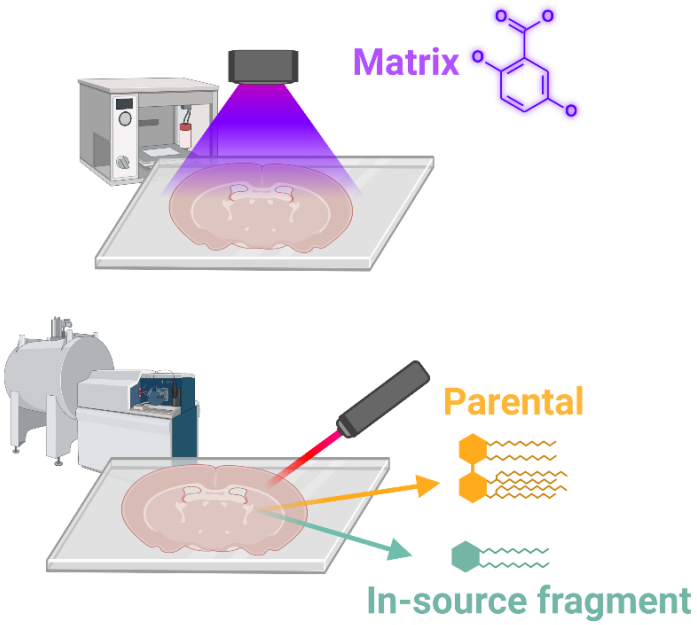
COMPUTATIONAL TOOLS FOR THE ANNOTATION OF IN-SOURCE FRAGMENTS AND MATRIX-RELATED SIGNALS IN MALDI MASS SPECTROMETRY IMAGING

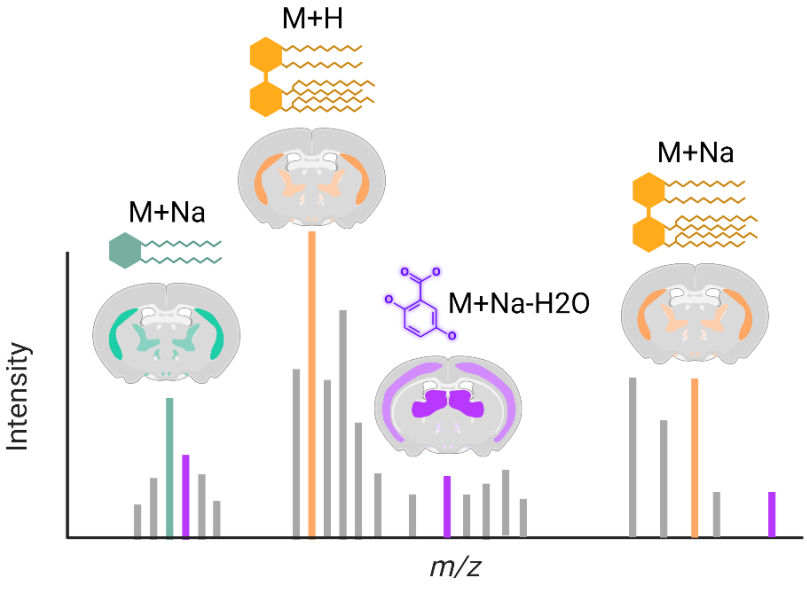
Gerard Baquer Gómez

UNIVERSITAT ROVIRA i VIRGILI

Computational tools for the annotation of in-source fragments and matrix-related signals in MALDI Mass Spectrometry Imaging

Gerard Baquer Gómez





TESI DOCTORAL / TESIS DOCTORAL / DOCTORAL THESIS
2023

UNIVERSITAT ROVIRA I VIRGILI
COMPUTATIONAL TOOLS FOR THE ANNOTATION OF IN-SOURCE FRAGMENTS AND MATRIX-RELATED SIGNALS IN MALDI MASS SPECTROMETRY IMAGING
Gerard Baquer Gómez

Gerard Baquer Gómez

Computational tools for the annotation of in-source fragments and matrix-related signals in MALDI Mass Spectrometry Imaging

DOCTORAL THESIS

supervised by Dr. Xavier Correig Blanchar and

Dr. Pere Ràfols Soler

Departament d'Enginyeria Electrònica, Elèctrica i Automàtica (DEEEA)

UNIVERSITAT ROVIRA i VIRGILI

Tarragona

2023

UNIVERSITAT ROVIRA i VIRGILI

Escola Tècnica Superior d'Enginyeria

Departament d'Enginyeria Electrònica, Electrica i Automatica

Av. Paisos Catalans 26

Campus Sescelades

43007 Tarragona

We STATE that the present study, entitled "**Computational tools for the annotation of in-source fragments and matrix-related signals in MALDI Mass Spectrometry Imaging**", presented by **Gerard Baquer Gómez** for the award of the degree of Doctor, has been carried out under our supervision at the Department of Electronic, Electric and Automatic Engineering of this university and meets the requirements to qualify for International Mention.

Tarragona, November 2022

Doctoral thesis supervisors:

Francesc Xavier Correig Blanchar - DNI 39849924D (AUT)
2022.11.20 13:04:34 +01'00'

Prof. Xavier Correig Blanchar

Dr. Pere Ràfols Soler

Dedicated to Daniel Escobar Solà.

Acknowledgements

[PENDING]

Abstract

Matrix-Assisted Laser Desorption Ionization Mass Spectrometry Imaging (MALDI-MSI) is an analytical technique used in biochemical and clinical studies to reveal the chemical composition and spatial information of organic tissues. It provides valuable information in many applications, including the understanding and diagnosis of complex diseases such as cancer, diabetes, Alzheimer's and infectious diseases.

Despite the surge of MALDI-MSI's popularity, associating each mass-to-charge (m/z) signal with univocal molecular identifications remains challenging: (1) samples include thousands of molecules; (2) each molecule is responsible for several MS signals (isotopes, adducts, in-source fragments, multiple charges…); and (3) isomers and isobars cannot be resolved using only MS1.

Traditional mass spectrometry techniques rely on chromatographic separation (LC-MS, GC-MS) for sample simplification. However, MALDI-MSI does not include such separation steps. Complementary, tandem mass spectrometry can augment the depth of the chemical analysis by providing fragmentation information on single molecules. Many MALDI-MSI instruments are equipped with tandem-MS capabilities (Bruker's ultrafleXtreme, Thermo Scientific's MALDI LTQ Orbitrap XL, or Waters' MALDI SYNAPT G2-Si) but untargeted imaging MS/MS is not routinely feasible due to (1) prohibitive running times, (2) limited parental ion selectivity and intensity, and (3) increased data size and complexity. For all these reasons, untargeted fragmentation of all ions in a sample is only possible using highly specialized instrumental setups.

In this complex scenario, the annotation of MS signals and putative identification of metabolites present in the sample is a daunting task. There are several software solutions to perform automatic annotation of MSI data. However, two types of signals have been traditionally overlooked and underestimated: in-source fragments and matrix-related signals.

In-Source Decay (ISD) or In-Source Fragmentation (ISF) (i.e. the natural and unavoidable generation of fragments inside the MALDI ion source) needs to be minimized. ISD depends mainly on the chemical structure of the analyte and ionization conditions such as ionization temperature or voltage and can be problematic in the study of lipids, as several fragmentation pathways lead to isobaric lipid species. These known lipid fragmentation pathways result in falsely low concentrations of lipids suffering from ISD and falsely high concentrations of lipids overlapping with isobaric in-source fragments. Additionally, if not properly annotated and removed, in-source fragments can yield an increased number of incorrect annotations using common MALDI-MSI annotation tools such as LipoStar, METASPACE, and rMSIannotation

As a first objective of this thesis, we first propose rMSIfragment, a software solution that exploits known in-source fragmentation pathways to increase confidence in lipid annotations. Our novel ranking score combines the times a given lipid has been found in the dataset (adducts and in-source fragments) and their spatial correlation to filter out unlikely lipids. After validation with HPLC and 2 different Target-Decoy approaches, rMSIfragment demonstrates good performance on multiple sample types and experimental conditions. We also find that ISD-agnostic annotation tools like METASPACE can falsely annotate in-source fragments as endogenous lipids.

Additionally, in the classical MALDI-MSI workflow, an organic compound (e.g. matrix) is deposited onto the sample to promote the desorption and ionization of endogenous analytes. Unfortunately, this low-weight exogenous compound adds several undesired MS signals to the MALDI-MSI spectra; including exogenous matrix signals (adducts, multiple charges, and in-source fragments) and matrix adducts with endogenous biomolecules. These signals add an undesired layer of complexity to core MSI processing pipelines like untargeted statistical analyses (Baquer et al. 2020) or molecular annotation (Baquer et al. 2022). This is particularly worrying in metabolomics and lipidomics, as matrix-related signals are densely concentrated in the low *m/z* range.

As a second objective of this thesis, we develop rMSIcleanup, a computational tool to automatically annotate matrix-related peaks. The algorithm also incorporates an overlapping peak detection feature to prevent misclassification of overlapped or isobaric ions. In a first iteration we validate rMSIcleanup on a well-understood LDI promoting material such as silver. Later, we demonstrate its use in annotating the most widely used organic matrix, DHB.

In further efforts we acknowledge that current automatic tools for the annotation of matrix-annotation tools suffer from multiple of the following pitfalls: (1) focus exclusively on the spatial distribution, (2) do not control the False Discovery Rate (FDR), (3) do not consider adducts with endogenous metabolites, and (4) rely on a predefined list of theoretical matrix adducts.

We develop an experimental and computational workflow to discover matrix-containing adducts using $13C^6$-labeled 2,5-Dihydroxybenzoic acid ($13C^6$-DHB). By exploiting the labeling-induced m/z shift and unique spatial distribution of matrix-containing ions we can discover and annotate matrix-containing adducts formed with exogenous and endogenous compounds.

In both cases, we demonstrate that the proper annotation of in-source fragments and matrix-related signals significantly improves key untargeted metabolomics workflows such as dimensionality reduction and metabolite annotation.

Overall, we conclude that rMSIfragment, rMSIcleanup and the use of SIL-MALDI-MSI for matrix-related signal discovery are three essential tools to incorporate in routinary MSI processing pipelines.

List of publications

Gerard Baquer*, Miguel Bernús* , Lluc Sementé, René van Zeil, Maria García-Altares, Bram Heijs, Christoph Bookmeyer, Omar Boutureira, Xavier Correig, Pere Ràfols. *“FDR-controlled discovery of matrix-related signals using labeled matrix improves statistical analysis and small molecule annotation in MALDI-MSI”* [Submitted]

Gerard Baquer, Lluc Sementé, Pere Ràfols, Lucía Martín-Saiz, Christoph Bookmeyer, José A. Fernández, Xavier Correig & María García-Altares. *“rMSIfragment: Automated in-source fragmentation and adduct annotation of lipids in MALDI-MSI”* [Submitted]

Gerard Baquer*, Lluc Sementé*, Toufik Mahamdi, Xavier Correig, Pere Ràfols, and María García-Altares. *"What are we imaging? Software tools and experimental strategies for annotation and identification of small molecules in mass spectrometry imaging."* Mass Spectrometry Reviews (2022): e21794.

Gerard Baquer, Lluc Sementé, María García-Altares, Young Jin Lee, Pierre Chaurand, Xavier Correig, and Pere Ràfols. *"rMSIcleanup: an open-source tool for matrix-related peak annotation in mass spectrometry imaging and its application to silver-assisted laser desorption/ionization."* Journal of Cheminformatics 12, no. 1 (2020): 1-13.

Lluc Sementé, **Gerard Baquer**, María García-Altares, Xavier Correig-Blanchar, and Pere Ràfols. *"rMSIannotation: A peak annotation tool for mass spectrometry imaging based on the analysis of isotopic intensity ratios." Analytica Chimica Acta* 1171 (2021): 338669.

Stefania-Alexandra Iakab, **Gerard Baquer**, Marta Lafuente, Maria Pilar Pina, José Luis Ramírez, Pere Ràfols, Xavier Correig-Blanchar, and María García-Altares. *"SALDI-MS and SERS Multimodal Imaging: One Nanostructured Substrate to Rule Them Both." Analytical chemistry* 94, no. 6 (2022): 2785-2793.

Giulia Notarangelo, Jessica B. Spinelli, Elizabeth M. Perez, Gregory J. Baker, Kiran Kurmi, Ilaria Elia, Sylwia A. Stopka, **Gerard Baquer**, Jia-Ren Lin, Alexandra J. Golby, Shakchhi Joshi, Heide F. Baron, Jefte M. Drijvers, Peter Georgiev, Alison E. Ringel, Elma Zaganjor, Samuel K. McBrayer, Peter K. Sorger, Arlene H. Sharpe, Kai W. Wucherpfennig, Sandro Santagata, Nathalie Y.R. Agar, Mario L. Suvà, and Marcia C. Haigis. *"Oncometabolite d-2HG alters T cell metabolism to impair CD8+ T cell function."* Science 377, no. 6614 (2022): 1519-1529.

Shannon Coy, Shu Wang, Sylwia A. Stopka, Jia-Ren Lin, Clarence Yapp, Cecily C. Ritch, Lisa Salhi, Gregory J. Baker, Rumana Rashid, **Gerard Baquer**, Michael Regan, Prasidda Khadka, Kristina A. Cole, Jaeho Hwang, Patrick Y. Wen, Pratiti Bandopadhayay, Mariarita Santi, Thomas De Raedt, Keith L. Ligon, Nathalie Y.R. Agar, Peter K. Sorger, Mehdi Touat, and Sandro Santagata. *"Single-cell spatial analysis reveals the topology of immunomodulatory purinergic signaling in glioblastoma."* Nature communications 13, no. 1 (2022): 1-24.

Lin Wang, Dan Wang, Olmo Sonzogni, Shizhong Ke, Qi Wang, Abhishek Thavamani, Felipe Batalini, Sylwia A. Stopka, Michael S. Regan, Steven Vandal, Shengya Tian, Jocelin Pinto, Andrew M. Cyr, Vanessa C. Bret-Mounet, **Gerard Baquer**, Hans P. Eikesdal, Min Yuan, John M. Asara, Yujing J. Heng, Peter Bai, Nathalie Y.R. Agar, and Gerburg M.Wulf. *"PARP-inhibition reprograms macrophages toward an anti-tumor phenotype."* Cell Reports 41, no. 2 (2022): 111462.

* These authors contributed equally

List of conferences

Gregory J. Baker*, **Gerard Baquer***, Sylwia A. Stopka, Shannon Coy, Michael Regan, Jia-Ren Lin, Peter K. Sorger, Sandro Santagata, Nathalie Y.R. Agar. “Integration of MALDI-MSI with Multiplexed Tissue Immunofluorescence on Serial Sections Enables Integrated Analysis of the Glioblastoma Microenvironment”, ASMS 37th Asilomar Conference on Mass Spectrometry - Single-cell Mass Spectrometry, Pacific Grove, CA, USA (2022) **[Oral communication]**

Gerard Baquer, Pere Ràfols, María García-Altares, Lluc Sementé, Esteban del Castillo, and Xavier Correig. *“rMSIcleanup: an open-source computational tool for matrix-related peak annotation in MALDI-MSI”*, OurCon VII, Saint-Malo, France (2019)

Gerard Baquer, Pere Ràfols, Maria Garcia-Altares, Maria Vinaixa, and Xavier Correig. “rMSIcleanup: an open-source computational tool for matrix-related peak annotation in MALDI-MSI”, METABOLOMICS, The Hague, Netherlands (2019) **[Oral communication]**

Gerard Baquer, Pere Ràfols, Maria Garcia-Altares, Maria Vinaixa, Xavier Correig. “Matrix-related peak annotation in MALDI-MSI”, Converging Imaging and Systems Medicine (ConISyM), Castle Ringberg, Germany (2019) **[Oral communication]**

List of international secondments

“MALDI-MSI experimental protocol training” supervised by René van Zeil and Bram Heijs, Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, The Netherlands (1 week, 2019)

“MALDI-MSI experimental protocol training” supervised by Nathalie Y.R. Agar, Surgical Molecular Imaging Laboratory, Department of Neurosurgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA (6 months, 2021-2022)

Table of Contents

CHAPTER 1:

Introduction

1. Spatial metabolomics in the era of personalized medicine

Precision medicine is one of the most ambitious goals of our times. The shift towards patient-tailored diagnostic, treatment and prognostic clinical practices promise to help millions of people worldwide (Ovchinnikova et al. 2020). However, precision medicine requires a deep and fundamental understanding of the complex interplay of genetic, phenotypic, and exposition factors (Wang et al. 2022). Complex diseases like cancer, diabetes, or neurodegenerative diseases still present plenty of mysteries (Dreisewerd 2013).

The mantra of the past decades has been that serious and chronic diseases have purely genetic origins (Wishart 2016). Thousands of genomics studies have been performed on population-wide cohorts aimed at uncovering disease-associated genes and Single nucleotide polymorphisms (SNPs).. However, the number of identified disease genes and SNPs has not matched the initial excitement and expectations (Wishart 2016). In recent years, it has become apparent the need for a system's approach capable of capturing downstream biological and biochemical processes closer to the phenotype. For example, RNA (transcriptomics) and protein levels (proteomics), have been associated with diseases like diabetes and cancer.

Metabolomics, the analysis of small biomolecules, has received particular attention recently. Genes suggest what could happen, but metabolites reveal what is actually happening (Wishart 2016). Metabolites serve a myriad of functions that include fuel, structure, signaling, and enzyme regulation and are thus key players in disease. The oncometabolite 2HG, for instance, has been recently shown to promote tumorigenesis by impairing CD8+ T cell function (Notarangelo et al. 2022). Knowledge about metabolites is used daily in medical practice in the form of disease biomarkers or inhibitory drugs.

While bulk biofluid studies help uncover an overall picture of the metabolic state of the body, spatial metabolomics techniques enable a more refined look into tissues and cells. Contextualizing each metabolite in space can help us understand the interaction between different metabolites, structures and cell types (Nikolaev et al. 2016).

2. MALDI Mass Spectrometry Imaging

Matrix-Assisted Laser Desorption Ionization Mass Spectrometry Imaging (MALDI-MSI) is the workhorse of spatial metabolomics. In the past two decades, it has become a pivotal analytical tool in the study of complex diseases such as cancer and diabetes (Fig 1).

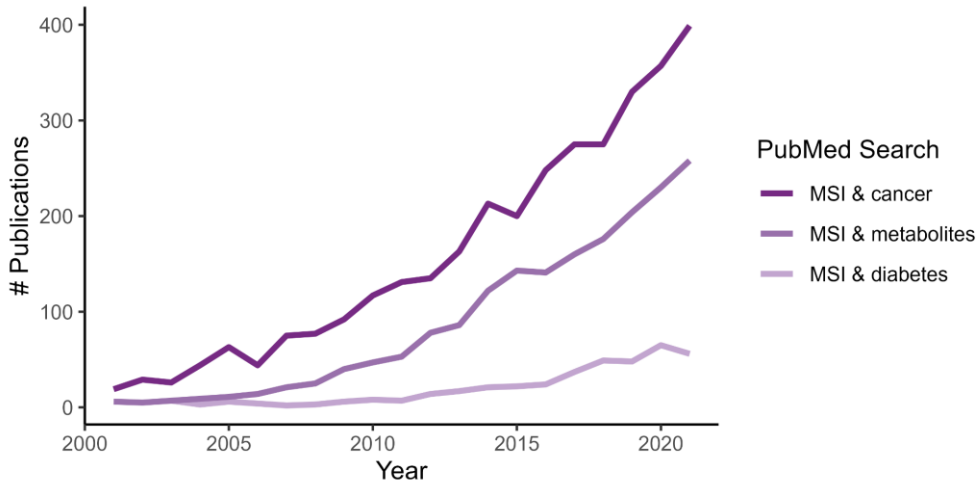


Figure 1. The number of PubMed publications per year in the field of MSI.

The typical MALDI-MSI workflow includes sectioning, sample preparation, and acquisition (Fig. 2). To prevent metabolite degradation, the sample (typically an animal model tissue or patient biopsy) is snap-frozen at -80 ºC after extraction. A cryostat is used to obtain thin sections which are then mounted onto Iridium Tin Oxide-coated glass slides. The next step is coating the sample with the MALDI matrix, an organic compound that absorbs the energy of the laser and promotes analyte desorption and ionization. Finally, a laser is rastered across the sample and the generated ions are analyzed by MS. For each pixel, we obtain an MS spectrum informing about the tissue composition at that location. Each of the MS features (e.g. peaks) in the spectrum can be mapped out to reveal their distribution over the tissue. The resulting dataset contains 100s of molecular images that are used to study the tissue.

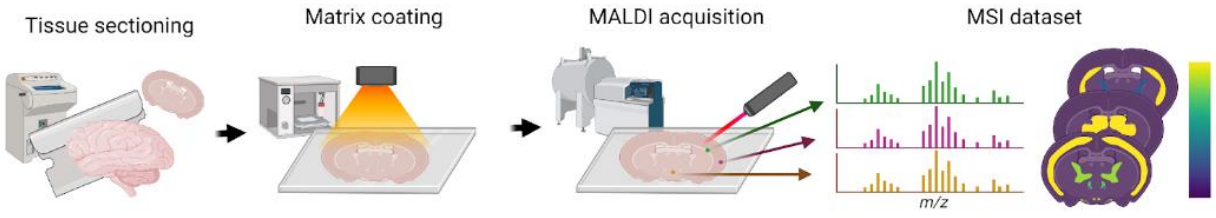


Figure 2. General MALDI-MSI. For each pixel, we obtain an MS spectrum (chemical composition). For each peak in the MS spectra (molecule of interest) we can image their distribution.

Chapter 2 will offer a detailed explanation of the instrumental aspects of MALDI-MSI. Here we provide a brief overview for the reader to gain basic familiarity with the technique.

