. Soc. Mass Spectrom.* 31, 768–772. doi:10.1021/jasms.9b00097

Tiquet, M., La Rocca, R., Kirnbauer, S., Zoratto, S., Van Kruining, D., Quinton, L., et al. (2022). FT-ICR mass spectrometry imaging at extreme mass resolving power using a dynamically harmonized ICR cell with 1ω or 2ω detection. *Anal. Chem.* 94, 9316–9326. doi:10.1021/acs.analchem.2c00754

Torres, G., Melzer, R. R., Spitzner, F., Šargač, Z., Harzsch, S., and Gimenez, L. (2021). Methods to study organogenesis in decapod crustacean larvae. I. Larval rearing, preparation, and fixation. *Helgol Mar. Res.* 75, 3. doi:10.1186/s10152-021-00548-x

Trabalon, M., and Garcia, C. F. (2021). Transport pathways of hydrocarbon and free fatty acids to the cuticle in arthropods and hypothetical models in spiders. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* 252, 110541. doi:10.1016/j.cbpb.2020.110541

Triebl, A., Trötzmüller, M., Hartler, J., Stojakovic, T., and Köfeler, H. C. (2017). Lipidomics by ultrahigh performance liquid chromatography-high resolution mass spectrometry and its application to complex biological samples. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* 1053, 72–80. doi:10.1016/j.jchromb.2017.03.027

Virmani, M. A., and Cirulli, M. (2022). The role of l-Carnitine in mitochondria, prevention of metabolic inflexibility and disease initiation. *Int. J. Mol. Sci.* 23, 2717. doi:10.3390/ijms23052717

Wang, Y., Yu, Z., Zhang, J., and Moussian, B. (2016). Regionalization of surface lipids in insects. *Proc. R. Soc. B Biol. Sci.* 283, 20152994. doi:10.1098/rspb.2015.2994

Wenk, M. R. (2005). The emerging field of lipidomics. *Nat. Rev. Drug Discov.* 4, 594–610. doi:10.1038/nrd1776

Whitehead, W. F., and Rempel, J. G. (1959). A study of the musculature of the black widow spider, Latrodectus mactans (fabr.). *Can. J. Zool.* 37, 831–870. doi:10.1139/z59-084

Wishart, D. S. (2019). Metabolomics for investigating physiological and pathophysiological processes. *Physiol. Rev.* 99, 1819–1875. doi:10.1152/physrev.00035.2018

Wishart, D. S., Guo, A. C., Oler, E., Wang, F., Anjum, A., Peters, H., et al. (2022). HMDB 5.0: the human metabolome Database for 2022. *Nucleic Acids Res.* 50, D622–D631. doi:10.1093/nar/gkab1062

Wrońska, A. K., Kaczmarek, A., Boguś, M. I., and Kuna, A. (2023). Lipids as a key element of insect defense systems. *Front. Genet.* 14, 1183659. doi:10.3389/fgene.2023.1183659

Xu, H., Chen, L., Tong, X. L., Hu, H., Liu, L. Y., Liu, G. C., et al. (2022). Comprehensive silk gland multi-omics comparison illuminates two alternative mechanisms in silkworm heterosis. *Zool. Res.* 43, 585–596. doi:10.24272/j.issn.2095-8137.2022.065

Xu, T., Molday, L. L., and Molday, R. S. (2023). Retinal-phospholipid schiff-base conjugates and their interaction with ABCA4, the ABC transporter associated with Stargardt disease. *J. Biol. Chem.* 299, 104614. doi:10.1016/j.jbc.2023.104614

Yang, K., and Han, X. (2016). Lipidomics: techniques, applications, and outcomes related to biomedical sciences. *Trends Biochem. Sci.* 41, 954–969. doi:10.1016/j.tibs.2016.08.010

Yang, F. Y., Chen, J. H., Ruan, Q. Q., Saqib, H. S. A., He, W. Y., and You, M. S. (2020). Mass spectrometry imaging: an emerging technology for the analysis of metabolites in insects. *Arch. Insect Biochem. Physiol.* 103, e21643. doi:10.1002/arch.21643

Zhao, H., and Wang, T. (2020). PE homeostasis rebalanced through mitochondria-ER lipid exchange prevents retinal degeneration in Drosophila. *PLoS Genet.* 16, e1009070. doi:10.1371/journal.pgen.1009070

Zheng, Y., Li, Y., Rimm, E. B., Hu, F. B., Albert, C. M., Rexrode, K. M., et al. (2016). Dietary phosphatidylcholine and risk of all-cause and cardiovascular-specific mortality among US women and men. *Am. J. Clin. Nutr.* 104, 173–180. doi:10.3945/ajcn.116.131771