MALDI is a soft ionization technique that relies on the absorbance of the energy of a laser by a matrix to desorb and ionize analytes (Dreisewerd 2003; Hillenkamp, Jaskolla, and Karas 2014; Knochenmuss and Zenobi 2003). It is characterized by its low in-source fragmentation and the predominant formation of single ion species. In MALDI, the analytes that reach the MS analyzer depend on two convolved processes: desorption and ionization. The dynamics of these two processes largely depend on the matrix and the laser, two analytical aspects that should be optimized together. The choice of matrix and its method of deposition will influence its co-crystalization with analytes. Other important parameters are laser wavelength, pulse duration, and laser fluence (energy per pulse and unit area). Interestingly, the dynamics of the particle plume are also an influential factor. The interactions between neutral molecules and ions in the plume will determine which ions make it into the analyzer (Dreisewerd 2003).

Once ions are formed, they need to be analyzed. The three main MS analyzer principles of operation are Time Of Flight (TOF), Fourier Transform Ion Cyclotron Resonance (FTICR) and Orbitrap.

TOF mass analyzers are vacuum tubes in which ions move to the detector via an electric field. Mass resolution increases with tube length, as the ions have more time to separate during flight. As a result, TOF mass analyzers have a moderate mass resolution because higher mass resolutions need larger instruments and longer sampling times. TOF analyzers, on the other hand, do not have a theoretical upper m/z limit of detection (Jurinke, Oeth, and van den Boom 2004).

FTICR mass analyzers use a magnetic field to resonate ions into cyclotron orbits, which are then converted into m/z using the Fourier Transform. A higher mass resolution requires a stronger magnetic field. These instruments are better suited for smaller molecules (< m/z 3000) because the mass resolution is inversely proportional to the m/z (Nikolaev, Kostyukevich, and Vladimirov 2016).

Orbitrap mass analyzers use electrically charged ion trap cells to excite ions into orbits. The orbits' longitudinal movement contains information about the cyclotron frequencies of each ion, which can be converted to m/z using the Fourier transform. Orbitraps can achieve high mass by increasing the electric field (Hu et al. 2005).

3. Thesis motivation and objectives

The spectra of a typical MALDI-MSI experiment are complex and include multiple signals pertaining to the same molecule (isotopes, adducts, multiple charges …). Multiple tools have been developed over the years to facilitate the automatic annotation of metabolites and lipids in MSI. Some examples include pySM (Palmer et al. 2016), massPix (Bond et al. 2017), LipostarMSI (Tortorella et al. 2020), and rMSIannotation (Sementé et al. 2021). However, two types of MS signals have been traditionally overlooked. Namely, in-source fragments and MALDI-matrix-related signals.

This thesis pursues the following 3 objectives:

Objective 1: Develop an automated tool for the annotation of in-source fragments in lipids.

Lipids are interesting molecules for the understanding of several diseases, due to their multiple functions including energy storage, structure, and signaling. However, their measurement and annotation in MALDI-MSI are complex due to their propensity to break down into multiple in-source fragments (Garate et al. 2020). This is particularly challenging when considering that multiple of these in-source fragments have been demonstrated to be isomeric to other parental (unfragmented) lipids.

We, therefore, aim to develop an automatic tool to annotate in-source fragments and increase the confidence in the annotation of lipids in MALDI-MSI.

Objective 2: Develop an automated tool for the annotation of matrix-related signals.

In MALDI-MSI, the matrix used to promote desorption and ionization of analytes introduces interferences in the low-mass range. This is a particular issue in small molecule applications such as metabolomics and lipidomics. The removal of these non-biologically relevant peaks promises to improve the downstream analysis of MSI data.

Additionally, the matrix forms adducts with endogenous compounds. Their correct annotation promises to clear up part of the dark metabolome in MSI experiments.

We aim to develop an automated tool to annotate matrix-related signals.

Objective 3: Study the influence of in-source fragments and matrix-related signals on untargeted metabolomics workflows.

If not properly annotated, a high presence of in-source fragments and matrix-related signals can have a deleterious effect on downstream analyses.

We aim to quantify both the prevalence of these signals in a typical MSI dataset and their influence on downstream untargeted metabolomics workflows.

4. Document structure

Chapter 1 introduces the role of spatial metabolomics in clinical research and describes MALDI-MSI as a crucial analytical technique. It concludes by outlining the motivation and main objectives of this thesis.

Chapter 2 reviews experimental and computational considerations for the annotation and identification of small molecules with MSI. We call attention to the report of confidence levels and use them throughout the review. The experimental section covers sample preparation, acquisition, orthogonal techniques, and validation against reference standards. The bioinformatics section first discusses key principles to guide the development of new annotation tools. It later reviews the most influential annotation and identification software tools in MSI. We finally provide perspectives on the future of automatic metabolite annotation in MSI.

Chapter 3 presents rMSIfragment, a novel automatic tool for the annotation of in-source fragments. We first introduce the topic of MALDI in-source fragmentation and present the algorithmic foundations of the annotation tool. We then provide several alternative validations to demonstrate its robustness and applicability under a wide range of experimental conditions. We justify the need for correct in-source fragment annotation by finding the overlap of several in-source fragments with endogenous lipids.

Chapter 4 and 5 center on the topic of automated annotation of matrix-related signals. Chapter 3 covers the first deployment of rMSIcleanup. After providing the algorithmic rationale we proceed to demonstrate its applicability to the annotation of matrix-related signals in Ag-LDI-MSI. Chapter 4 expands the functionality of rMSIcleanup to DHB, the most widely used organic MALDI matrix. We present a novel approach for matrix-related signal discovery based on the use of Stable Isotope Labeled matrix analog. We validate our findings with a set of positive (DHB) and negative (other matrices) controls as well as publicly available MSI datasets.

Finally, Chapter 6 discusses the collective impact of the presented work and provides a set of conclusions and avenues for future work.

5. References

Baquer, Gerard, Lluc Sementé, María García-Altares, Young Jin Lee, Pierre Chaurand, Xavier Correig, and Pere Ràfols. 2020. "rMSIcleanup: An Open-Source Tool for Matrix-Related Peak Annotation in Mass Spectrometry Imaging and Its Application to Silver-Assisted Laser Desorption/ionization." *Journal of Cheminformatics* 12 (1): 45.

Baquer, Gerard, Lluc Sementé, Toufik Mahamdi, Xavier Correig, Pere Ràfols, and María García-Altares. 2022. "What Are We Imaging? Software Tools and Experimental Strategies for Annotation and Identification of Small Molecules in Mass Spectrometry Imaging." *Mass Spectrometry Reviews*, July, e21794.

Bond, Nicholas J., Albert Koulman, Julian L. Griffin, and Zoe Hall. 2017. "massPix: An R Package for Annotation and Interpretation of Mass Spectrometry Imaging Data for Lipidomics." *Metabolomics: Official Journal of the Metabolomic Society* 13 (11): 128.

Coy, Shannon, Shu Wang, Sylwia A. Stopka, Jia-Ren Lin, Clarence Yapp, Cecily C. Ritch, Lisa Salhi, et al. 2022. "Single Cell Spatial Analysis Reveals the Topology of Immunomodulatory Purinergic Signaling in Glioblastoma." *Nature Communications* 13 (1): 4814.

Dreisewerd, Klaus. 2003. "The Desorption Process in MALDI." *Chemical Reviews*. https://doi.org/10.1021/cr010375i.

Dührkop, Kai, Markus Fleischauer, Marcus Ludwig, Alexander A. Aksenov, Alexey V. Melnik, Marvin Meusel, Pieter C. Dorrestein, Juho Rousu, and Sebastian Böcker. 2019. "SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information." *Nature Methods* 16 (4): 299–302.

Garate, Jone, Sergio Lage, Lucía Martín-Saiz, Arantza Perez-Valle, Begoña Ochoa, M. Dolores Boyano, Roberto Fernández, and José A. Fernández. 2020. "Influence of Lipid Fragmentation in the Data Analysis of Imaging Mass Spectrometry Experiments." *Journal of the American Society for Mass Spectrometry* 31 (3): 517–26.

Hillenkamp, Franz, Thorsten W. Jaskolla, and Michael Karas. 2014. "The MALDI Process and Method." *MALDI MS. A Practical Guide to Instrumentation, Methods, and Applications, 2nd Ed. (Ed. : F. Hillenkamp, J. Peter-Katalinic), Wiley Blackwell, Weinheim, Germany*. https://onlinelibrary.wiley.com/doi/pdf/10.1002/9783527335961#page=16.

Hoffmann, Nils, Joel Rein, Timo Sachsenberg, Jürgen Hartler, Kenneth Haug, Gerhard Mayer, Oliver Alka, et al. 2019. "mzTab-M: A Data Standard for Sharing Quantitative Results in Mass Spectrometry Metabolomics." *Analytical Chemistry* 91 (5): 3302–10.

Hu, Qizhi, Robert J. Noll, Hongyan Li, Alexander Makarov, Mark Hardman, and R. Graham Cooks. 2005. "The Orbitrap: A New Mass Spectrometer." *Journal of Mass Spectrometry: JMS* 40 (4): 430–43.

Iakab, Stefania-Alexandra, Gerard Baquer, Marta Lafuente, Maria Pilar Pina, José Luis Ramírez, Pere Ràfols, Xavier Correig-Blanchar, and María García-Altares. 2022. "SALDI-MS and SERS Multimodal Imaging: One Nanostructured Substrate to Rule Them Both." *Analytical Chemistry* 94 (6): 2785–93.

Jurinke, Christian, Paul Oeth, and Dirk van den Boom. 2004. "MALDI-TOF Mass Spectrometry." *Molecular Biotechnology* 26 (2): 147–63.

Knochenmuss, R., and R. Zenobi. 2003. "MALDI Ionization: The Role of in-Plume Processes." *Chemical Reviews* 103 (2): 441–52.

Nikolaev, Eugene N., Yury I. Kostyukevich, and Gleb N. Vladimirov. 2016. “Fourier Transform Ion Cyclotron Resonance (FT ICR) Mass Spectrometry: Theory and Simulations.” *Mass Spectrometry Reviews* 35 (2): 219–58.

Notarangelo, Giulia, Jessica B. Spinelli, Elizabeth M. Perez, Gregory J. Baker, Kiran Kurmi, Ilaria Elia, Sylwia A. Stopka, et al. 2022. “Oncometabolite D-2HG Alters T Cell Metabolism to Impair CD8+ T Cell Function.” *Science* 377 (6614): 1519–29.

Ovchinnikova, Katja, Lachlan Stuart, Alexander Rakhlin, Sergey Nikolenko, and Theodore Alexandrov. 2020. “ColocML: Machine Learning Quantifies Co-Localization between Mass Spectrometry Images.” *Bioinformatics* 36 (10): 3215–24.

Palmer, Andrew, Prasad Phapale, Ilya Chernyavsky, Regis Lavigne, Dominik Fay, Artem Tarasov, Vitaly Kovalev, et al. 2016. “FDR-Controlled Metabolite Annotation for High-Resolution Imaging Mass Spectrometry.” *Nature Methods* 14 (1): 57–60.

Ruttkies, Christoph, Emma L. Schymanski, Sebastian Wolf, Juliane Hollender, and Steffen Neumann. 2016. “MetFrag Relaunched: Incorporating Strategies beyond in Silico Fragmentation.” *Journal of Cheminformatics* 8: 3.

Schramm, Thorsten, Alfons Hester, Ivo Klinkert, Jean Pierre Both, Ron M. A. Heeren, Alain Brunelle, Olivier Laprévote, et al. 2012. “ImzML - A Common Data Format for the Flexible Exchange and Processing of Mass Spectrometry Imaging Data.” *Journal of Proteomics* 75 (16): 5106–10.

Schymanski, Emma L., Junho Jeon, Rebekka Gulde, Kathrin Fenner, Matthias Ruff, Heinz P. Singer, and Juliane Hollender. 2014. “Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence.” *Environmental Science & Technology* 48 (4): 2097–98.

Sementé, Lluc, Gerard Baquer, María García-Altares, Xavier Correig-Blanchar, and Pere Ràfols. 2021. “rMSIannotation: A Peak Annotation Tool for Mass Spectrometry Imaging Based on the Analysis of Isotopic Intensity Ratios.” *Analytica Chimica Acta* 1171 (August): 338669.

Tortorella, Sara, Paolo Tiberi, Andrew P. Bowman, Britt S. R. Claes, Klára Ščupáková, Ron M. A. Heeren, Shane R. Ellis, and Gabriele Cruciani. 2020. “LipostarMSI: Comprehensive, Vendor-Neutral Software for Visualization, Data Analysis, and Automated Molecular Identification in Mass Spectrometry Imaging.” *Journal of the American Society for Mass Spectrometry* 31 (1): 155–63.

Wang, Fei, Jaanus Liigand, Siyang Tian, David Arndt, Russell Greiner, and David S. Wishart. 2021. “CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification.” *Analytical Chemistry* 93 (34): 11692–700.

Wang, Lin, Dan Wang, Olmo Sonzogni, Shizhong Ke, Qi Wang, Abhishek Thavamani, Felipe Batalini, et al. 2022. “PARP-Inhibition Reprograms Macrophages toward an Anti-Tumor Phenotype.” *Cell Reports* 41 (2): 111462.

Wishart, David S. 2016. “Emerging Applications of Metabolomics in Drug Discovery and Precision Medicine.” *Nature Reviews. Drug Discovery* 15 (7): 473–84.

CHAPTER 2:

What are we imaging? Software tools and experimental strategies for annotation and identification of small molecules in Mass Spectrometry Imaging

Gerard Baquer[1,†], Lluc Sementé[1,2,†], Toufik Mahamdi[1], Xavier Correig[1,2,3], Pere Ràfols[1,2,3,*], María García-Altares [1,2]

[1] University Rovira I Virgili, Department of Electronic Engineering, Tarragona, Spain.

[2] Spanish Biomedical Research Center in Diabetes and Associated Metabolic Disorders (CIBERDEM), 28029, Madrid, Spain.

[3] Institut D'Investigacio Sanitaria Pere Virgili, Tarragona, Spain.

† These authors contributed equally to this work.

*Correspondence:

Pere Ràfols, Department of Electronic Engineering, University Rovira i Virgili, Tarragona, Spain. Email: pere.rafols@urv.cat

Mass Spectrometry Reviews e21794 (2022)

https://doi.org/10.1002/mas.21794

Abstract

Mass Spectrometry Imaging (MSI) has become a widespread analytical technique to perform non-labeled spatial molecular identification. The Achilles' heel of MSI is the annotation and identification of molecular species due to intrinsic limitations of the technique (lack of chromatographic separation or the difficulty to apply tandem MS). Successful strategies to perform annotation and identification combine extra analytical steps, like using orthogonal analytical techniques to identify compounds; with algorithms that integrate the spectral and spatial information.

In this review, we discuss different experimental strategies and bioinformatics tools to annotate and identify compounds in MSI experiments. We target strategies and tools for small molecule applications, such as lipidomics and metabolomics.

First, we explain how sample preparation and the acquisition process influences annotation and identification, from sample preservation to the use of orthogonal techniques. Then, we review twelve software tools for annotation and identification in MSI. Finally, we offer perspectives on two current needs of the MSI community: the adaptation of guidelines for communicating confidence levels in identifications; and the creation of a standard format to store and exchange annotations and identifications in MSI.

Keywords: Mass Spectrometry Imaging, metabolomics, molecular annotation, molecular identification, software, identification confidence levels

List of Abbreviations

CCS - Collision Cross-Section, DDA - Data Dependent Acquisition, DESI - Desorption Electrospray Ionization, ESI - Electrospray ionization, FDR - False Discovery Rate, FFPE - Formalin-Fixed Paraffin-Embedded, FT-IR - Fourier-Transform Infrared, FTICR - Fourier-transform Ion Cyclotron Resonance, GC-MS - Gas Chromatography-Mass Spectrometry, HCD - Higher-energy Collision-induced Dissociation, IMS - Ion Mobility Spectrometry, IT - Ion Trap, KMD - Kendrick Mass Defect, LA-ICP - Laser Ablation Inductively Coupled Plasma, LC-MS - Liquid Chromatography-Mass Spectrometry, LCM - Laser-Capture Microdissection, m/z - mass to charge, MALDI - Matrix-Assisted Laser Desorption/Ionization, MS - Mass Spectrometry, MS/MS - Tandem Mass Spectrometry, MSI - Mass Spectrometry Imaging, NMR - Nuclear Magnetic Resonance, NP - Nanoparticle, ROI - Region of Interest, RT- Retention Time, SIL - Stable Isotope Labeling, SIMS - Secondary Ion Mass Spectrometry, t-MALDI - transmission MALDI, TOF - Time-Of-Flight

1. Introduction: the challenge of annotation and identification in MSI

Mass Spectrometry Imaging (MSI) is an analytical technique capable of spatially resolving the chemical composition of biological tissues (Buchberger *et al.*, 2018). Over recent years, MSI has become a key technique in diverse fields such as biochemistry, pharmaceutics, and medical diagnostics (Patti, Yanes and Siuzdak, 2012; Vaysse *et al.*, 2017; Ren *et al.*, 2018; Schulz *et al.*, 2019). Its use in metabolomics, the study of small molecules in biological specimens (Clish, 2015), is of particular interest as metabolites

serve a wide variety of biological purposes such as structural, signaling, immune modulators, endogenous toxins, and environmental sensors (Wishart, 2019).

To draw meaningful biological and diagnostic conclusions from MSI experiments, the mass to charge (*m/z*) ratios obtained need to be traced back to unique compound identifications. This is a non-trivial task considering that spectra in mass spectrometry (MS) are often cluttered with signals from isotopes, adducts, in-source fragments, multiple-charges, matrix, and other exogenous compounds. It is estimated that monoisotopic endogenous peaks only represent 5% of the MS signals in an MSI experiment (Wang *et al.*, 2019). This is particularly challenging in metabolomics since matrix signals and in-source fragments are densely concentrated in the low mass range (Baquer *et al.*, 2020; Janda *et al.*, 2021). The vast amount of MS signals leaves research groups using MSI around the world struggling with the question: “What are we detecting in MSI experiments?”.

Workflows for identification of compounds by other MS-based techniques such as Gas or Liquid Chromatography-Mass Spectrometry (GC-MS and LC-MS) mostly rely on chromatographic separation, followed by MS analysis and often MS/MS experiments. However, these workflows cannot be directly applied to MSI experiments:

1) MSI lacks chromatographic separation: GC-MS and LC-MS use chromatographic columns to separate compounds by their chemical properties (such as polarity) (Lisec *et al.*, 2006; Pitt, 2009) and use retention times (RT) as complementary information to aid compound identification. This information is not available in MSI experiments (Amstalden van Hove, Smith and Heeren, 2010; Yagnik, Korte and Lee, 2013; Buchberger *et al.*, 2018).
2) Most MSI experiments are only performed in Full MS scan: multiple isobars and isomers with different chemical, physical and functional properties can be associated with a given monoisotopic mass (Kyle *et al.*, 2016). Tandem mass spectrometry (MS/MS) can distinguish them by their fragmentation spectra (McLafferty, 1981). Similarly, ion mobility instruments use ion drift times to facilitate the identification of isomers (Mesa Sanchez *et al.*, 2020). In MSI it is still not routinary to perform MS/MS fragmentation and ion mobility separation on-tissue in an untargeted fashion (Amstalden van Hove, Smith and Heeren, 2010; Yagnik, Korte and Lee, 2013; Buchberger *et al.*, 2018).

On the flip side, peak annotation in MSI experiments is statistically more robust given the higher number of data points (each pixel contains a unique spectrum). Spatial correlations between different ion MS signals add statistical confidence to ion annotations (Sementé *et al.*, 2021).

This complex analytical context calls for well-designed experimental strategies and automated software-based solutions to perform robust molecular annotation and identification in MSI metabolomics.

In this review, we explain how each step of the sample preparation and acquisition process influences annotation and identification, from artifacts that may be introduced during sample preservation, to the use of orthogonal techniques like LC-MS/MS with the same tissue. Later, we discuss how different bioinformatics tools annotate and identify compounds in MSI experiments. We specifically target tools for small molecule applications such as lipidomics and metabolomics. This review offers an analytical

background for the bioinformatician to understand the influence of each experimental step on annotation and identification. In turn, analytical chemists will discover the possibilities that bioinformatics offers to support compound annotation and identification in MSI. We also point out how the MSI community struggles to communicate confidence levels for identification and lacks a standard format to report annotations and identifications. As a solution, we propose to adopt the 5 Level scheme by *Schymanski et al*. (Schymanski *et al.*, 2014), and we draft a file format annex to imzML based on mzTab-M (Hoffmann *et al.*, 2019) to report annotations and identifications in MSI.

2. The need for reporting standards in MSI

2.1. A word about the terms *annotation* and *identification*

According to the Metabolomics Standards Initiative, a non-novel molecule is considered "identified" when its experimental data is compared to a standard by at least two types of orthogonal data (for instance, RT and MS/MS), while a compound would be considered "annotated" if identification is not achieved (Sumner *et al.*, 2007). A common problem in metabolomics (Salek *et al.*, 2013) and MSI scientific articles is that the terms annotation and identification are sometimes used interchangeably, at times even accompanied by the adjectives "putative" or "tentative". This confusion impedes the comparison of different annotation/identification strategies and the interpretation results.

To seize the impact of this problem in the MSI community, we reviewed the usage of the terms "annotation" and "identification" in 58 papers published in the last 5 years (Table S1 in supplementary materials) dealing with annotation/identification from several perspectives (bioinformatics, experimental protocol, instrumental and application).

We found that 52% of the papers use the term "identification" to refer to exact mass matching at least once (when "annotation" should be used). Moreover, the adjectives "putative" and "tentative" are used in 31% of the papers. When they appear, they accompany the terms annotation and identification indistinctly to refer to exact mass matching.

2.2. Adaptation of identification confidence levels for MSI

Communicating the degree of confidence in compound identification is essential to avoid misinterpretation of the results, and to compare identification strategies. While the MSI community has its own initiative for improving standardization and reproducibility (MALDISTAR, https://www.maldistar.org/), now the aims of this initiative do not include the definition of guidelines for reporting the confidence of compound annotation and identification. Besides, current reporting standards for mass spectrometry imaging (McDonnell *et al.*, 2015; Gustafsson *et al.*, 2018) do not explicitly mention identification confidence levels. The 2015 guideline proposed by *McDonnell et al*. (McDonnell *et al.*, 2015) defines the minimum reporting standards for identifications as (1) experimental and theoretical *m/z*, (2) mass tolerance, (3) MS/MS on-tissue, and (4) orthogonal measurements (i.e. LC-MS/MS). However, this scheme does not communicate different degrees of confidence in MSI identifications and annotations.

On the other hand, the metabolomics community does have well-accepted guidelines for communicating identification confidence based on the four-level system suggested by the Metabolomics Standards Initiative in 2007 (Sumner *et al.*, 2007). In 2014,

Schymanski et al. (Schymanski *et al.*, 2014) proposed a 5 level system to rank levels of confidence in identification: (Level 1) Confirmed structure matched against a reference standard (MS, MS/MS, and RT); (Level 2) Probable structure matched against literature or library spectrum (MS, MS/MS, and RT); (Level 3) Tentative candidates matched against literature or library spectrum (MS, MS/MS, and RT); (Level 4) Unequivocal molecular formula (MS with adduct and isotope information); (Level 5) Exact mass (MS). Later, *Schrimpe-Rudletge et al*. (Schrimpe-Rutledge *et al.*, 2016) expanded the model by proposing the use of orthogonal techniques, such as Nuclear magnetic resonance (NMR) or ion mobility, to reach level 2 and level 3 identifications.

The scheme of 5 confidence levels used in metabolomics (Schymanski *et al.*, 2014; Schrimpe-Rutledge *et al.*, 2016) shown in Figure 1 could be adopted to report identification confidences in MSI experiments. As the information obtained by MSI is different from the data collected by common metabolomics techniques (usually based on chromatographic separation), we suggest the following adjustment of the 5 level system (from highest to lowest confidence) to report identification confidence in MSI experiments:

Level 1 Confirmed structure: Reporting exact mass, unequivocal molecular formula, and a single confirmed structure. At this level, a unique structure is confirmed by comparing all experimental data from Levels 2-5 to reference standards. The use of reference standards for confirming identifications in MSI may include spotting the standard on the glass slide or substrate, on a replicate tissue, or spiking a homogenized replicated tissue. Alternatively, one can perform LC-MS/MS measurements of tissue homogenates or microdissection of the tissue to compare against standards dissolved in solvents or in tissue extracts (matrix-matched comparison).

Level 2 Probable structure: Reporting exact mass, unequivocal molecular formula, and a single possible structure. This level is achieved when only one unambiguous possible structure results after following the procedures described in Level 3.

Level 3 Tentative candidates: Reporting exact mass, unequivocal molecular formula, and a list of possible structures. This level requires information complementary to the MS measurement that can be obtained using orthogonal data, obtained during the MSI experiment (like ion mobility or MS/MS fragmentation) or by orthogonal techniques such as LC-MS/MS on homogenized tissues, or complementary molecular imaging techniques. If MS/MS is used, the obtained experimental spectra are matched against experimental, *in-silico* or literature libraries.

Level 4 Unequivocal molecular formula: Reporting exact mass and unequivocal molecular formula. This requires the integration of MS information such as isotopes, adducts, and/or in-source fragments. In MSI, the annotation of isotopes, adducts, and in-source fragments benefit from the high number of sampling points over the tissue. The spatial correlation of signals (not available in other MS methods) ensures robust Level 4 annotation.

Level 5 Exact mass of interest: Reporting only the exact mass of the compound, together with the mass tolerance of the MSI method. Unable to distinguish between different molecular formulas within the mass tolerance of the method.

3. Influence of the sample preparation and spectra acquisition procedures for molecular annotation and identification

This section covers the influence of experimental procedures in compound annotation and identification in MSI. It describes experimental strategies regarding sample preparation, instrumental setups, and combinations of MSI with other techniques. It provides a solid analytical background for bioinformaticians working in MSI annotation and identification. For a deeper explanation of MSI experimental procedures, the reader is referred to more extensive reviews (Amstalden van Hove, Smith and Heeren, 2010; Chatterji and Pich, 2013; Gode and Volmer, 2013; Norris and Caprioli, 2013; Buchberger *et al.*, 2018). Table 1 contains a compendium of the principal effects in annotation/identification of all the procedures and instruments covered in this section.

3.1. Effects of the sample preparation in MSI annotations and identifications

Sample preparation is a critical step in any MSI experiment, as it largely influences which compounds will be ionized and detected. Proper sample preparation will also reduce ion suppression, adduct formation, matrix interferences, and in-source fragmentation. Besides, the use of calibrants improves the mass axis calibration and increases the confidence of annotations by exact mass.

3.1.1. Sample preservation

Sample preservation is the first decision that affects an MSI experiment, as it determines what type of compounds will remain in the tissue. There are three main preservation options: formalin-fixed paraffin-embedded (FFPE) tissues, fresh-frozen tissues, and formalin-fixed frozen tissues.

FFPE tissues have been the gold standard for the fixation and storage of samples for histopathological analyses. FFPE tissues can be preserved at room temperature for years without degradation and are easy to section and transport thanks to the wax embedding. Nevertheless, paraffine induces ion suppression during the ionization process in MSI, and formalin fixation (which cross-links proteins together) hampers the desorption/ionization of proteins and peptides. Moreover, both compounds contaminate the spectra by adding more signals. Thus, the use of FFPE tissues for MSI requires the removal of the paraffine before MSI analysis (by a series of xylene and ethanol washing steps); and the reversal of the cross-linking of proteins (by antigen retrieval protocols). These washing steps lead to the loss of lipids and metabolites, thus FFPE tissues are better suited for peptide and protein analysis by MSI.(Wisztorski *et al.*, 2010; Ly *et al.*, 2016; Hermann *et al.*, 2020)

Fresh-frozen tissues have the advantage of stopping post-mortem decay (autolysis) without using any chemical agent that may induce changes in the tissue. In principle, this allows the preservation of all the molecular species in the tissue, thus enabling the detection of metabolites, lipids, and proteins. This makes fresh-frozen the standard

sample preservation for MSI. Nevertheless, fresh-frozen samples are costly to store, as they require -80ºC freezers to avoid the rapid deterioration in room temperature. This makes the sample vulnerable to power outages and mechanical failures in the closing door.

Formalin-fixed frozen tissue is a combination of both previous approaches. In this case, the sample is fixed by formalin, but it is stored as fresh-frozen tissue without paraffin embedding. Heat-induced antigen retrieval protocols can be used to avoid metabolite loss (Groseclose *et al.*, 2008), but formalin may reduce the ionization yield of amine-containing lipids, and generate [M+HSO_4]- adducts (Vos *et al.*, 2019). Using this sample preservation, it is possible to measure compounds in all mass ranges although with lower effectiveness than fresh-frozen tissues for the low mass range (Pietrowska *et al.*, 2016).

3.1.2. On-tissue enzymatic digestion of intact proteins

MSI analysis of intact proteins is usually restricted to those molecules below 25 kDa (although some MALDI matrices like ferulic acid can extend this range (Mainini *et al.*, 2013)), thus classical top-down proteomic strategies may not be efficient in MSI. Thus, on-tissue enzymatic digestion is included in most protein identification routines, which allow larger proteome coverage identification. This bottom-down approach is based on spraying or spotting enzymes (usually trypsin) over the tissue to cleave the proteins into their peptides, followed by an incubation step (Cillero-Pastor and Heeren, 2014; Diehl *et al.*, 2015). Besides trypsin, other enzymes can be used to digest proteins, such as the enzyme peptide-N-glycosidase F for N-glycan profiling (Drake *et al.*, 2018). Sequencing the detected peptides by common MS/MS approaches can help both identify and spatially locate proteins directly on the tissue. Previous reviews on protein identification in MSI (Mascini and Heeren, 2012; Ryan, Spraggins and Caprioli, 2019) have covered this topic in depth.

3.1.3. On-tissue chemical derivatization

Some compounds are difficult to detect using MSI due to their low ionization efficiency, ion suppression, low concentration, and/or small molecular weight. Sample preparation steps (i.e. the proper matrix selection in MALDI MS or solvent selection in DESI-MS) might alleviate this concern. On-tissue chemical derivatization applies reagents over a tissue section to modify the chemical structure of specific compounds and enhance their detectability, by adding moieties with specific properties. For instance, adding a charged moiety often counteracts low ionization efficiency problems. Ion suppression due to low molecular weight can also be avoided by the reaction of the target compound and a derivatization molecule, which increases the analyte *m/*z ratio. All these mechanisms alter the detectability of specific compounds and therefore, the capacity of annotating and identifying them. *Harkin et al*. review concrete examples of these procedures (Harkin *et al.*, 2021). For instance, pyrylium salts react selectively with primary amines in neurotransmitters, thus they can be incorporated into matrices (Shariatgorji *et al.*, 2015) or synthesized as bromopyrylium to introduce a distinctive isotopic pattern only in targeted neurotransmitters (Shariatgorji *et al.*, 2020).

3.1.4. Matrix selection and deposition in MALDI MSI

In MALDI MSI, matrices are compounds that assist the desorption/ionization of analytes from the tissue. Most common applications use small organic compounds as matrices

that are either sprayed or sublimated over the tissue (Gemperline, Rawson and Li, 2014). MALDI matrix application techniques should ensure good homogeneity of the deposited layer and minimize in-tissue compound delocalization to get high-quality images.

The selection of appropriate matrices and optimization of the deposition method greatly affect the outcome of MALDI MSI analysis and the annotation and identification of analytes.

Matrices may introduce undesired effects that clutter the mass spectra and hamper compound annotation, such as matrix clusters, matrix adduct formation, and detector saturation. This is a particular issue in the low mass range where matrix-metabolite adducts can explain a considerable amount of non-annotated peaks (Janda *et al.*, 2021). Lipidomics and metabolomics identification routines are very sensitive to the matrix method used (Fernández *et al.*, 2011)(Thomas *et al.*, 2012).

The selection of the matrix will define the ionization polarity mode. For instance, MALDI matrices with an acidic group (like benzoic acid and cinnamic acid derivatives) are mostly used in positive ionization mode, while matrices that are basic and contain amino functions tend to be used in negative ionization mode. The ionization mode will favor the detection of specific compounds, for example, lipids with a polar headgroup like phosphatidylcholines will be detected in positive mode, while glycophosphoinositol will have better ionization yield in negative mode (Leopold *et al.*, 2018). To increase the coverage of the lipidome, several research groups opt for the use of matrices and acquisition modes that allow dual polarity MALDI MSI analysis on the same sample (Kaya *et al.*, 2018; Li *et al.*, 2019; Huang *et al.*, 2020).

Developing new matrices is a hot research field in MSI. While classical first-generation matrices like alpha-Cyano-4-hydroxycinnamic acid and 2,5-Dihydroxybenzoic acid are still widely used, the design of second-generation and reactive matrices (simultaneously a derivatization reagent and a matrix) allow the selective desorption/ionization of specific analytes. The analytes of interest are detected with higher signal-to-noise ratios and sometimes present specific spectra features (such as a distinctive isotopic pattern) that facilitate their annotation and identification. The reviews by *Zhou* et al. and *Calvano* et al. provide an excellent reference on selective matrices for MSI metabolomics and lipidomics (Calvano *et al.*, 2018; Zhou, Fülöp and Hopf, 2021).

On the other hand, inorganic nanoparticles (NPs) (of gold and silver, among others), as well as some metal-oxides (TiO_2, CeO_2, etc.), have been proposed as an alternative to organic matrices for the analysis of small molecules by MSI (Abdelhamid, no date; Basu *et al.*, 2019). They often produce fewer matrix clusters and adducts, leading to a cleaner background spectrum. Additionally, their distinctive carbon-free isotopic pattern and easily identifiable peaks can serve as internal calibrants during data processing (Nizioł and Ruman, 2013; Ràfols, Castillo, *et al.*, 2018; Ràfols, Vilalta, Torres, *et al.*, 2018).

Matrix deposition is one of the most important sample preparation steps toward the production of high-quality ion images. Researchers use different techniques to apply matrices onto the target tissue, including spray (Khatib-Shahidi *et al.*, 2006; Norris *et al.*, 2007) and sublimation (Hankin, Barkley and Murphy, 2007; Thomas *et al.*, 2012) for organic matrices, and sputtering for NPs (Dufresne *et al.*, 2013; Ràfols, Vilalta, Torres, *et al.*, 2018). The spray method is based on applying the matrix solution into the tissue section manually (DeKeyser *et al.*, 2007; Ye *et al.*, 2013) or using automated spray devices allowing controllable solvent flow rate and matrix layers number (Mounfield and

Garrett, 2012; Gemperline, Rawson and Li, 2014; Phan *et al.*, 2016). Sublimation is a dry deposition technique (the transition of one chemical substance from the solid phase to the gas phase without passing through the intermediate liquid phase), in which matrices are sublimated and deposited under reduced pressure and specific elevated temperature parameters, leading to the deposition of dry matrix layer on tissue target (Hankin, Barkley and Murphy, 2007; Nakamura *et al.*, 2017). However, sublimation alone is not sufficient for the ionization of some compound species, such as proteins, therefore a re-hydration or re-crystallization step is needed in order to promote the integration of these molecules with the matrix crystals (Yang and Caprioli, 2011).

Sputtering is a thin film deposition process where inorganic NPs or metal-oxide targets (such as gold or silver) are bombarded with high-energy ions in a vacuum chamber resulting in the condensation of the target atoms on the substrate tissue section as thin layers (Ogrinc Potočnik *et al.*, 2014; Hansen, Dueñas and Lee, 2019).

3.1.5. Stable Isotope Labeling

Stable Isotope Labeling (SIL) consists of the synthesis of compounds containing atoms with artificial isotopic abundances highly dissimilar to the ones that occur in nature. Common isotope labels include ^{13}C, ^{15}N, and deuterium (^{2}H). This technique has many applications in several aspects of MSI (Grey *et al.*, 2021) such as tracing of drugs and metabolites (Eckelmann, Kusari and Spiteller, 2018; Ellis *et al.*, 2021). Additionally, the labeled compounds introduced in the sample can be used as internal standards to normalize signal intensity (Chumbley *et al.*, 2016; Barry *et al.*, 2019) and provide quantitative results (Grey *et al.*, 2019).

For annotation, one of the most relevant applications is SIL MALDI matrices. By isotopically labeling the matrix, their background signals can be shifted and uncover relevant endogenous signals. Additionally, their distinct isotopic pattern can be exploited to develop more robust annotation tools. As an example, *Shariatgorji et al*. (Shariatgorji *et al.*, 2012) managed to shift the matrix peaks by using deuterated CHCA to uncover and annotate several neurotransmitters.

3.2. MSI image acquisition

Mass spectrometers intrinsically affect the annotation and identification procedures, as they determine which species of ions will be generated in the ion source, and the *m/z* resolving power and accuracy. The parts of the mass spectrometer that affect the annotation/identification process are the ion source, responsible for the desorption and ionization of the molecules, and the mass analyzer, responsible for the determination and counting of the *m/z* ratio of the ions. Figure 2 shows a broad comparison between the main ion sources and mass analyzers.

3.2.1. Ion source

The ion source induces the desorption of the analytes from the tissue, and the ionization of compounds that will be transferred into the mass analyzer. Depending on the polarity of the electrical field applied in the ion source, the ions formed will be positive (usually protonated adducts and adducts with cations, such as Na^+ and K^+) or negative (like deprotonated adducts and adducts with anions, such as Cl^-). The different technologies

result in differences in the mass range analyzed, the number of charges of the produced ions, the number of in-source fragments generated, and the sensitivity to detect low concentration compounds. Spatial resolution and sensitivity are related concepts, as increasing the spatial resolution results in decreasing the ablated area and therefore, reduces the sensitivity. In MSI, the most used ion sources are Matrix-assisted Laser Desorption/Ionization (MALDI), Desorption Electrospray Ionization (DESI), Secondary Ion Mass Spectrometry (SIMS) and Laser Ablation Inductively Coupled Plasma (LA-ICP).

MALDI sources ionize the sample using a pulsating laser (usually UV or IR) inside a vacuum or low-pressure chamber with the assistance of the previous matrix deposition. The laser strikes the sample and generates a plume of charged ions that are directed to the mass analyzer. MALDI sources tend to produce low fragmentation and singly charged ions (Karas, Glückmann and Schäfer, 2000; Jaskolla and Karas, 2011), which enable the ionization of metabolites ('Compound and metabolite distribution measured by MALDI mass spectrometric imaging in whole-body tissue sections', 2007), lipids (Züllig and Köfeler, 2021), peptides (Phillips, Gill and Baxter, 2019) and proteins ('Ultra-high resolution MALDI-FTICR-MSI analysis of intact proteins in mouse and human pancreas tissue', 2019), and usually achieve spatial resolutions in the range of 100 to 10 µm and close to 1 µm with specific setups (Hansen and Lee, 2017; Kompauer, Heiles and Spengler, 2017; Wäldchen *et al.*, 2020). In recent years, enhanced versions of MALDI sources have been proposed, like MALDI-2 (Soltwisch *et al.*, 2015; Heijs *et al.*, 2020), which increases the sensitivity of the MALDI source by adding a second post-ionization laser that ionizes the neutral molecules in the ion plume; transmission MALDI (t-MALDI) (Trimpin *et al.*, 2009; Zavalin *et al.*, 2012, 2015; Steven *et al.*, 2019), which increases the later resolution up to 1 µm and below by changing the laser focus geometry; and more recently t-MALDI-2 (Niehaus *et al.*, 2019; Bien *et al.*, 2021; Dreisewerd, Bien and Soltwisch, 2022), which combines the benefits of both improved designs.

DESI sources produce ions at atmospheric pressure conditions directing a spray of charged microdroplets directly into the tissue. DESI sources require minimal sample preparation. They are commonly used to analyze small molecules and lipids, but bigger compounds like peptides and proteins can also be analyzed (Towers *et al.*, 2018), although most solvents used with DESI denature proteins, affecting the three-dimensional structure (Hale and Cooper, 2021). Typically, DESI sources achieve spatial resolutions in the range of 200 to 20 µm (Ifa *et al.*, 2007; Claude, Jones and Pringle, 2017; Nguyen *et al.*, 2018; Towers *et al.*, 2018; Zhang *et al.*, 2020) and are known to produce little fragmentation and singly charged ions (Towers *et al.*, 2018).

SIMS sources bombard samples using an ion beam, ionizing molecules from the sample surface and ejecting them into the vacuum environment but, due to the high energy of the beam, SIMS sources easily cause the fragmentation of the molecular ions (Yoon and Lee, 2018). Currently, SIMIS sources provide the greatest spatial resolution for MSI, reaching the nanometer scale (Gamble and Anderton, 2016), but have less sensitivity, as the area ablated is lower than other technologies. Applications of SIMS sources are principally focused on small metabolites and lipids (Touboul and Brunelle, 2016) and like DESI, require minimal sample preparation.

LA-ICP sources use an inductively heated plasma to atomize molecules ablated from a specific region, generating atomic composition maps over the sample. LA-ICP is

generally used to track metals in biological sections with a spatial resolution between 200 and 10 µm (Becker *et al.*, 2011, 2012; Pornwilard *et al.*, 2013; Sabine Becker, 2013). In terms of fragmentation, LA-ICP fragments all the compounds in the sample to their atomic composition, resulting in null preservation of precursor ions.

3.2.2. Mass analyzer

The mass analyzer detects the ions generated by the source, determines the mass-to-charge ratio of them, and composes the spectrum at each sample position or pixel of the image. There are three parameters that influence the identification of compounds for each mass analyzer: (1) mass range, the lowest and highest *m/z* that the mass analyzer can detect; (2) mass accuracy, the difference between the measured *m/z* of an ion and the real *m/z* (usually described in ppm); and (3) mass resolution, the ability to distinguish between ions separated by small *m/z* values, often defined as the *m/z* of a peak divided by the peak width at 10% or 50% of peak height. The most common mass analyzers in MSI systems are time-of-flight (TOF), Fourier-transform ion cyclotron resonance (FTICR), and Orbitrap.

TOF mass analyzers are vacuum tubes in which ions travel through an electric field to the detector. The longer the tube, the higher the mass resolution of the spectra, as the ions have more time to gain distance between them during the flight. Despite this, TOF mass analyzers tend to have lower mass resolution compared to other mass analyzers used in MSI, as enhancing it implies an increase in the physical size of the whole MSI system and in the sampling time. With reflectron set-ups, the mass resolution can be increased, but still lower than other analyzers. Moreover, TOFs are very susceptible to temperature changes, as the metal tube may suffer expansions and contractions that affect the mass accuracy of sampled ions. On the other hand, TOF analyzers do not have a theoretical upper *m/z* detection limit like other mass analyzers (Xian, Hendrickson and Marshall, 2012), and their mass resolution increases within the mass range. TOF mass analyzers are extensively used with MALDI ion sources to image almost any kind of compounds, with a preference for compounds in the high mass range like peptides and proteins, with a typical upper limit of *m/z* 30,000 (Spengler, 2015). Common set-ups of TOF mass analyzers are MALDI-TOF, MALDI-TOF/TOF, MALDI-Q-TOF, and TOF-SIMS.

FTICR mass analyzers use a magnetic field to resonate the ions into cyclotron orbits and transduce the orbiting frequencies into *m/z* using the Fourier Transform. These mass analyzers are built around powerful magnets; the stronger the magnetic field, the greater the mass resolution, reaching values of up to 1,600,000 at *m/z* 400 for a 21T magnet (Bowman *et al.*, 2020) with mass accuracies below 1 ppm. FTICR mass analyzers are used to analyze all families of compounds, but preferably not higher than *m/z* 3000, as the mass resolution decreases as the *m/z* ratio increases (Almeida *et al.*, 2015) and the magnetic field and sampling time required to detect these ions are high. Still, there are examples of high mass protein MSI investigations up to *m/z* 30,000 using a 15T FTICR mass analyzer with a mass accuracy below 10 ppm and transients close to 4 seconds per pixel (M. Dilillo *et al.*, 2017). Common set-ups of FTICR mass analyzers are MALDI-FTICR and DESI-FTICR.

Orbitrap mass analyzers use electrically charged ion trap cells to excite the ions into orbits. The longitudinal movement of the orbits contains the information of the cyclotron frequencies of each ion, which can be converted to mass using the Fourier transform.

Orbitraps achieve high mass resolution values by increasing the electric field. With Orbitraps it is possible to analyze a wide range of compounds but, as FTICR, high mass compounds are typically excluded as the mass resolution decreases by the square root of the *m/z* ratio and require long sampling times and strong fields to compensate for this (Bielow *et al.*, 2017). Common set-ups of Orbitrap mass analyzers are DESI-Orbitrap and MALDI-Orbitrap.

3.3. Combinations of MSI with other analytical techniques (Level 2-3 Identification)

To ensure high levels of confidence in molecular identification with MSI, a common strategy is to examine the tissue with additional or orthogonal techniques (those based on fundamentally different principles). LC-MS and tandem mass spectrometry (MS/MS) are the most used confirmatory techniques. Recently, ion mobility has been included in commercial MSI instruments to provide an additional dimension for metabolite analysis and resolve isomers (Meier *et al.*, 2015, 2020; Łącki *et al.*, 2021). Finally, the combination of different imaging techniques coupled to MSI has been used to improve the identification process. Multimodal imaging combines non-destructive orthogonal analysis like immunohistochemistry, immunofluorescence, or vibrational spectroscopy imaging techniques with MSI (Iakab *et al.*, 2021; Tuck *et al.*, 2021).

3.3.1. MS/MS

MS/MS uses a combination of ion traps, mass analyzers, and fragmentation chambers to measure fragments of molecules and reveal their structure. The typical setup is two consecutive mass spectrometers separated by a fragmentation chamber. The first mass spectrometer is in charge of recording the ionization product of an ion source that keeps the precursor compounds with low fragmentation. Later, some of the precursor ions are directed to a collision chamber to achieve a controlled fragmentation. The resulting fragments are registered in a second mass analyzer to obtain the fragmentation spectra of all the selected precursors. By knowing the precursor *m/z* value and examining the fragmentation spectrum, it is possible to provide hypotheses about the structure of the compound and hence its identification.

In MSI, MS/MS analysis can be performed in some instruments achieved either by sampling consecutive slides in MS/MS mode (Dueñas *et al.*, 2017) or adjacent regions in the same slide (Zhan *et al.*, 2021), which can be a problem if there are very localized compounds or limited sample material. Common set-ups are based on TOF/TOF and Q-TOF devices, commonly used for top-down proteomics (Alam, Kumar and Kamboj, 2012; Ye *et al.*, 2014; Xu *et al.*, 2019).

To overcome these limitations, new methods have been investigated in recent years. Multiplex MSI has achieved to overlap scans of MS and MS/MS in the same place using a spiral pattern and proved to be used for 10 µm high-spatial-resolution imaging of maize leaf cross-sections in both the high and low mass ranges for a variety of metabolites (Perdian and Lee, 2010; Yagnik, Korte and Lee, 2013; Hansen and Lee, 2017). *Ellis et al*. developed an automatic structural identification workflow consisting of parallel acquisition of a MALDI-Orbitrap instrument with an ion trap (IT)-MS/MS (Ellis *et al.*, 2018). *Lanekoff et al.* coupled a nano-DESI source with a high-resolution Q-Exactive Orbitrap and a Higher-energy Collision-induced Dissociation (HCD) cell to identify and image isobaric and isomeric species combining the MSI and the MS/MS data (Lanekoff

et al., 2013). Finally, *Fu et al.* were able to analyze and image by tandem MS the molecular products of natural biosynthesis of rubrynolide and rubrenolide in Amazonian trees using a TOF-SIMS and a triple ion focusing time-of-flight (TRIFT) analyzer with a precursor selection window of a monoisotopic ion, which allow the parallel and lossless collection of MS and MS/MS data (Fu *et al.*, 2018).

Despite all the efforts, MS/MS is rarely used with MSI data as many commercial instruments still do not include this option. Moreover, the concentration of precursors is limited to the area covered by the scans, which might be low for some compounds.

3.3.2. LC-MS

LC-MS incorporates chromatographic separation before the mass analyzer. RT allows differentiation of the compounds based on criteria other than *m/z*, like polarity or compound size. Most LC-MS systems use tandem MS and can provide fragmentation information on the analytes.

The combination of LC-MS with MSI is one of the most common approaches used to identify and spatially visualize a compound in all kinds of metabolomics and lipidomics experiments (Garate *et al.*, 2020). The identification workflow usually consists of homogenizing some of the tissue samples to identify as many compounds as possible with the LC-MS instrument (Baijnath *et al.*, 2016; Shobo *et al.*, 2016; Ntshangase *et al.*, 2019). Later, the identified compounds are searched in the MSI spectra by exact mass matching.

Other approaches combine LC-MS with laser-capture microdissection (LCM), which allows the isolation and compound profiling of specific cells or tissue regions of interest (ROIs) determined by MSI (Marialaura Dilillo *et al.*, 2017; Dewez *et al.*, 2019). This approach ensures that the LC-MS identifications come from the same region in the tissue that was mapped by MSI.

3.3.3. Ion mobility spectrometry

Ion mobility spectrometry (IMS) is a technology that separates ions according to their size, shape, and weight by directing and colliding them into a chamber filled with an inert gas. The collision cross-section (CCS) value is computed from the time each ion takes to reach the end of the chamber. In combination with MS, IMS can be used as an additional dimension of information to resolve isomeric species, improve selectivity and get structural information of compounds, including metabolites (Lapthorn, Pullen and Chowdhry, 2013). *Sans et al*. reviewed an extensive number of applications and advances combining MSI and IMS for biological applications (Sans, Feider and Eberlin, 2018).

3.3.4. Multimodal molecular imaging

Other molecular imaging techniques can provide the orthogonal chemical information needed to provide structural identification of *m/z* features (Porta Siegel *et al.*, 2018).

Vibrational Spectroscopy Imaging techniques (i.e. Raman and Fourier-Transform Infrared (FT-IR)) measure the energy scattering and absorption of different lasers to determine functional groups and other chemical features (Harrison and Berry, 2017). This structural information is rarely enough to fully resolve isomers, but it can be used to discard candidates and achieve Level 3 annotation. As an example, Lasch and Noda

(Lasch and Noda, 2017) applied Raman, FT-IR, and MSI to study the composition of the hamster brain. They could identify and spatially locate several lipids by the spectral correlation between Raman bands (for instance, bands 548 and 703 cm^{-1} for cholesterol) and *m/z* features (*m/z* 369.30 for [Cholesterol–H_2O+H]$^+$).

Fluorescence Microscopy techniques enable imaging of specific compounds by labeling them with fluorescent probes (Lichtman and Conchello, 2005). Cyclic or multiplexed immunofluorescence images the same sample with dozens of different fluorescent probes (Lin *et al.*, 2016). Highly selective fluorescent probes (Li, Liu and Wang, 2011; Uslu *et al.*, 2017; Dong *et al.*, 2020) can target specific isomers and enable Level 3-2 annotation. For instance, *Fuch et al.* (Fuchs *et al.*, 2018) monitored the biodistribution of the anticancer drug sunitinib and its metabolites in rabbit liver tissue using fluorescence to measure the total amount of the drug, and MSI to characterize *in-situ* the presence of its metabolites.

3.4. Validation against reference standards in MSI (Level 1 Identification)

According to the system for reporting identification confidence in MSI (section 2.2.), to achieve Level 1 identification (highest level of confidence), the experimental data (MSI and orthogonal technique of choice) must be matched against a reference standard. One common strategy in MSI experiments is to homogenize the tissue, spike it with the reference standard of the compound of interest, and measure it with LC-MS/MS (Baijnath *et al.*, 2016; Shobo *et al.*, 2016; Ntshangase *et al.*, 2019). Using LCM, the tissue homogenates can be obtained from specific tissue ROIs selected by MSI (Marialaura Dilillo *et al.*, 2017; Dewez *et al.*, 2019). Nevertheless, even when using LCM, homogenizing the tissue leads to the loss of the spatial information provided by MSI. Additionally, due to differences in their ionization, LC-MS/MS and MSI data may not be directly comparable (i.e. the analytes of interest may form different adducts in each system, etc.).

Full confirmation of MSI identifications requires strategies to measure reference standards directly in MSI. Most of the developments in this area have been conducted for the study of synthetic drugs and their metabolites *in-situ* (Buck *et al.*, 2015; Groseclose *et al.*, 2015) but they are largely applicable to endogenous neurotransmitters (Shariatgorji *et al.*, 2014), metabolites (Pirman *et al.*, 2013), lipids (Jadoul *et al.*, 2015), and peptides (Zhang, Kuang and Li, 2013). In general there are three strategies (Rzagalinski and Volmer, 2017; Unsihuay, Mesa Sanchez and Laskin, 2021): (1) “in-solution” (2) “on/under tissue” and (3) “mimetic tissue”.

The “in-solution” strategy is the most straightforward of the three, as the standard is spotted directly on the substrate next to the sample. This method will inform about isotopic patterns, general adducts, matrix adducts, and in-source fragments that can be formed with the analyte of interest during the MSI experiment. However, it fails to capture endogenous adduct formation and ion suppression effects. As an example, the in-solution strategy was used for identifying the drug Erlotinib and its metabolites in rat tissue sections (Signor *et al.*, 2007).

The “on/under tissue” strategy alleviates these limitations by spotting the standard beneath or on top of the tissue. Normally, this is performed on a control tissue, preferably

a consecutive slice. If allowed by the application (i.e. in synthetic drug applications), the control tissue should be blank and not contain the endogenous compound to be compared to the reference standard. As a variation of this approach, some studies apply the standard mixed with the MALDI matrix. As an example, the "on-tissue" approach has been used to identify the drug paclitaxel in the study of pleural tumors (Giordano *et al.*, 2016), glutathione in ovarian tissue (Nazari *et al.*, 2018), and raclopride and SCH 23390 in rat brain tissue (Goodwin *et al.*, 2011).

Finally, the "mimetic tissue" approach relies on homogenizing the tissue and spiking it with the standard. This mixture is then deposited on the MSI slice and treated with the same sample preparation protocol. This approach provides a more realistic scenario on how the analyte behaves during the MSI experiment, as the standard is fully mixed within the sample. One drawback is that it fails to capture differences in matrix and suppression effects across anatomical regions. The mimetic tissue approach has been successfully used for identifying the drugs lapatinib and nevirapine in rat liver (Groseclose and Castellino, 2013), GSH in human ocular lens tissue (Grey *et al.*, 2019), and clozapine and norclozapine in rat liver (Barry *et al.*, 2019).

4. Bioinformatics strategies for annotation and identification in MSI

In this section, we discuss automated data processing strategies for annotation and identification in MSI. We first start by discussing the importance of preprocessing to ensure robust annotation and identification. Later, we provide a wide picture of the basic principles in the development of software-based annotation and identification. We close the section with a comprehensive comparison of twelve software tools developed in the last 5 years.

4.1. Data-preprocessing

Good quality MSI data is crucial to conduct successful molecular annotation and identification (Norris *et al.*, 2007). As stated in the previous section, careful analytical design is key, as it will set the boundaries of what is possible in compound identification. But even when the analytical procedure is carefully designed and executed, variability due to experimental factors can worsen data quality. Chemical noise and variations in the intensity and exact mass of each MS feature are some of the examples of unwanted experimental variability. Additionally, when dealing with large samples and high spatial resolution, MS intensities and *m/z* values can drift during the long acquisition (Ràfols, Vilalta, Brezmes, *et al.*, 2018). Proper data preprocessing mitigates these negative effects and enhances the chances of correct identification.

The typical preprocessing workflow includes the following steps: baseline correction, noise reduction, spectral alignment, normalization, peak picking, and binning (Ràfols, Vilalta, Brezmes, *et al.*, 2018). Depending on the experiment, some steps may be performed in a different order or even be omitted. The resulting processed data can come in two forms: (1) profile data retains the continuous shape of the spectra, as no peak picking is performed, and (2) centroid data only retains certain features of each peak (commonly the *m/z* and maximum intensity value) after peak picking.

Calibration (a form of spectral alignment) is the most relevant step for annotation and identification, as it increases the mass accuracy of the measured *m/z*. In calibration, a list of known *m/z* values is used to compute a warping function that minimizes the *m/z* error in the MSI dataset. The calibration *m/z* values can come from reference standards spotted on the plate (phosphorus red) (Paine *et al.*, 2019), the matrix or ionization promoter (Ràfols, Castillo, *et al.*, 2018; Ràfols, Vilalta, Torres, *et al.*, 2018), or well-characterized endogenous compounds (He *et al.*, 2019). Additionally, label-free alignment can further improve data quality. In this case, a reference spectrum from within the sample is used to minimize the *m/z* errors between pixels.

All MSI instrumentation vendors provide in-house software capable of performing to some extent this preprocessing pipeline. SCiLS (Trede *et al.*, 2012) by Bruker is one of the most widely used commercial solutions. Several open-access alternatives such as CARDINAL (Bemis *et al.*, 2015), rMSIproc (Ràfols *et al.*, 2020), and MALDIQuant (Gibb and Strimmer, 2012) have gained importance over recent years.

4.2. Basic software-related principles in annotation and identification of MSI

Figure 3 shows the general workflow of annotation and identification software tools in MSI. Each of the steps increases the level of confidence and relies on different experimental data and libraries.

There are various basic concepts to consider while designing or choosing an annotation tool for MSI data. How input data represent each *m/z* feature, the direction of the flow of information between data and libraries, how to match the information in the libraries, and how to use the annotation or share them. The following section comments on various of these topics.

4.2.1. Working with profile vs. centroided data

Molecular annotation and identification can either be performed on profile or centroided spectra. Profile spectra provide richer information: (1) they keep potentially relevant small and noisy peaks, (2) they retain peak shape, and (3) they enable overlapped peaks to be recognized and eventually deconvoluted (Polanska *et al.*, 2012). The main problem with data in profile mode is the higher computational load, which is oftentimes prohibitive in terms of memory and CPU time requirements. For this reason, most annotation and identification software tools work on centroided spectra. Centroided mode retains only the most relevant features of a peak (*m/z* and maximum intensity or peak area) to dramatically reduce the size of the dataset, which leads to relaxed memory and CPU time requirements.

4.2.2. Library-centric vs. feature-centric strategies

There are two general approaches to determine chemical composition in MSI: library-centric or feature-centric. These approaches are applicable to both annotation (using only exact mass matching) and identification (combining MSI with orthogonal techniques and reference standards).

Library-centric approaches match library information to experimental data. For each candidate compound in the library, the algorithm will generate an *in-silico* theoretical

spectrum (with isotopes, adducts, or ion fragments) using the molecular formula, and will determine its presence in the sample by matching them against the experimental spectra (usually the mean spectra) (Alexandrov and Bartels, 2013; Novák, Škríba and Havlíček, 2020; Tortorella *et al.*, 2020). These approaches tend to be computationally consuming in terms of time and memory, as the algorithm will try to fit all the compounds in the libraries. Besides, the results are limited to the compounds existing in the libraries (if the compound does not exist in the library, the associated *m/z* signals will not be annotated).

Feature-centric approaches look for patterns in the data (adducts, isotopes, or fragments) to create several networks of related MS signals. In general, this strategy gathers information from the data and tries to construct isotopic patterns of unknown compounds taking into account the spatial correlation, the intensity profile, and the mass error between features (Bond *et al.*, 2017; Janda *et al.*, 2021; Sementé *et al.*, 2021). This approach also includes using the Kendrick mass defect (KMD) to assign families of compounds (Kune *et al.*, 2019). At the end of these procedures, some *m/z* features are confidently annotated as monoisotopic ion candidates, considering all the information gathered, and can be searched against libraries of compounds. These approaches tend to be faster to run but require extra steps to assign compounds to the *m/z* features. Additionally, they are less generalizable, as they make certain assumptions about the data that might be specific only to a certain family of compounds, like the shape of the isotopic pattern due to the elemental composition; or about the experimental procedure, like searching for specific adducts or labeled moieties.

4.2.3. Isotopic pattern generation

Tools that follow the library-centric approach tend to generate the *in silico* pattern of the compounds in the libraries to compare with the spectra. This can be achieved using in-house algorithms or with enviPat (Loos *et al.*, 2015), an R-package that generates the profile spectrum and the centroids of sum formulas simulating different resolving power; and Rdispo (Böcker *et al.*, 2006), an R-package that generates isotopic patterns and elucidates molecular formulas for a given mass.

4.2.4. Match scores

Regardless of the approach followed (library-centric or feature-centric) all software tools rely on several match scores to determine the fitness of each hit. The two main metrics are (1) spectral similarity (to compare experimental data against theoretical isotopic ratios, fragmentation spectra, or CCS) and (2) spatial similarity (to determine if isotopes, adducts, and fragments are colocalized, using the ion images of the tissues).

The most widely used spectral similarity metrics are Pearson's correlation (McDonnell *et al.*, 2008), and cosine similarity. *Smets et al*. proposed histogram matching as an alternative (Smets *et al.*, 2019). Recently, a new metric inspired by natural language processing algorithms (Spec2Vec) (Huber *et al.*, 2021) has been shown to outperform cosine similarity.

Spatial similarity can be determined using Pearson's/Spearman's correlation, cosine similarity, hypergeometric similarity measure (Kaddi, Parry and Wang, 2011) or Structural Similarity Index (SSIM) (Ekelöf *et al.*, 2018). *Ovchinnikova et al*. (Ovchinnikova, Stuart, *et al.*, 2020) used as a gold standard 2210 ion images ranked by similarity by 42 imaging experts to quantitatively compare several spatial similarity metrics. One of the machine learning models (Pi-Model) included in their software

ColocML obtained the highest performance (0.797 correlation to the gold standard) closely followed by cosine similarity (0.794) and Pearson's correlation (0.788).

The match score can be further refined using other metrics such as mass error (Sementé *et al.*, 2021), spatial chaos (Palmer *et al.*, 2016; Tortorella *et al.*, 2020), or False Discovery Rate (FDR) estimates (Palmer *et al.*, 2016).

4.2.5. Library matching

Both in annotation and identification, it is crucial to compare the MS signals obtained in the experiment to a list of known compounds or references. To obtain the highest degree of confidence in annotation, the experimental data must be matched against a reference standard. Nevertheless, reference standards are not always available or compatible with the experimental workflow of choice. Reference standard matching is particularly challenging in untargeted studies, where tens or even hundreds of compounds are analyzed at the same time. To aid compound annotation in these cases, several libraries compile and index thousands of previous experimental MS and MS/MS measurements of standards from laboratories around the world. Libraries offer a reliable, automatable, and easy-to-use substitute to real standards. They can be considered compound-centric or spectra-centric, depending on their content.

Compound-centric (or metadata-centric) libraries such as HMDB (Wishart *et al.*, 2018), ChEBI (Hastings *et al.*, 2016), PubChem (Kim *et al.*, 2019) include information such as the monoisotopic mass of the compound, molecular formula, SMILES, Inchy Key, molecular structure, and in some cases, other relevant metadata such as compound origin (plant, animal, bacterial, etc.) or even metabolic function. This first type of library is mainly useful for exact mass matching.

Spectra-centric libraries store MS and MS/MS spectra of thousands of compounds. Identification is obtained by matching experimental data to the spectra in the library. Several libraries are available for MS/MS, some examples include METLIN (Smith *et al.*, 2005), NIST (Lemmon *et al.*, 2010) and MassBank (Horai *et al.*, 2010). For ion-mobility, the most ambitious projects include the Online Collision Cross Section Compendium (Picache *et al.*, 2019) and the AllCCS atlas (Zhou *et al.*, 2020). Most databases in this category have been developed with traditional MS technologies in mind (mainly LC-MS and GC-MS) and almost exclusively include fragments of the $[M+H]^+$ and $[M-H]^-$ adducts. There is a lack of databases of experimental spectra acquired by MSI.

With the advent of machine learning and cheminformatics techniques, *in-silico* libraries have emerged. They typically generalize from experimental data of pure compounds and rely on advanced algorithms to generate relevant information of unknown or unmeasured compounds. This can include information such as monoisotopic mass, molecular formula, chemical structure and even MS and MS/MS spectra. Some of these in-silico tools for tandem MS include Sirius (Dührkop *et al.*, 2019), MetFrag (Ruttkies *et al.*, 2016) and CFM-ID (Djoumbou-Feunang *et al.*, no date). For ion mobility, AllCCS (Zhou *et al.*, 2020) uses machine learning to predict CCS values from SMILES. These tools should be carefully evaluated and used in a case-by-case scenario. Blindly trusting them in untargeted studies can lead to incorrect annotations.

4.2.6 Peak Filtering

Peak annotation results can be used to filter out redundant or non-biologically-relevant peaks from downstream statistical analyses.

A common peak filtering strategy is deisotoping (Bond *et al.*, 2017; Sementé *et al.*, 2021), which consists of localizing monoisotopic peaks in the spectra to remove all the subsequent isotopic peaks. This eliminates redundancy in the data, as all the isotopic peaks in a pattern are highly correlated and facilitates the posterior identification of the monoisotopic peaks. In this same line, another peak filtering strategy is de-adducting, which consists in discovering as many adducts as possible for each compound to combine them as a unique feature. The identification of adducts is mainly based on the mass difference between them; which could lead to the detection of false adducts since there may exist multiple mass differences between ions that match with several possible adducts. Additionally, the images produced by adducts are not necessarily co-localized among them (like in the case of isotopes) because the natural abundance of the adduct-forming elements over the tissue sample (Hankin *et al.*, 2011) (i.e. Na^+ or K^+ ions) may be not homogeneous and dependent on the tissue type.

Finally, spectra contain exogenous peaks, (coming from the substrate, the matrix, the embedding medium, etc.) that may be desirable to exclude from the analysis. In the ideal case, these peaks should be annotated and discarded, although sometimes they could be useful for calibration purposes, like some inorganic matrix peaks(Ràfols, Vilalta, Torres, *et al.*, 2018). Most of the strategies behind annotating these off-sample ion peaks are based on exact mass matches by knowing which compounds are expected to appear in the sample preparation (Niedermeyer and Strohalm, 2012; Baquer *et al.*, 2020), but there are also software programs that rely on machine learning methods to annotate them (Ovchinnikova, Kovalev, *et al.*, 2020).

4.2.7. Data sharing and repositories

Data repositories are an essential tool for data sharing. The vast amount of experimental data available allow two main benefits to the community: experimental results are easily accessible to the whole community, and the data can be used to validate and develop software tools.

METASPACE (Alexandrov *et al.*, 2019) is the main repository available in MSI. To date, METASPACE holds over 6000 experimental studies. Both the experimental data (in .imzML (Schramm *et al.*, 2012) format and centroid mode) and resulting annotations using PySM (Palmer *et al.*, 2016) can be freely downloaded.

More generic repositories include Metabolights (Haug *et al.*, 2013) or Metabolomics Workbench (Sud *et al.*, 2016). Nevertheless, their coverage of MSI experiments is rather limited. Only 50 (0.2% of all entries) and 4 (0.25% of all entries) of their respective entries correspond to MSI experiments.

4.3. Specific software packages

The MSI community has dedicated their efforts to developing several software tools for the compound annotation/identification of MSI data. In this section, we review 12 current software tools to guide the readers in selecting the most suitable ones for their application. Table 2 contains a summary of the main characteristics of each tool including the confidence levels of the annotations/identifications they can provide, the target features, the output, and the general type of annotation. We have defined three types of annotation: (1) “general annotation” if all the peaks in the spectra are targeted; (2) “specific annotation” if specific peaks (e.g. matrix) are annotated; and (3) “identification” if MSI is combined with MS/MS or other orthogonal techniques.

4.3.1. Alex[123]

Alex[123] (Ellis *et al.*, 2018) is a software for the automated identification of lipids. It relies on a unique experimental setup multiplexing an FTMS Orbitrap for high-mass resolution MSI and an IT-MS/MS for data-dependent acquisition (DDA) on-tissue fragmentation of almost every detected *m/z* value. By alternating the two acquisitions in 20µm steps, they can effectively determine high-mass MSI and structural information *in-situ*. This tool achieves Level 3 and 2 identification confidence.

They rely on an in-house library that contains more than 430k molecular lipid species and their adduct-specific fragments. They use different adducts based on the lipid family. To annotate a sum-composition lipid species from the FTMS data, the peak must be present in all 3 replicates and at least 1 fragment must be detected by IT-MS/MS. To identify the lipid species, three conditions must be met: (1) at least 50% of the fragments must be detected, (2) two complementary pairs of fragments (adding to the parental ion) must be detected, and (3) the parental ion must be found by FTMS.

Using the MS data, they managed to annotate 165 unique sum-composition lipid species in rat brain tissue. From these sum-compositions, they managed to structurally identify 113 lipid species using the parallel IT-MS/MS run. A total of 92% of the identified lipids could be validated with HPLC-MS/MS.

4.3.2. CycloBranch 2

CycloBranch 2 (Novák, Škríba and Havlíček, 2020) is a standalone software package implemented in C++ that can annotate LC-MS, MSI, and MS/MS data independently or combine all of them. CycloBranch 2 generates molecular formulas from an input list of chemical elements to form a database of compounds, optimized for peptides and some small molecules. Later, all the molecules in the database are tested using various rules like the nitrogen to oxygen ratio, the Senior's rules(Kind and Fiehn, 2007) and matching the *m/z* in an experimental input spectrum. Additionally, CycloBranch 2 supports fine isotope structure annotation, being able to resolve $^{34}S/^{13}C_2$ and $^{41}K/^{13}C_2$ peaks. Moreover, CycloBranch 2 includes a tool to visualize the annotations over the MSI image combined with multiple microscopy or histology images, which can be shifted and adjusted manually to increase the overlap between them. The output of the software consists of a list of interactive tables that show the annotations over the spectra and images.

The tool was used to annotate an MSI dataset consisting of a mixture of three commercial siderophores standards of bis(methylthio)gliotoxin, ferrioxamine, and triacetylfusarinine C ferriform over an ITO glass. Cyclobranch 2 predicted elemental compositions of all three compounds reported in at least 50 spectra from a total of 1215. Later, the peaks were searched in a library of 709 siderophores and secondary metabolites as a positive control.

4.3.3. HIT-MAP

HIT-MAP (Guo *et al.*, 2021) is an R package that annotates peptides and proteins in high mass resolution MSI datasets using peptide mass fingerprint analysis and a scoring system. To annotate, HIT-MAP generates a customized local database of digested proteolytic peptides *in silico* from a protein sequence file in FASTA format containing the proteome of the species under investigation and a complete *in silico* digestion

framework. Moreover, HIT-MAP generates a decoy database to produce FDR-controlled annotations.

To match the reference database with the experimental data, three principal scores are used. First, the number of peaks in the experimental isotopic pattern found in the theoretical pattern, discarding those peaks below 2.5% of the most intense isotopic peak; second, the intensity profile of the patterns; and third the mass error between peaks. Once a list of annotated peptides is generated, protein annotation is achieved by grouping peptides into the target proteins computing an FDR. The output of HIT-MAP consists of two sub-folders, one containing all the identification data and the other a summary with peptide and protein lists as well as ion images.

4.3.4. LipostarMSI

LipostarMSI (Tortorella *et al.*, 2020) is a commercial software for targeted and untargeted MSI data analysis with automated annotation of lipids, metabolites, and drug metabolites. It annotates by accurate *m/z* ratio matching within user-defined tolerances in libraries of compounds like the HMDB or LIPID MAPS. In-house libraries are also supported. Each hit to the database is ranked based on a mass score (proximity to the theoretical mass), an isotopic pattern score (compliance to theoretical intensity ratios and mass distances), and chaos score (spatial distribution of the *m/z* density image).

The software also allows the inclusion of MS/MS data to reach higher levels of confidence in identification. Each experimental MS/MS spectra can be compared to fragmentation libraries or to in-silico fragments produced by a set of proprietary lipid fragmentation rules. In addition to the scores used in annotation, a new fragment score is introduced. This score is based on (1) the percentage of theoretical fragments found in the experimental data, and (2) the ratio between experimental and theoretical fragment intensities. Each theoretical fragment can be labeled as “mandatory” or “recommended” either manually or based on user-defined intensity thresholds. This allows the fragment score to only focus on relevant fragments.

The output of the software consists of a list of compounds assigned to each *m/z* ratio and ranked by the LipostarMSI score. Each annotation/identification is color-coded based on the confidence of annotation. Green indicates successful structural identification; orange indicates the presence of conflicts that need to be manually reviewed and approved; and red indicates unsuccessful identification. Figure 3 shows the general workflow of annotation and identification software tools in MSI. Each of the steps increases the level of confidence and relies on different experimental data and libraries.

Finally, after approving correct identifications, all adducts assigned to the same compound are merged in a unique identification.

4.3.5. Mass2adduct

Mass2adduct (Janda *et al.*, 2021) is an R tool that follows a feature-centric approach to automatically annotate common alkali metal adducts, matrix adducts, and isotopes. The tool computes the mass difference between all *m/z* feature pairs available in the dataset and plots them in a histogram. The most common mass differences are matched against a list of common adducts to determine their identity. Finally, the Pearson’s correlation of each candidate adduct to their parental ion is used to discard unlikely adducts. Bonferroni

correction and false-discovery rate analysis based on q-value cutoff are applied to Pearson's correlation values.

To validate their approach, they conducted on-tissue tandem MS on mouse brain tissue using DHB as the matrix. They focused on four pairs of *m/z* values with a mass difference of 136.016 Da (DHB-H_2O) and found that they showed identical MS/MS fragments.

They showcase their annotation tool on several tissue types, sources, mass analyzers and two matrices (DHB and CHCA). Comparable $[M+Na]^+$ and $[M+K]^+$ adduct frequencies were found across tissue types and experimental setups. Abundant matrix peaks were found for DHB (up to 30% of the total amount of features). CHCA was less abundant (up to 10% of all features).

As a final validation, they compare their results to METASPACE (Alexandrov *et al.*, 2019). Out of the 604 *m/z* features annotated as matrix adducts by Mass2adduct for a mussel dataset, a total of 103 were annotated as metabolites by METASPACE. This highlights that matrix adducts can cause false-positive annotations and they should be considered for library searches. They also conclude that exact mass matching is not enough for identification and the use of orthogonal techniques is required.

4.3.6. massPix

massPix (Bond *et al.*, 2017) is an R package that combines data analysis functionalities with deisotoping and exact mass matching against generated lipid libraries. The deisotoping algorithm finds monoisotopic ions (M+0) and removes the first and second isotopes (M+1 and M+2) which are within a calculated proportion of M+0. To achieve lipid annotation, first, a library of lipids is generated by combining common fatty acids, lipid head-groups and adducts; and second, the M+0 ions previously found are matched against the generated library. The output consists of various CSV files with annotations.

4.3.7. MSKendrickFilter

MSKendrickFilter (Kune *et al.*, 2019) is a python software capable of exploiting the benefits of KMD analysis to classify chemically related compounds in their corresponding families. It is based on the conversion of exact mass measurements to a Kendrick Mass (KM) scale (linear conversion factor computed with the nominal and exact mass of a reference molecule of choice). The KMD is later obtained by keeping the decimal part of the KM. Their results show how using CH_2 as a reference molecule, different tetraalkylammonium, lipids, and lipopeptides families can be identified. When using C_2H_4O as a reference molecule, different polymers groups could be separated. Their results were validated on bacteria cocultures and brain tissue sections.

4.3.8. OffsampleAI

OffsampleAI (Ovchinnikova, Kovalev, *et al.*, 2020) is an artificial intelligence approach to recognize ion images localized outside of the sample (off-sample). The authors initially compiled a database of 23,238 ion images from 87 public MSI datasets manually labeled as on-sample and off-sample by 5 experts (using a custom web app). This database is used as a validation for the three algorithms proposed. The two first methods proposed, the “Spatio-molecular biclustering method” and the “Molecular co-localization method” rely on the spatial correlation between ions and clustering of pixels to identify off-sample ions. The top-performing method is based on a Deep residual Learning approach trained on part of the gold standard.

4.3.9. pySM (METASPACE)

Palmer et al. (Palmer *et al.*, 2016) proposed a novel approach to annotate metabolite data in MSI with a confidence estimation approach. Using the compound-specific databases selected by the user, as well as a list of possible adducts, a list of all possible monoisotopic molecular matches is compiled. These molecular matches are then ranked based on the so-called metabolite-signal match score (MSM score), a composite score that relies on three metrics: (1) the "spatial chaos metric" quantifies the informativeness of the monoisotopic peak (2) the "spectral isotope metric" indicates the degree of similarity between the theoretical isotopic pattern and the experimental one and (3) the "spatial isotope metric" indicates the degree of similarity between the ionic images for all isotopes.

The MSM score values will depend largely on the sample at hand, making it difficult to specify a stable MSM cutoff. This is addressed using an FDR value estimation using a Target-Decoy approach. The main database with normal adducts is referred to as the Target database and it is extended with a Decoy database of the same size. In this case, the decoy is composed by randomly selecting an implausible adduct. For each search in the Target database (using plausible adducts) a search in the Decoy database is conducted (using implausible adducts). All hits, from both the target and the decoy databases, are ranked based on MSM. The number of Decoy hits and Target hits above a certain MSM cutoff is used to estimate the FDR. This allows converting an MSM cutoff to a much more easily interpretable FDR cutoff.

pySM is currently integrated in the online annotation platform METASPACE (Alexandrov *et al.*, 2019), which allows users to submit high-resolution datasets to be annotated using four libraries: CoreMetabolome (an in-house library), HMDB(Wishart *et al.*, 2018), LipidMaps(Sud *et al.*, 2007) and SwissLipids(Aimo *et al.*, 2015). Moreover, METASPACE allows sharing the results online by storing all data online, both the MSI data and the annotations, and includes options for privacy and teamwork. METASPACE contains nowadays close to 6000 downloadable MSI datasets, being one of the biggest MSI data repositories in the world.

4.3.10. ReSCORE METASPACE

Some strategies try to extract more information from the annotations and identifications rather than only speculating with the identity of peaks for MSI datasets. One of them is annotation rescoring, which implies a verification step after the initial annotation to increase the precision of the workflow. In this line, *Silva et al.* (C Silva *et al.*, 2018) applied this strategy with METASPACE (Alexandrov *et al.*, 2019) to increase the FDR of the target-decoy approach. The strategy consists of various recursive iterations of selecting some of the annotations with higher scores from the target set and some annotations from the decoy set to train a linear classifier using a collection of 34 features extracted for each annotation. At each iteration, the annotations are rescored using the linear classifier until a certain number of iterations is reached. The result of this procedure increases the number of annotated compounds for a given FDR in METASPACE.

4.3.11. rMSIannotation

rMSIannotation (Sementé *et al.*, 2021) is an annotation workflow integrated into the MSI processing R package rMSIproc (Ràfols *et al.*, 2020) and implemented in C++. The algorithm annotates monoisotopic ions from metabolites and peptides by directly searching in the spectra peaks that accomplish three rules: spatial correlation, isotopic

mass distance, and intensity profile of the isotopic pattern, which can be extracted with confidence thanks to the great number of sampling points in an MSI experiment. To avoid direct searches in libraries, rMSIannotation uses a previous modelization of the intensity profiles of different compounds found in the HMDB (Wishart *et al.*, 2018) and the Peptide Atlas (Desiere, 2006), which allow to predict variations in the intensity profile along the *m/z* axis. After detecting monoisotopic peaks, the algorithm groups them creating networks of adducts using spatial correlation as a criterion. The output of the algorithm consists of different R structures containing all the annotations in tables, information about the isotopic patterns, and the adduct networks. Moreover, it retrieves structures to facilitate the inclusion or exclusion of monoisotopic and isotopic peaks from the data analysis and there are visualization options included in rMSIproc.

4.3.12. rMSIcleanup

rMSIcleanup (Baquer *et al.*, 2020) is an R package that annotates matrix-related signals in MSI datasets. It annotates them by computing all the theoretical isotopic patterns related to the matrix clusters and matching them to the spectra using cluster spectral similarity and intra-cluster morphological similarity. Moreover, it detects overlapped peaks in the isotopic pattern using bisecting k-means based on the correlation of their spatial distribution. The output of rMSIcleanup consists of an informative visual report in PDF with all the patterns detected, ion images, and cluster names.

5. Extending the imZML format to include annotations and identifications

The imzML is a data format (Schramm *et al.*, 2012) created to enable the exchange of MSI data between different software and instruments. It uses two files linked by a universally unique identifier (UUID): (1) an XML file that stores experimental metadata that expands on the HUPO-PSI mzML standard format, and (2) a binary file to store spectral data efficiently. The spectral data can be stored in continuous mode, where all pixel MS measurements share the same *m/z* values, or in processed mode, where each pixel has its *m/z* values.

The imzML format is currently the gold standard for MSI data storage and sharing. Nevertheless, it does not contemplate a standard way of including molecular annotations and identifications.

The MS community has recognized the importance of storing annotations and identifications in a reproducible manner to stimulate data sharing and accountability. This interest promoted the creation of several file formats that complement the popular mzML file format (Martens *et al.*, 2011), a standard format developed by HUPO Proteomics Standards Initiative (Hermjakob, 2006) to “capture the use of a mass spectrometer, the data generated, and the initial processing of that data (to the level of the peak list)”. Although these file formats are not compatible with MSI experiments, the current and most relevant formats to store annotations and identifications in MS are mzTab, mzTab-M, and mzIdentML.

mzTab (Griss *et al.*, 2014) was first released in 2014 and it is intended to store only the final reported results of an MS proteomics experiment and to provide a simple way to share data with MS proteomics repositories. It can contain protein, peptide, and small molecule identifications with basic quantitative information. Using the same core as

mzTab, a new format to better support small molecule experiments was developed by the end of 2019 as the 20th version of mzTab, the mzTab-M (Hoffmann *et al.*, 2019). This file format is intended to extend the concept of mzTab to include more details for quantification, including different charge states or adducts, and was developed specifically for experiments on small molecules like metabolites and lipids. In the future, mzTab-M might be adopted to create a specific version of mzTab for proteomics only (mzTab-P (Salek, 2019)), but now, mzTab version 1.0 remains active for proteomics. Both standard file formats are structured as tab-delimited text files and are intended to share part of the results of an experiment (not all the MS data), which make them suitable for searches in libraries and to be the output of library searches. The files are structured as big tables of compound identifications with fields like database identifier, chemical formula, theoretical neutral mass, adduct ions, and various study variables that can be defined by the user. A heading containing metadata and some defining words are also included.

mzIdentML (Jones *et al.*, 2012) is an XML-based format that was first released in August 2009 and reached the current version 1.2 in March 2017. It is intended for the systematic description of polypeptide identification and characterization based upon MS. The format was originally named AnalysisXML to encapsulate different computational analyses on proteomics performed with mass spectra, but it was decided to split the development into two branches: mzIdentML for peptide and protein identification, and mzQuantML (Walzer *et al.*, 2013), to describe quantification experiments. mzIdentML can store MS data by itself, but it is expected to be accompanied by an mzML file (there is an mzML unique identifier camp inside mzIdentML) containing the complete dataset, as mzIdentML is best suited for results and not the complete experiment. Polypeptide's identifications can be stored in different ways depending on the identification procedure, but the information usually consists of the sequence accession, the length of the sequence, information about the enzyme used, and fragmentation information among many others.

6. Perspectives

6.1. Identification confidence levels for MSI

As MSI matures into an analytical technique frequently used in untargeted metabolomic studies, the scientific community expects the same level of accuracy and accountability in MSI experiments as in studies with LC or GC coupled to MS or NMR. Thus, we propose the adoption of the identification confidence levels used in LC-MS metabolomics [19] to the field of MSI as described in Section II.B. and Figure 1.

MSI lacks the chromatographic separation available in LC and GC metabolomics, which impedes the acquisition of orthogonal information (i.e. RT). Nevertheless, the high number of pixels enables image and peak intensity correlations to reliably annotate isotopes, adducts, and in-source fragments. A typical MSI experiment contains tens of thousands of pixels that are usually treated as individual data points during statistical analyses. It is important to use statistical models that compensate for spatial autocorrelation (pixels are not independent of each other and depend on their neighbors) (Cassese *et al.*, 2016).

We are confident that the MSI community, especially in the field of software development for annotation and identification, would benefit from this proposal. Firstly, we encourage

the community to be consistent with the term *annotation* and *identification*. As shown in Supplementary Table 1, more than 50% of the papers reviewed used the term identification to refer to exact mass matching. Assignments based on only exact mass matching (Level 4-5) should be referred to as annotation. “Annotation” should still be used even when using orthogonal information to distinguish isomers and isobars (Level 2-3). The term “identification” should be used when all experimental data is matched against a reference standard (Level 1).

Secondly, we claim that users of software tools would appreciate a clear indication of the level of confidence the tool provides. The list of annotations and identifications produced by the software should include a field indicating the level of confidence (Level 1-5). Furthermore, we consider that they should also be clearly specified in any accompanying publication.

The adoption of these guidelines will provide a clear framework to communicate confidence in annotation and identification and ensure correct biological interpretation of the results. This initiative will also encourage the community to strive for higher identification confidence in their studies by adjusting their experimental and software workflows.

6.2. Incorporation of annotations and identifications to the imzML format

Table 1 shows that imZML (Schramm *et al.*, 2012) is the default input format in the overwhelming majority of software tools for annotation and identification in MSI. This clearly indicates the full commitment of the community to the idea of cross-instrument, open protocol, and standardized data sharing. The imZML format has been a clear success. At the same time, Table 1 also shows a clear disparity of output formats (.csv, .xlsx, .Rdata, .pdf, …). The resulting annotations and identifications are usually reported in loosely-defined in-house formats with different fields that impede data sharing, integration, and reusability.

Thus, we identified an imperative need for a standard format to report MSI annotations and identifications easily integrable with imZML.

We have observed that most data formats for MS that contain identifications (mzIdentML (Jones *et al.*, 2012), mzTab (Griss *et al.*, 2014), and mzTab-M (Hoffmann *et al.*, 2019)) were not designed to contain all the spectral data but as an annex to the mzML (Martens *et al.*, 2011) data storing format. Additionally, these data formats answer the needs of specific research fields, like proteomics and metabolomics. There is no universal data format to report MSI annotations and identifications.

We propose adopting this same strategy to define a new file format to include annotations and identifications as an annex to the imzML standard. We consider that in the field of metabolomics the format mzTab-M should be used as a reference. Each dataset would now be described by three key files: the common .ibd and .imzML files containing the spectral data and a new .mzTab-M file containing annotations, identifications, and supporting evidence. All these files would be linked using the same Universally Unique Identifier (UUID). The .mzTab-M file could contain multiple UUIDs in

studies with multiple .imzML files. Figure 4 shows a high-level abstraction of the imzML format, the mzTab-M format, and their integration by a list of UUIDs.

mzTab-M is the result of years of collaborative work between the Metabolomics Standards Initiative, Proteomics Standards Initiative, and Metabolomics Society. It relies on a well-defined structure and controlled vocabulary, and it can be read, written, and validated using mzTab-M (Hoffmann, Hartler and Ahrends, 2019). It has successfully been adopted by some of the main MS annotation software such as Lipid Data Analyzer (Hartler *et al.*, 2011), GNPS (Nothias *et al.*, 2020), MS-Dial (Tsugawa *et al.*, 2015), and MetaboAnalyst (Chong *et al.*, 2018).

The main drawback of this format is the fact that it is based on a plain text tab-separated file and relies on a custom structure defined by its own specification. This format makes it easy to read and understand visually as it is in plain text. Additionally, it is arranged as a set of tables, making it an easy replacement of excel and CSV files commonly used in publications and statistical programming languages like R. Nevertheless, using Extensible Markup Language (XML), a ubiquitous file format in all fields of computer science, would be more desirable. All major programming languages and platforms have plenty of reliable tools to read, write and validate XML. And its well-defined structure makes it extend. In the long run, basing it on XML ensures a robust adoption by more developers and easier maintenance. We consider that one of the priorities when adopting mzTab-M for MSI applications is to redefine it in XML format.

Additionally, to adapt it to the field of MSI, part of the controlled vocabulary and fields defined by the mzTab-M format would need to be updated or removed. New fields would also need to be defined. As an example, all columns regarding RT in the Small Molecule Feature table (SMF) should be removed. The general structure of metadata, small molecule table (SML), Small Molecule Feature table (SMF), and Small Molecule Evidence table would remain unchanged.

Finally, the most crucial point to consider is how to include the spatial information of the identified compounds. The same MS signal can correspond to different molecules in different areas of the tissue, especially when working with low-resolution MS analyzers, like peptides with the same *m/z* belonging to different proteins (Guo *et al.*, 2021). Accounting for this phenomenon is a non-trivial task. We suggest including a column to specify the ROI of a specific MS feature. The representation and storage of ROIs are not properly solved in MSI, and multiple vendors and software tools use their own custom-built formats.

6.3. The future of automatic annotation and identification in MSI

We have extensively reviewed twelve software tools available between 2016 and 2022 to perform automatic identification and annotation of MSI data. Tools specialize in different target molecules (i.e. metabolites, lipids, peptides, or proteins), different experimental data (i.e. MS, tandem MS, ion mobility or integrate other orthogonal techniques), and different approaches (i.e. library-centric or feature centric). Most of the tools available to date only focus on annotation and only reach identification level 4 as they rely on exact mass matching. ALEX[123] (Ellis *et al.*, 2018), CycloBranch 2 (Novák, Škríba and Havlíček, 2020) and Lipostar (Tortorella *et al.*, 2020) are the only tools that can consistently provide Level 3 or Level 2 identifications. There is a clear need for

automatic tools that can provide identifications with a confidence level over 3. Combining structural information obtained from orthogonal techniques is an important area of research that needs to be further explored.

For a confident identification, it is important to highlight the importance of proper mass calibration (Ràfols, Vilalta, Brezmes, *et al.*, 2018), using internal standard compounds or matrix peaks, and the use of high-resolution mass analyzers with mass accuracy below 5 ppm.

The future of automated annotation and identification in MSI relies not only on instrumental development but also on creativity in the application of strategies inspired by more established MS-based techniques such as LC-MS and GC-MS. We have identified the following challenges where software developers have an opportunity to make an impact in the field of annotation and identification by MSI:

- *In-source fragmentation*

To date, there is no automatic tool that directly addresses the annotation of in-source fragments (fragments generated naturally during ionization or desorption) in MSI. Their correct annotation is key, as in-source fragments clutter the spectra and can be wrongfully annotated as other parental ions ((Garate *et al.*, 2020)). This is particularly problematic in ion sources like SIMS and LA-ICP, but it is still a problem in soft-ionization sources like DESI or MALDI. At the same time, if properly dealt with, in-source fragments promise to increase confidence in annotation as they can provide insights into the structure of a molecule (much like tandem MS). In a recent LC-MS study, *Xue et al.* (Xue *et al.*, 2020) proposed adjusting the ESI source to produce in-source fragmentation patterns comparable to the MS/MS spectra available in METLIN (Smith *et al.*, 2005). They found that 90% of 50 mixed metabolites showed in-source fragmentation patterns consistent with METLIN. This could lead to potentially high levels of confidence (above level 3) only using MS1 data.

- Exogenous compounds

Similarly, although several efforts have been presented in recent years (Baquer *et al.*, 2020; Ovchinnikova, Kovalev, *et al.*, 2020; Janda *et al.*, 2021), a comprehensive and reliable tool for the annotation of matrix-related signals of all widely used matrices is still missing. The use of inorganic matrices limits the presence of matrix fragments in the low range of the spectrum, but its use is far from being widespread.

Another area needing further research is the annotation of exogenous compounds. Various MSI workflows contemplate the use of FFPE slides as sampling material but identifying all the peaks that originated during the sample processing is still an open issue. Here we see an opportunity for researchers to develop software tools dealing with the identification and removal of all the peaks related to FFPE, OCT, or other cutting materials, which would require an in-depth analysis of the chemical processes produced by the sample processing.

- *Stable Isotope Labeling (SIL) annotation*

Following this line, SIL methods for MSI would benefit from the development of annotation tools specially designed for targeting different compounds with distinct or artificial isotopic patterns. There are various annotation tools for LC-MS data that are able to target SIL compounds (Neumann *et al.*, 2014; Capellades *et al.*, 2016; ‘Evaluation

of freely available software tools for untargeted quantification of 13C isotopic enrichment in cellular metabolome from HR-LC/MS data', 2020). These tools could be used to inspire the development of new software for MSI. Even better, contributing to the development of this software to include MSI data would allow combining both LC-MS and MSI SIL methods, which would benefit both disciplines and open the door for more collaboration between techniques in the SIL field.

- *Pathways in LC/GC-MS how to apply them in MSI*

Metabolic pathway analysis (a.k.a. metabolic pathway enrichment analysis) compares two sample classes (i.e. control vs. treatment or condition vs wildtype) to produce a list of dysregulated (upregulated or downregulated) metabolic pathways. Data about metabolic pathways is obtained from databases such as KEGG (Kanehisa and Goto, 2000), HMDB (Wishart *et al.*, 2018), or BioCyc (Caspi *et al.*, 2014). For each pathway found, the coverage percentage is given (the percentage of metabolites in the pathway annotated). For each feature annotation, the dysregulation (up or down), fold-change and p-value are given. Additionally, an overview of all pathways can be represented in a variety of plots showing overall significance (p-value) or metabolite overlap percentage. This process is typically performed on the list of annotations but using the mummichog algorithm it can be applied directly to MS features. XCMS (Forsberg *et al.*, 2018) and MetaboAnalyst (Chong *et al.*, 2018), two major MS metabolomics processing platforms, implement pathway analysis.

Additionally, to facilitate the generation of hypotheses, several software tools also include interactive network explorers. Metaboanalyst (Chong *et al.*, 2018), for example, allows the user to show the metabolite annotations on the KEGG (Kanehisa and Goto, 2000) global metabolic network and other networks.

To date, there is no automatic tool that can provide pathway analysis in MSI. Currently, pathway analysis in MSI is typically done by (1) running annotation/identification, (2) exporting a list of significant metabolites when comparing two ROIs, and (3) conducting pathway analysis using non-MSI targeted tools data such as XCMS or MetaboAnalyst. As an example, *Sun et al.* (Sun *et al.*, 2018) followed this approach (using KEGG and MetaboAnalyst) to metabolically compare the cortex and medulla in human adult adrenal gland samples. Among other pathways, the purine metabolism pathway was upregulated in the medulla while the biosynthesis of unsaturated fatty acids was upregulated in the cortex.

- *The role of AI and DL in annotation*

Finally, we conclude the review addressing the hot topic on every researcher's lip: Deep Learning (DL). DL has already achieved science-fiction-like results in a wide range of fields such as robotics (Sünderhauf *et al.*, 2018), natural language processing (Otter, Medina and Kalita, 2021), and medical image processing (Minaee *et al.*, 2021). In recent years, MSI has seen some developments in Machine Learning (ML) and DL in applications such as tumor classification (Behrmann *et al.*, 2018), clustering (Zhang *et al.*, 2021), image registration (Race *et al.*, 2021), and peak picking (Abdelmoula *et al.*, 2021). In the field of molecular annotation and identification, OffSample AI (Ovchinnikova, Kovalev, *et al.*, 2020) used several DL models for the annotation of matrix-related and off-sample MS features. Nevertheless, the adoption of these technologies for MSI metabolomics is slow and we seem to be missing out on this Artificial Intelligence revolution ('Why the metabolism field risks missing out on the AI

revolution', 2019). The two main drawbacks that are holding the community back are (1) the lack of result transparency and accountability, and (2) the lack of big data for training.

MSI is used in fields such as biochemistry, pharmaceutics, and medical diagnostics where reliable annotations and identifications are crucial. Since their inception, ML and DL have struggled with their inability to transparently justify their learning-based non-linear results (black-box problem) (Castelvecchi, 2016). This inherent problem leaves scientists and funding bodies unable to fully interpret and trust DL results (von Eschenbach, 2021). There are three strategies to open the black box (Azodi, Tang and Shiu, 2020). In the field of MSI molecular annotation and identification, the black-box problem could be mitigated by coupling DL models with more traditional score-based methods (i.e. spectral similarity, spatial similarity, spectral chaos, FDR estimates, etc.). Only annotations and identifications ranking high in both approaches would be accepted automatically, while mismatching annotations and identifications would be manually curated by the user.

The second bottleneck limiting the adoption of DL is the lack of big, labeled, and curated sets of MSI data ("ground truth") needed to train the models (Alexandrov, 2020). Ideally, for training DL models, in the task of annotation and identification, we would need thousands of MSI datasets with a complete list of Level 1 identifications. Additionally, for the DL model to generalize, it should be exposed to enough sample types (specimen, condition, tissue) and instrumental setups (ion source, ion mode, and mass analyzer). METASPACE (Ovchinnikova, Kovalev, *et al.*, 2020) includes thousands of publicly available datasets, but it does not include a complete list of confident annotations. The creation of this ground truth could follow two approaches (Alexandrov, 2020). The first approach relies on expert crowdsourcing to manually annotate MSI datasets and has successfully been used in MSI to estimate quality (Palmer *et al.*, 2015), off-sample signals (Ovchinnikova, Kovalev, *et al.*, 2020), and colocalization (Ovchinnikova, Stuart, *et al.*, 2020). Nevertheless, expert annotation could prove unfeasible and unreliable in the task of molecular annotation and identification. Following the success of MS/MS libraries like METLIN (Smith *et al.*, 2005), NIST (Lemmon *et al.*, 2010), or MassBank (Horai *et al.*, 2010), the second approach involves the creation of an MSI metabolite spectral library using tissue mimetics (or alternative approaches described in Section III..). This is certainly one of the biggest challenges ahead for our community, but Deep Learning promises to give birth to the next generation of automated tools to answer the question more reliably "what are we imaging?".

Acknowledgments

The authors acknowledge the financial support of the Spanish Ministry of Economy and Competitivity through project RTI2018-096061-B-100. GB acknowledges the financial support of the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 713679 and the Universitat Rovira i Virgili (URV). LS acknowledges the financial support of Universitat Rovira i Virgili through the predoctoral grant 2017PMF-PIPF-60. TM acknowledges the financial support of the Universitat Rovira i Virgili through the predoctoral grant ref. PRE2019-089374. MGA acknowledges the financial support from the Agency for Management of University and Research Grants of the Generalitat de Catalunya (AGAUR) through the post-doctoral grant 2018 BP 00188.

7. Figures and tables

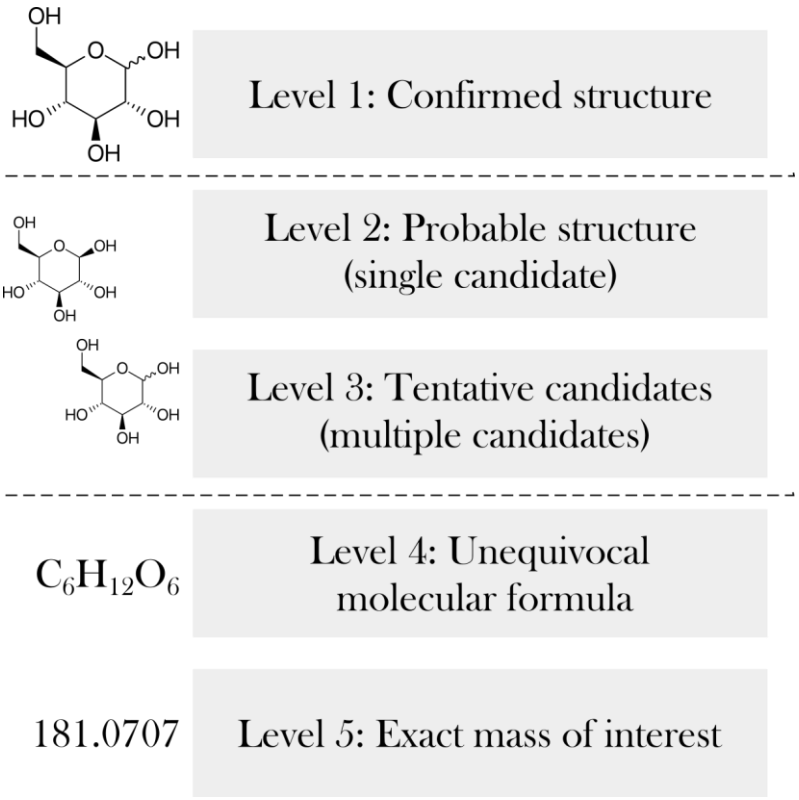


Figure 1. Identification confidence levels proposed by Schymansky et al (Schymanski *et al.*, 2014). Adapted with permission from (Schymanski *et al.*, 2014). Copyright 2022 American Chemical Society.

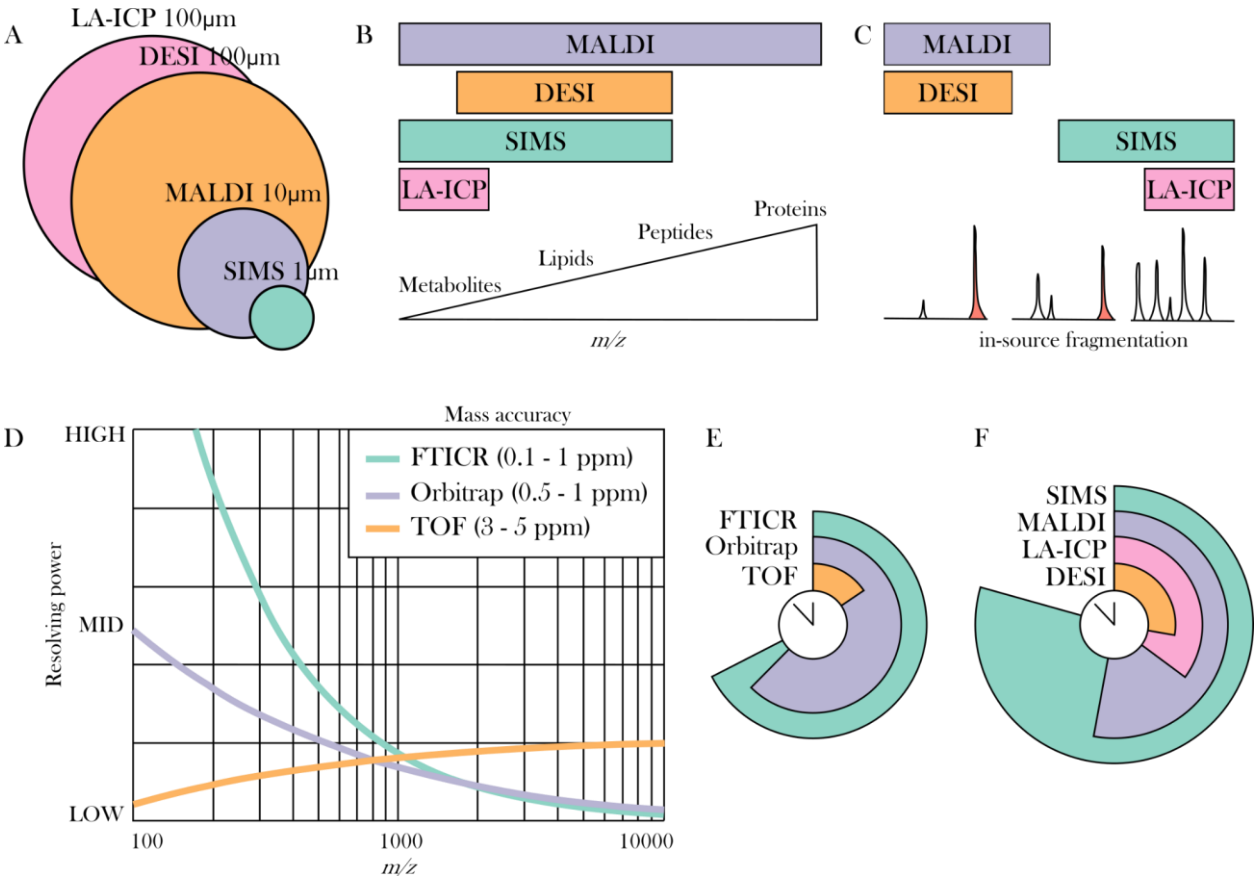


Figure 2. General comparison of the most widely used ion sources and mass analyzers for MSI. **A:** Spatial resolution, **B:** Mass range, and **C:** In-source fragmentation of the four most common ion sources. **D:** Resolving power and mass accuracy and **E:** Acquisition time of the three most common mass analyzers. F: Acquisition time of the four most common ion sources. Adapted with permission from (Evers *et al.*, 2019) (A,B,F), and (Zubarev and Makarov, 2013; Ayet San Andrés *et al.*, 2019) (D). Copyright 2022 American Chemical Society. CC-BY license https://creativecommons.org/licenses/by/4.0/.

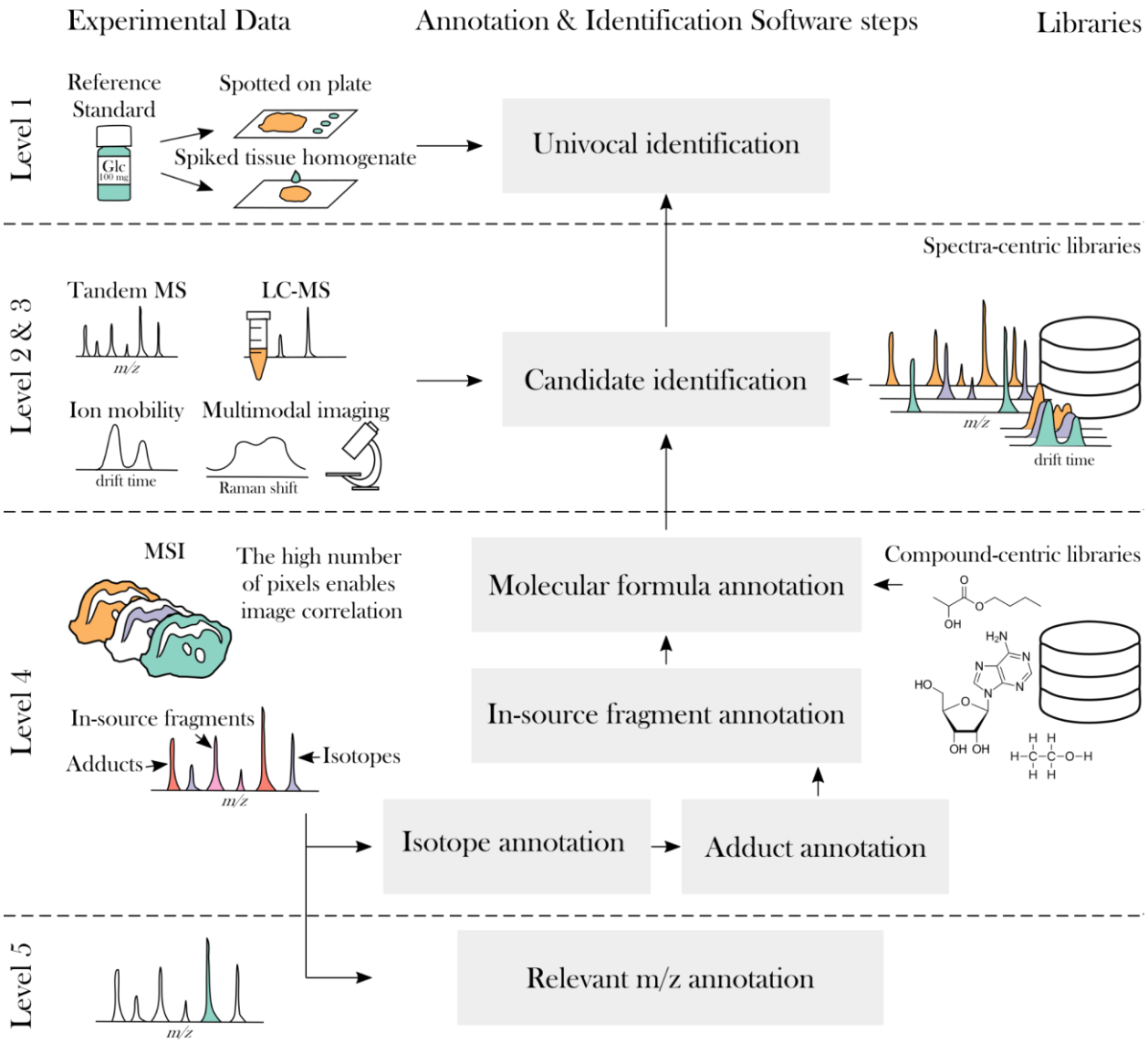


Figure 3. General steps in software annotation & identification in MSI experiments

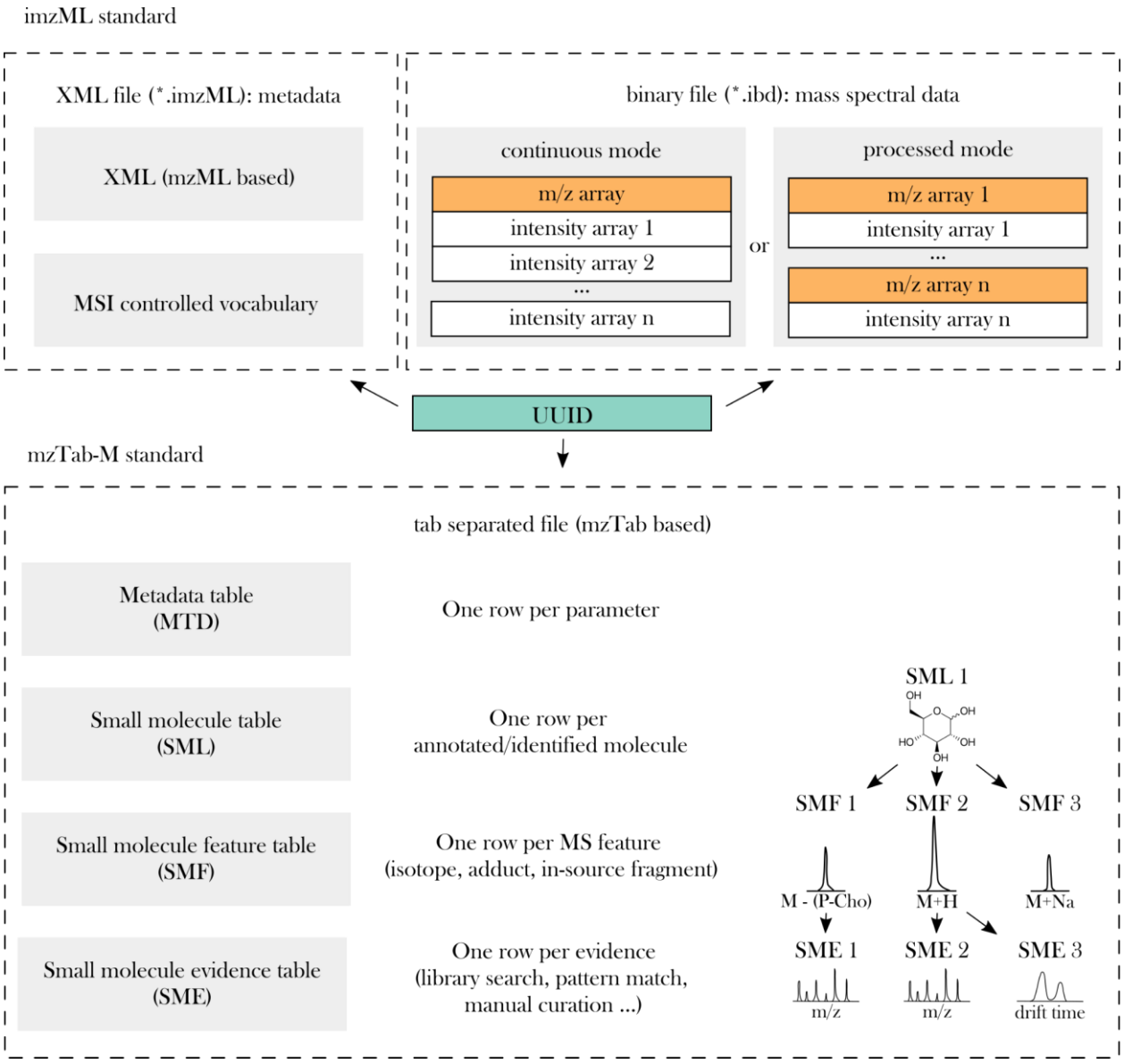


Figure 4. Adaptation of mzTab-M format to be compatible with imzML. A list of Unique Universally Identifiers (UUIDs) would link multiple imzML files from the same study to a single mzTab-M file containing annotations and identifications. Adapted with permission from (Schramm *et al.*, 2012) and (Hoffmann *et al.*, 2019). Copyright 2022 Elsevier. CC-BY license https://creativecommons.org/licenses/by/4.0/.

Table 1. Summary of the principal effects of experimental steps in compound annotation and identification in MSI.

Procedure	Effect in annotation/identification
Sample preservation	
FFPE tissue	● Severe contamination of the spectra. ● Requires deparaffinization. ● Suitable for protein and peptide detection.
Formalin-fixed fresh-frozen tissue	● Formalin may suppress the ionization of amine-containing lipids and introduce $[M+HSO4]^-$ adducts. ● Suitable for sampling all families of compounds but less effective than fresh-frozen in the low mass range.
Fresh-frozen tissue	● No chemical changes in the tissue. ● Risk of shattering and degradation during transport. ● Suitable for sampling all families of compounds.
On-tissue sample treatment	
On-tissue enzymatic digestion	● Proteins are broken down into their peptides, which are easier to ionize and detect than intact proteins. ● Peptides are used to elucidate possible proteins. ● Enzymes hydrolyze proteins in specific bonds.
On-tissue chemical derivatization	● Added moieties increase ionization efficiency and the mass of targeted compounds.
Matrix application (Only for MALDI sources)	
Organic matrices	● Introduce matrix signals in the low mass spectra region and matrix adducts. ● Matrix selection influences which ionization polarity should be used.
Reactive matrices	● More selective measurement. ● Act as derivatization agents.
Isotopically labeled matrices	● Controlled isotopic pattern used to annotate matrix signals and matrix-endogenous adducts.
Inorganic matrices	● Introduce fewer matrix signals. ● In general, produce more fragmentation peaks. ● Some inorganic matrix peaks can be used as calibrants.
Spraying matrix deposition	● Small amount of matrix used. ● Solvent required for desorption of some molecules (such as proteins). ● Higher risk of analyte delocalization.

Sublimation matrix deposition	● More matrix amount required. ● More homogenous layer and less analyte delocalization.
Sputtering matrix deposition	● Requires inorganic material. ● More homogenous layer and less analyte delocalization.
Stable Isotope Labeling	
SIL Matrices	● Shift matrix signals to uncover endogenous signals. ● Distinct isotopic pattern that helps annotation.
Ion Source	
MALDI	● Requires matrix, which might contaminate the spectra. ● Board mass range (up to several kDa). ● Common spatial resolution range from 100 to 10 µm. ● Both ionization polarities (influences type of adducts). ● MALDI-2 increases sensitivity. ● t-MALDI increases routine spatial resolution to 1 µm and below.
DESI	● Minimal sample preparation (dopants may be added to the spray solvent). ● Preference for detecting low molecular weight molecules. ● Spatial resolution range from 200 to 20 µm. ● Both ionization polarities (influences type of adducts).
SIMS	● Minimal sample preparation. ● Suitable for detecting low molecular weight molecules (hard ionization). ● Highest spatial resolution (sub µm). ● Both ionization polarities (influences type of adducts).
LA-ICP	● Used to map atomic composition. ● Spatial resolution range from 200 to 10 µm.
Mass analyzer	
TOF	● Theoretically unlimited mass range. ● Mass resolution increases as *m/z* increases. ● Fastest scan rate.
FTICR	● Ultra-high mass resolution for low-weight compounds. ● Mass resolution decreases linearly as *m/z* increases.
Orbitrap	● Very-high mass resolution for low-weight compounds. ● Mass resolution decreases linearly as *m/z* increases.
Combining MSI with other analytical techniques	

MS/MS	● Structural hypothesis using fragments of precursors. ● Fragmentation patterns may be poor quality or precursor intensity is too low. ● MS/MS libraries mostly contain $[M+H]^+$ fragmentation patterns.
LC-MS and LC-MS/MS	● Most common approach for identification. ● Chromatographic separation allows better spectra interpretability. ● Can use homogenization of the sample or other related biofluids. ● LCM allows the LC-MS analysis of specific tissue regions. ● Usually, Electrospray Ionization (ESI), may generate different adducts than MSI.
IMS	● CCS can be used to resolve isomeric species and get structural information.
Multimodal molecular imaging	● Vibrational spectroscopy can determine functional groups. ● Fluorescence Microscopy enables labeled imaging. ● Registration of images is required.
Reference Standards	
In-solution	● Easy sample preparation. ● Fails to capture matrix effects, ion suppression effects, and endogenous adducts.
On-tissue	● Easy sample preparation. ● Captures matrix effects, ion suppression effects, and endogenous adducts. ● Low extraction efficiency. The standard only interacts with the surface.
Tissue mimetics	● Complex sample preparation. ● Captures matrix effects, ion suppression effects, and endogenous adducts. ● High extraction efficiency. ● Loses spatial context.

Table 2. Summary of software tools for annotation and identification.

Name	**Confidence level**	- **Annotation type** - **Target features**	- **Approach** - **Library**	- **Input data format** - **Output**	- **Programming language** - **Installation** - **License**	**Ref.**
Alex[123]	2-3	- Specific to lipids. - On-tissue MS/MS fragments	- Library-centric - In-house	- .RAW (Thermo Fisher Scientific) - List of identified lipids in 2 levels (annotated by exact mass matching and identified by MS/MS)	- Python - Install from repository - GNU GPL v3.0	(Ellis et al., 2018)
CycloBranch 2	2-4	- General identification and annotation. - Isotopes, adducts and molecular formulas.	- Library-centric - *In silico* library of molecular formulas tunable by the user.	- PROFILE IN imzML, mzML and some proprietary formats. - Interactive tables and spectra.	- C++ - Download and install a standalone package. - GNU GPL v3.0	(Novák et al. 2020)
HIT-MAP	4	- Specific to peptides and proteins. - Full isotopic pattern	- Library-centric - *In silico* library from a protein sequence file in FASTA format.	- imzML - Two sub-folders: one with identification data and the other with containing peptide and protein lists as well as the corresponding ion images	- R - Install from github or docker image - GNU GPL v3.0	(Guo et al., 2021)

LipostarMSI	2-4	- Specific to lipids. - Lipids, metabolites and drug metabolites	- Library-centric - HMDB, LIPID MAPS and in-house.	- imzML (MSI) and csv (MS/MS) - Interactive tables and spectra	- Not specified - Download and install a standalone package. - Private Software	(Tortorella et al., 2020)
mass2adduct	4	- Specific to metabolite-matrix adducts. - Isotopes, adducts, and matrix adducts	- Feature-centric	- imzML - List of adduct masses	- R - Install from github - GNU GPL v3.0	(Janda et al., 2021)
MassPix	4	- General annotation. Specific molecular formulas of lipids. - Isotopes, adducts and molecular formulas.	- Feature-centric	- CENTROID IN imzML - Excel tables	- R - Install from github - GNU GPL v3.0	(Bond et al., 2017)
MSKendrickFilter	5	- Specific compound family annotation. - Suggests compound family based on KMD	- Feature-centric	- imzML - Images of MS signals classified as a user defined compound family.	- Python - Available under request to the authors - Unlicensed	(Kune et al., 2019)

OffsampleAI	5	- Specific to compounds outside of the sample. - MS signals outside of the sample	- Feature-centric	- imzML - Indication of of-sample ion in data	- Python - Install from github / Built-in functionality in METASPACE - Apache 2.0	(Ovchinnikova, et al., 2020)
pySM (METASPACE)	4	- General annotation. - Metabolites high-resolution imaging	- Library-centric - In-house, HMDB, LipidMaps and SwissLipids.	- imzML - CSV table with annotations and FDR level of confidence	- Python - Install from github/ Built-in functionality in METASPACE - Apache 2.0	(Palmer et al., 2016)
ReSCORE METASPACE	4	- General annotation - Improve sensitivity of annotation of metabolites with pySM	- Feature-centric	- Annotations from pySM - CSV table with annotations and q values	- Python - Install from github - Apache 2.0	(C Silva et al., 2018)
rMSIannotation	4	- General annotation. - Isotopes and adducts of metabolites and peptides.	- Feature centric - Modeled after HMDB and Peptide Atlas	- imzML - R objects containing isotopes and adducts.	- R/C++ - Install from github - GNU GPL v3.0	(Sementé et al., 2021)

rMSIcleanup	4	- Specific to matrix peaks - Matrix-related MS signals	- Library-centric - In-house	- imzML - R object containing matrix clusters & PDF with spectra, ion images and matrix clusters	- R - Install from github - GNU GPL v3.0	(Baquer et al., 2020)

8. References

Abdelhamid, H.N. (208) 'Nanoparticle assisted laser desorption/ionization mass spectrometry for small molecule analytes', *Mikrochim Acta,* Mar 1;185(3):200. doi:10.1007/s00604-018-2687-8.

Abdelmoula, W.M. *et al.* (2021) 'Peak learning of mass spectrometry imaging data using artificial neural networks', *Nature communications*, 12(1), p. 5544. doi:10.1038/s41467-021-25744-8.

Aimo, L. *et al.* (2015) 'The SwissLipids knowledgebase for lipid biology', *Bioinformatics* , 31(17), pp. 2860–2866. doi:10.1093/bioinformatics/btv285.

Alam, S.I., Kumar, B. and Kamboj, D.V. (2012) 'Multiplex detection of protein toxins using MALDI-TOF-TOF tandem mass spectrometry: application in unambiguous toxin detection from bioaerosol', *Analytical chemistry*, 84(23), pp. 10500–10507. doi:10.1021/ac3028678.

Alexandrov, T. *et al.* (2019) 'METASPACE: A community-populated knowledge base of spatial metabolomes in health and disease', *bioRxiv* [Preprint]. doi:10.1101/539478.

Alexandrov, T. (2020) 'Spatial Metabolomics and Imaging Mass Spectrometry in the Age of Artificial Intelligence', *Annual review of biomedical data science*, 3, pp. 61–87. doi:10.1146/annurev-biodatasci-011420-031537.

Alexandrov, T. and Bartels, A. (2013) 'Testing for presence of known and unknown molecules in imaging mass spectrometry', *Bioinformatics* , 29(18), pp. 2335–2342. doi:10.1093/bioinformatics/btt388.

Almeida, R. *et al.* (2015) 'Comprehensive Lipidome Analysis by Shotgun Lipidomics on a Hybrid Quadrupole-Orbitrap-Linear Ion Trap Mass Spectrometer', *Journal of the American Society for Mass Spectrometry*, pp. 133–148. doi:10.1007/s13361-014-1013-x.

Amstalden van Hove, E.R., Smith, D.F. and Heeren, R.M.A. (2010) 'A concise review of mass spectrometry imaging', *Journal of chromatography. A*, 1217(25), pp. 3946–3954. doi:10.1016/j.chroma.2010.01.033.

Ayet San Andrés, S. *et al.* (2019) 'High-resolution, accurate multiple-reflection time-of-flight mass spectrometry for short-lived, exotic nuclei of a few events in their ground and low-lying isomeric states', *Physical review C: Nuclear physics*, 99(6), p. 064313. doi:10.1103/PhysRevC.99.064313.

Azodi, C.B., Tang, J. and Shiu, S.-H. (2020) 'Opening the Black Box: Interpretable Machine Learning for Geneticists', *Trends in genetics: TIG*, 36(6), pp. 442–455. doi:10.1016/j.tig.2020.03.005.

Baijnath, S. *et al.* (2016) 'Small molecule distribution in rat lung: a comparison of various cryoprotectants as inflation media and their applicability to MSI', *Journal of molecular histology*, 47(2), pp. 213–219. doi:10.1007/s10735-016-9658-3.

Baquer, G. *et al.* (2020) 'RMSIcleanup: An open-source tool for matrix-related peak annotation in mass spectrometry imaging and its application to silver-assisted laser desorption/ionization', *Journal of cheminformatics*, 12(1), p. 45. doi:10.1186/s13321-020-00449-0.

Barry, J.A. *et al.* (2019) 'Multicenter Validation Study of Quantitative Imaging Mass Spectrometry', *Analytical chemistry*, 91(9), pp. 6266–6274. doi:10.1021/acs.analchem.9b01016.

Basu, S.S. *et al.* (2019) 'Metal Oxide Laser Ionization Mass Spectrometry Imaging (MOLI MSI) Using Cerium(IV) Oxide', *Analytical chemistry*, 91(10), pp. 6800–6807. doi:10.1021/acs.analchem.9b00894.

Becker, J.S. *et al.* (2011) 'Mass spectrometric imaging (MSI) of metals using advanced BrainMet techniques for biomedical research', *International journal of mass spectrometry*, 307(1-3), pp. 3–15. doi:10.1016/j.ijms.2011.01.015.

Becker, J.S. *et al.* (2012) 'Mass spectrometry imaging (MSI) of metals in mouse spinal cord by laser ablation ICP-MS', *Metallomics: integrated biometal science*, 4(3), pp. 284–288. doi:10.1039/c2mt00166g.

Behrmann, J. *et al.* (2018) 'Deep learning for tumor classification in imaging mass spectrometry', *Bioinformatics* , 34(7), pp. 1215–1223. doi:10.1093/bioinformatics/btx724.

Bemis, K.D. *et al.* (2015) 'Cardinal: an R package for statistical analysis of mass spectrometry-based imaging experiments', *Bioinformatics* , 31(14), pp. 2418–2420. doi:10.1093/bioinformatics/btv146.

Bielow, C. *et al.* (2017) 'On Mass Ambiguities in High-Resolution Shotgun Lipidomics', *Analytical chemistry*, 89(5), pp. 2986–2994. doi:10.1021/acs.analchem.6b04456.

Bien, T. *et al.* (2021) 'Transmission-mode MALDI mass spectrometry imaging of single cells: Optimizing sample preparation protocols', *Analytical chemistry*, 93(10), pp. 4513–4520. doi:10.1021/acs.analchem.0c04905.

Böcker, S. *et al.* (2006) 'Decomposing Metabolomic Isotope Patterns', *Lecture Notes in Computer Science*, pp. 12–23. doi:10.1007/11851561_2.

Bond, N.J. *et al.* (2017) 'massPix: an R package for annotation and interpretation of mass spectrometry imaging data for lipidomics', *Metabolomics: Official journal of the Metabolomic Society*, 13(11), p. 128. doi:10.1007/s11306-017-1252-5.

Bowman, A.P. *et al.* (2020) 'Ultra-High Mass Resolving Power, Mass Accuracy, and Dynamic Range MALDI Mass Spectrometry Imaging by 21-T FT-ICR MS', *Analytical chemistry*, 92(4), pp. 3133–3142. doi:10.1021/acs.analchem.9b04768.

Buchberger, A.R. *et al.* (2018) 'Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights', *Analytical Chemistry*, pp. 240–265. doi:10.1021/acs.analchem.7b04733.

Buck, A. *et al.* (2015) 'Distribution and quantification of irinotecan and its active metabolite SN-38 in colon cancer murine model systems using MALDI MSI', *Analytical*

and bioanalytical chemistry, 407(8), pp. 2107–2116. doi:10.1007/s00216-014-8237-2.

Calvano, C.D. *et al.* (2018) 'MALDI matrices for low molecular weight compounds: an endless story?', *Analytical and bioanalytical chemistry*, 410(17), pp. 4015–4038. doi:10.1007/s00216-018-1014-x.

Capellades, J. *et al.* (2016) 'GeoRge: A computational tool to detect the presence of stable isotope labeling in LC/MS-based untargeted metabolomics', *Analytical chemistry*, 88(1), pp. 621–628. doi:10.1021/acs.analchem.5b03628.

Caspi, R. *et al.* (2014) 'The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases', *Nucleic acids research*, 42(Database issue), pp. D459–71. doi:10.1093/nar/gkt1103.

Cassese, A. *et al.* (2016) 'Spatial Autocorrelation in Mass Spectrometry Imaging', *Analytical chemistry*, 88(11), pp. 5871–5878. doi:10.1021/acs.analchem.6b00672.

Castelvecchi, D. (2016) 'Can we open the black box of AI?', *Nature*, 538(7623), pp. 20–23. doi:10.1038/538020a.

Chatterji, B. and Pich, A. (2013) 'MALDI imaging mass spectrometry and analysis of endogenous peptides', *Expert review of proteomics*, 10(4), pp. 381–388. doi:10.1586/14789450.2013.814939.

Chong, J. *et al.* (2018) 'MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis', *Nucleic acids research*, 46(W1), pp. W486–W494. doi:10.1093/nar/gky310.

Chumbley, C.W. *et al.* (2016) 'Absolute Quantitative MALDI Imaging Mass Spectrometry: A Case of Rifampicin in Liver Tissues', *Analytical chemistry*, 88(4), pp. 2392–2398. doi:10.1021/acs.analchem.5b04409.

Cillero-Pastor, B. and Heeren, R.M.A. (2014) 'Matrix-assisted laser desorption ionization mass spectrometry imaging for peptide and protein analyses: a critical review of on-tissue digestion', *Journal of proteome research*, 13(2), pp. 325–335. doi:10.1021/pr400743a.

Claude, E., Jones, E.A. and Pringle, S.D. (2017) 'DESI Mass Spectrometry Imaging (MSI)', *Methods in molecular biology* , 1618, pp. 65–75. doi:10.1007/978-1-4939-7051-3_7.

Clish, C.B. (2015) 'Metabolomics: an emerging but powerful tool for precision medicine', *Cold Spring Harbor molecular case studies*, 1(1), p. a000588. doi:10.1101/mcs.a000588.

'Compound and metabolite distribution measured by MALDI mass spectrometric imaging in whole-body tissue sections' (2007) *International journal of mass spectrometry*, 260(2-3), pp. 195–202. doi:10.1016/j.ijms.2006.10.007.

C Silva, A.S. *et al.* (2018) 'Data-Driven Rescoring of Metabolite Annotations Significantly Improves Sensitivity', *Analytical chemistry*, 90(19), pp. 11636–11642. doi:10.1021/acs.analchem.8b03224.

DeKeyser, S.S. *et al.* (2007) 'Imaging mass spectrometry of neuropeptides in decapod

crustacean neuronal tissues', *Journal of proteome research*, 6(5), pp. 1782–1791. doi:10.1021/pr060603v.

Desiere, F. (2006) 'The PeptideAtlas project', *Nucleic Acids Research*, pp. D655–D658. doi:10.1093/nar/gkj040.

Dewez, F. *et al.* (2019) 'Precise co-registration of mass spectrometry imaging, histology, and laser microdissection-based omics', *Analytical and bioanalytical chemistry*, 411(22), pp. 5647–5653. doi:10.1007/s00216-019-01983-z.

Diehl, H.C. *et al.* (2015) 'The challenge of on-tissue digestion for MALDI MSI- a comparison of different protocols to improve imaging experiments', *Analytical and bioanalytical chemistry*, 407(8), pp. 2223–2243. doi:10.1007/s00216-014-8345-z.

Dilillo, M. *et al.* (2017) 'Mass Spectrometry Imaging, Laser Capture Microdissection, and LC-MS/MS of the Same Tissue Section', *Journal of proteome research*, 16(8), pp. 2993–3001. doi:10.1021/acs.jproteome.7b00284.

Dilillo, M. *et al.* (2017) 'Ultra-High Mass Resolution MALDI Imaging Mass Spectrometry of Proteins and Metabolites in a Mouse Model of Glioblastoma', *Scientific reports*, 7(1), p. 603. doi:10.1038/s41598-017-00703-w.

Djoumbou-Feunang, Y. *et al.* (no date) 'metabolites CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification'. doi:10.3390/metabo9040072.

Dong, F. *et al.* (2020) 'Highly selective isomer fluorescent probes for distinguishing homo-/cysteine from glutathione based on AIE', *Talanta*, 206, p. 120177. doi:10.1016/j.talanta.2019.120177.

Drake, R.R. *et al.* (2018) 'MALDI Mass Spectrometry Imaging of N-Linked Glycans in Tissues', *Advances in experimental medicine and biology*, 1104, pp. 59–76. doi:10.1007/978-981-13-2158-0_4.

Dreisewerd, K., Bien, T. and Soltwisch, J. (2022) 'MALDI-2 and t-MALDI-2 mass spectrometry imaging', *Methods in molecular biology* , 2437, pp. 21–40. doi:10.1007/978-1-0716-2030-4_2.

Dueñas, M.E. *et al.* (2017) 'High spatial resolution mass spectrometry imaging reveals the genetically programmed, developmental modification of the distribution of thylakoid membrane lipids among individual cells of maize leaf', *The Plant journal: for cell and molecular biology*, 89(4), pp. 825–838. doi:10.1111/tpj.13422.

Dufresne, M. *et al.* (2013) 'Silver-assisted laser desorption ionization for high spatial resolution imaging mass spectrometry of olefins from thin tissue sections', *Analytical chemistry*, 85(6), pp. 3318–3324. doi:10.1021/ac3037415.

Dührkop, K. *et al.* (2019) 'SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information', *Nature methods*, 16(4), pp. 299–302. doi:10.1038/s41592-019-0344-8.

Eckelmann, D., Kusari, S. and Spiteller, M. (2018) 'Stable isotope labeling of prodiginines and serratamolides produced by Serratia marcescens directly on agar and simultaneous visualization by matrix-assisted laser desorption/ionization imaging high-

resolution mass spectrometry', *Analytical chemistry*, 90(22), pp. 13167–13172. doi:10.1021/acs.analchem.8b03633.

Ekelöf, M. *et al.* (2018) 'Evaluation of Digital Image Recognition Methods for Mass Spectrometry Imaging Data Analysis', *Journal of the American Society for Mass Spectrometry*, 29(12), pp. 2467–2470. doi:10.1007/s13361-018-2073-0.

Ellis, S.R. *et al.* (2018) 'Automated, parallel mass spectrometry imaging and structural identification of lipids', *Nature methods*, 15(7), pp. 515–518. doi:10.1038/s41592-018-0010-6.

Ellis, S.R. *et al.* (2021) 'Mass spectrometry imaging of phosphatidylcholine metabolism in lungs administered with therapeutic surfactants and isotopic tracers', *Journal of lipid research*, 62, p. 100023. doi:10.1016/j.jlr.2021.100023.

von Eschenbach, W.J. (2021) 'Transparency and the Black Box Problem: Why We Do Not Trust AI', *Philosophy & technology*, 34(4), pp. 1607–1622. doi:10.1007/s13347-021-00477-0.

'Evaluation of freely available software tools for untargeted quantification of 13C isotopic enrichment in cellular metabolome from HR-LC/MS data' (2020) *Metabolic Engineering Communications*, 10, p. e00120. doi:10.1016/j.mec.2019.e00120.

Evers, T.M.J. *et al.* (2019) 'Deciphering Metabolic Heterogeneity by Single-Cell Analysis', *Analytical chemistry*, 91(21), pp. 13314–13323. doi:10.1021/acs.analchem.9b02410.

Fernández, J.A. *et al.* (2011) 'Matrix-assisted laser desorption ionization imaging mass spectrometry in lipidomics', *Analytical and bioanalytical chemistry*, 401(1), pp. 29–51. doi:10.1007/s00216-011-4696-x.

Forsberg, E.M. *et al.* (2018) 'Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online', *Nature protocols*, 13(4), pp. 633–651. doi:10.1038/nprot.2017.151.

Fuchs, K. *et al.* (2018) 'Mapping of drug distribution in the rabbit liver tumor model by complementary fluorescence and mass spectrometry imaging', *Journal of controlled release: official journal of the Controlled Release Society*, 269, pp. 128–135. doi:10.1016/j.jconrel.2017.10.042.

Fu, T. *et al.* (2018) 'Tandem Mass Spectrometry Imaging and in Situ Characterization of Bioactive Wood Metabolites in Amazonian Tree Species Sextonia rubra', *Analytical chemistry*, 90(12), pp. 7535–7543. doi:10.1021/acs.analchem.8b01157.

Gamble, L.J. and Anderton, C.R. (2016) 'Secondary Ion Mass Spectrometry Imaging of Tissues, Cells, and Microbial Systems', *Microscopy today*, 24(2), pp. 24–31. doi:10.1017/S1551929516000018.

Garate, J. *et al.* (2020) 'Influence of Lipid Fragmentation in the Data Analysis of Imaging Mass Spectrometry Experiments'. doi:10.1021/jasms.9b00090.

Gemperline, E., Rawson, S. and Li, L. (2014) 'Optimization and comparison of multiple MALDI matrix application methods for small molecule mass spectrometric imaging',

Analytical chemistry, 86(20), pp. 10030–10035. doi:10.1021/ac5028534.

Gibb, S. and Strimmer, K. (2012) 'MALDIquant: a versatile R package for the analysis of mass spectrometry data', *Bioinformatics* , 28(17), pp. 2270–2271. doi:10.1093/bioinformatics/bts447.

Giordano, S. *et al.* (2016) '3D Mass Spectrometry Imaging Reveals a Very Heterogeneous Drug Distribution in Tumors', *Scientific reports*, 6, p. 37027. doi:10.1038/srep37027.

Gode, D. and Volmer, D.A. (2013) 'Lipid imaging by mass spectrometry--a review', *The Analyst*, 138(5), pp. 1289–1315. Available at: https://pubs.rsc.org/en/content/articlehtml/2013/an/c2an36337b.

Goodwin, R.J.A. *et al.* (2011) 'Qualitative and quantitative MALDI imaging of the positron emission tomography ligands raclopride (a D2 dopamine antagonist) and SCH 23390 (a D1 dopamine antagonist) in rat brain tissue sections using a solvent-free dry matrix application method', *Analytical chemistry*, 83(24), pp. 9694–9701. doi:10.1021/ac202630t.

Grey, A.C. *et al.* (2019) 'A quantitative map of glutathione in the aging human lens', *International journal of mass spectrometry*, 437, pp. 58–68. doi:10.1016/j.ijms.2017.10.008.

Grey, A.C. *et al.* (2021) 'Applications of stable isotopes in MALDI imaging: current approaches and an eye on the future', *Analytical and bioanalytical chemistry*, 413(10), pp. 2637–2653. doi:10.1007/s00216-021-03189-8.

Griss, J. *et al.* (2014) 'The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience', *Molecular & cellular proteomics: MCP*, 13(10), pp. 2765–2775. doi:10.1074/mcp.O113.036681.

Groseclose, M.R. *et al.* (2008) 'High-throughput proteomic analysis of formalin-fixed paraffin-embedded tissue microarrays using MALDI imaging mass spectrometry', *PROTEOMICS*, pp. 3715–3724. doi:10.1002/pmic.200800495.

Groseclose, M.R. *et al.* (2015) 'Imaging MS in Toxicology: An Investigation of Juvenile Rat Nephrotoxicity Associated with Dabrafenib Administration', *Journal of the American Society for Mass Spectrometry*, 26(6), pp. 887–898. doi:10.1007/s13361-015-1103-4.

Groseclose, M.R. and Castellino, S. (2013) 'A mimetic tissue model for the quantification of drug distributions by MALDI imaging mass spectrometry', *Analytical chemistry*, 85(21), pp. 10099–10106. doi:10.1021/ac400892z.

Guo, G. *et al.* (2021) 'Automated annotation and visualisation of high-resolution spatial proteomic mass spectrometry imaging data using HIT-MAP', *Nature communications*, 12(1), p. 3241. doi:10.1038/s41467-021-23461-w.

Gustafsson, O.J.R. *et al.* (2018) 'Balancing sufficiency and impact in reporting standards for mass spectrometry imaging experiments', *GigaScience*, 7(10). doi:10.1093/gigascience/giy102.

Hale, O.J. and Cooper, H.J. (2021) 'Native Mass Spectrometry Imaging of Proteins and Protein Complexes by Nano-DESI', *Analytical chemistry*, 93(10), pp. 4619–4627. doi:10.1021/acs.analchem.0c05277.

Hankin, J.A. *et al.* (2011) 'MALDI mass spectrometric imaging of lipids in rat brain injury models', *Journal of the American Society for Mass Spectrometry*, 22(6), pp. 1014–1021. doi:10.1007/s13361-011-0122-z.

Hankin, J.A., Barkley, R.M. and Murphy, R.C. (2007) 'Sublimation as a method of matrix application for mass spectrometric imaging', *Journal of the American Society for Mass Spectrometry*, 18(9), pp. 1646–1652. doi:10.1016/j.jasms.2007.06.010.

Hansen, R.L., Dueñas, M.E. and Lee, Y.J. (2019) 'Sputter-Coated Metal Screening for Small Molecule Analysis and High-Spatial Resolution Imaging in Laser Desorption Ionization Mass Spectrometry', *Journal of the American Society for Mass Spectrometry*, 30(2), pp. 299–308. doi:10.1007/s13361-018-2081-0.

Hansen, R.L. and Lee, Y.J. (2017) 'Overlapping MALDI-Mass Spectrometry Imaging for In-Parallel MS and MS/MS Data Acquisition without Sacrificing Spatial Resolution', *Journal of the American Society for Mass Spectrometry*, pp. 1910–1918. doi:10.1007/s13361-017-1699-7.

Harkin, C. *et al.* (2021) 'On-tissue chemical derivatization in mass spectrometry imaging', *Mass spectrometry reviews* [Preprint]. doi:10.1002/mas.21680.

Harrison, J.P. and Berry, D. (2017) 'Vibrational Spectroscopy for Imaging Single Microbial Cells in Complex Biological Samples', *Frontiers in microbiology*, 8, p. 675. doi:10.3389/fmicb.2017.00675.

Hartler, J. *et al.* (2011) 'Lipid Data Analyzer: unattended identification and quantitation of lipids in LC-MS data', *Bioinformatics* , 27(4), pp. 572–577. doi:10.1093/bioinformatics/btq699.

Hastings, J. *et al.* (2016) 'ChEBI in 2016: Improved services and an expanding collection of metabolites', *Nucleic acids research*, 44(D1), pp. D1214–D1219. doi:10.1093/nar/gkv1031.

Haug, K. *et al.* (2013) 'MetaboLights - An open-access general-purpose repository for metabolomics studies and associated meta-data', *Nucleic acids research*, 41(D1). doi:10.1093/nar/gks1004.

He, H. *et al.* (2019) '3,4-Dimethoxycinnamic Acid as a Novel Matrix for Enhanced In Situ Detection and Imaging of Low-Molecular-Weight Compounds in Biological Tissues by MALDI-MSI', *Analytical chemistry*, 91(4), pp. 2634–2643. doi:10.1021/acs.analchem.8b03522.

Heijs, B. *et al.* (2020) 'MALDI-2 for the Enhanced Analysis of -Linked Glycans by Mass Spectrometry Imaging', *Analytical chemistry*, 92(20), pp. 13904–13911. doi:10.1021/acs.analchem.0c02732.

Hermann, J. *et al.* (2020) 'Sample preparation of formalin-fixed paraffin-embedded tissue sections for MALDI-mass spectrometry imaging', *Analytical and bioanalytical chemistry*, 412(6), pp. 1263–1275. doi:10.1007/s00216-019-02296-x.

Hermjakob, H. (2006) 'The HUPO Proteomics Standards Initiative - Overcoming the Fragmentation of Proteomics Data', *PROTEOMICS*, pp. 34–38. doi:10.1002/pmic.200600537.

Hoffmann, N. *et al.* (2019) 'mzTab-M: A Data Standard for Sharing Quantitative Results in Mass Spectrometry Metabolomics', *Analytical chemistry*, 91(5), pp. 3302–3310. doi:10.1021/acs.analchem.8b04310.

Hoffmann, N., Hartler, J. and Ahrends, R. (2019) 'jmzTab-M: A Reference Parser, Writer, and Validator for the Proteomics Standards Initiative mzTab 2.0 Metabolomics Standard', *Analytical chemistry*, 91(20), pp. 12615–12618. doi:10.1021/acs.analchem.9b01987.

Horai, H. *et al.* (2010) 'MassBank: a public repository for sharing mass spectral data for life sciences', *Journal of mass spectrometry: JMS*, 45(7), pp. 703–714. doi:10.1002/jms.1777.

Huang, P. *et al.* (2020) 'Toward the Rational Design of Universal Dual Polarity Matrix for MALDI Mass Spectrometry', *Analytical Chemistry*, pp. 7139–7145. doi:10.1021/acs.analchem.0c00570.

Huber, F. *et al.* (2021) 'Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships', *PLoS computational biology*, 17(2), p. e1008724. doi:10.1371/journal.pcbi.1008724.

Iakab, S.A. *et al.* (2021) 'Perspective on Multimodal Imaging Techniques Coupling Mass Spectrometry and Vibrational Spectroscopy: Picturing the Best of Both Worlds', *Analytical chemistry*, 93(16), pp. 6301–6310. doi:10.1021/acs.analchem.0c04986.

Ifa, D.R. *et al.* (2007) 'Development of capabilities for imaging mass spectrometry under ambient conditions with desorption electrospray ionization (DESI)', *International journal of mass spectrometry*, 259(1-3), pp. 8–15. doi:10.1016/j.ijms.2006.08.003.

Jadoul, L. *et al.* (2015) 'A spiked tissue-based approach for quantification of phosphatidylcholines in brain section by MALDI mass spectrometry imaging', *Analytical and bioanalytical chemistry*, 407(8), pp. 2095–2106. doi:10.1007/s00216-014-8232-7.

Janda, M. *et al.* (2021) 'Determination of Abundant Metabolite Matrix Adducts Illuminates the Dark Metabolome of MALDI-Mass Spectrometry Imaging Datasets', *Analytical chemistry*, 93(24), pp. 8399–8407. doi:10.1021/acs.analchem.0c04720.

Jaskolla, T.W. and Karas, M. (2011) 'Compelling evidence for Lucky Survivor and gas phase protonation: the unified MALDI analyte protonation mechanism', *Journal of the American Society for Mass Spectrometry*, 22(6), pp. 976–988. doi:10.1007/s13361-011-0093-0.

Jones, A.R. *et al.* (2012) 'The mzIdentML data standard for mass spectrometry-based proteomics results', *Molecular & cellular proteomics: MCP*, 11(7), p. M111.014381. doi:10.1074/mcp.M111.014381.

Kaddi, C., Parry, R.M. and Wang, M.D. (2011) 'Hypergeometric Similarity Measure for Spatial Analysis in Tissue Imaging Mass Spectrometry', *Proceedings. IEEE International Conference on Bioinformatics and Biomedicine*, 2011, pp. 604–607.

doi:10.1109/BIBM.2011.113.

Kanehisa, M. and Goto, S. (2000) 'KEGG: kyoto encyclopedia of genes and genomes', *Nucleic acids research*, 28(1), pp. 27–30. doi:10.1093/nar/28.1.27.

Karas, M., Glückmann, M. and Schäfer, J. (2000) 'Ionization in matrix-assisted laser desorption/ionization: singly charged molecular ions are the lucky survivors', *Journal of mass spectrometry: JMS*, 35(1). doi:10.1002/(SICI)1096-9888(200001)35:1<1::AID-JMS904>3.0.CO;2-0.

Kaya, I. *et al.* (2018) 'Dual polarity MALDI imaging mass spectrometry on the same pixel points reveals spatial lipid localizations at high-spatial resolutions in rat small intestine', *Analytical Methods*, pp. 2428–2435. doi:10.1039/c8ay00645h.

Khatib-Shahidi, S. *et al.* (2006) 'Direct molecular analysis of whole-body animal tissue sections by imaging MALDI mass spectrometry', *Analytical chemistry*, 78(18), pp. 6448–6456. doi:10.1021/ac060788p.

Kim, S. *et al.* (2019) 'PubChem 2019 update: Improved access to chemical data', *Nucleic acids research*, 47(D1), pp. D1102–D1109. doi:10.1093/nar/gky1033.

Kind, T. and Fiehn, O. (2007) 'Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry', *BMC bioinformatics*, 8, p. 105. doi:10.1186/1471-2105-8-105.

Kompauer, M., Heiles, S. and Spengler, B. (2017) 'Atmospheric pressure MALDI mass spectrometry imaging of tissues and cells at 1.4-μm lateral resolution', *Nature methods*, 14(1), pp. 90–96. doi:10.1038/nmeth.4071.

Kune, C. *et al.* (2019) 'Rapid Visualization of Chemically Related Compounds Using Kendrick Mass Defect As a Filter in Mass Spectrometry Imaging', *Analytical chemistry*, 91(20), pp. 13112–13118. doi:10.1021/acs.analchem.9b03333.

Kyle, J.E. *et al.* (2016) 'Uncovering biologically significant lipid isomers with liquid chromatography, ion mobility spectrometry and mass spectrometry', *The Analyst*, 141(5), pp. 1649–1659. doi:10.1039/c5an02062j.

Łącki, M.K. *et al.* (2021) 'OpenTIMS, TimsPy, and TimsR: Open and Easy Access to timsTOF Raw Data', *Journal of proteome research*, 20(4), pp. 2122–2129. doi:10.1021/acs.jproteome.0c00962.

Lanekoff, I. *et al.* (2013) 'High-speed tandem mass spectrometric in situ imaging by nanospray desorption electrospray ionization mass spectrometry', *Analytical chemistry*, 85(20), pp. 9596–9603. doi:10.1021/ac401760s.

Lapthorn, C., Pullen, F. and Chowdhry, B.Z. (2013) 'Ion mobility spectrometry-mass spectrometry (IMS-MS) of small molecules: separating and assigning structures to ions', *Mass spectrometry reviews*, 32(1), pp. 43–71. doi:10.1002/mas.21349.

Lasch, P. and Noda, I. (2017) 'Two-Dimensional Correlation Spectroscopy for Multimodal Analysis of FT-IR, Raman, and MALDI-TOF MS Hyperspectral Images with Hamster Brain Tissue', *Analytical chemistry*, 89(9), pp. 5008–5016. doi:10.1021/acs.analchem.7b00332.

Lemmon, E.W. *et al.* (2010) 'NIST standard reference database 23', *Reference fluid thermodynamic and transport properties (REFPROP), version*, 9. Available at: https://www.nist.gov/system/files/documents/srd/REFPROP8_manua3.htm.

Leopold, J. *et al.* (2018) 'Recent Developments of Useful MALDI Matrices for the Mass Spectrometric Characterization of Lipids', *Biomolecules*, p. 173. doi:10.3390/biom8040173.

Li, B. *et al.* (2019) '3-Aminophthalhydrazide (Luminol) As a Matrix for Dual-Polarity MALDI MS Imaging', *Analytical chemistry*, 91(13), pp. 8221–8228. doi:10.1021/acs.analchem.9b00803.

Lichtman, J.W. and Conchello, J.-A. (2005) 'Fluorescence microscopy', *Nature methods*, 2(12), pp. 910–919. doi:10.1038/nmeth817.

Lin, J.-R. *et al.* (2016) 'Cyclic immunofluorescence (CycIF), A highly multiplexed method for single-cell imaging', *Current protocols in chemical biology*, 8(4), pp. 251–264. doi:10.1002/cpch.14.

Lisec, J. *et al.* (2006) 'Gas chromatography mass spectrometry–based metabolite profiling in plants', *Nature protocols*, 1(1), pp. 387–396. doi:10.1038/nprot.2006.59.

Li, X., Liu, D. and Wang, Z. (2011) 'Highly selective recognition of naphthol isomers based on the fluorescence dye-incorporated SH-β-cyclodextrin functionalized gold nanoparticles', *Biosensors & bioelectronics*, 26(5), pp. 2329–2333. doi:10.1016/j.bios.2010.10.005.

Loos, M. *et al.* (2015) 'Accelerated isotope fine structure calculation using pruned transition trees', *Analytical chemistry*, 87(11), pp. 5738–5744. doi:10.1021/acs.analchem.5b00941.

Ly, A. *et al.* (2016) 'High-mass-resolution MALDI mass spectrometry imaging of metabolites from formalin-fixed paraffin-embedded tissue', *Nature protocols*, 11(8), pp. 1428–1443. doi:10.1038/nprot.2016.081.

Mainini, V. *et al.* (2013) 'Detection of high molecular weight proteins by MALDI imaging mass spectrometry', *Molecular bioSystems*, 9(6), pp. 1101–1107. doi:10.1039/c2mb25296a.

Martens, L. *et al.* (2011) 'mzML—a community standard for mass spectrometry data', *Molecular & cellular proteomics: MCP*, 10(1). Available at: https://www.mcponline.org/article/S1535-9476(20)31387-6/abstract.

Mascini, N.E. and Heeren, R.M.A. (2012) 'Protein identification in mass-spectrometry imaging', *Trends in analytical chemistry: TRAC*, 40, pp. 28–37. doi:10.1016/j.trac.2012.06.008.

McDonnell, L.A. *et al.* (2008) 'Mass spectrometry image correlation: quantifying colocalization', *Journal of proteome research*, 7(8), pp. 3619–3627. doi:10.1021/pr800214d.

McDonnell, L.A. *et al.* (2015) 'Discussion point: reporting guidelines for mass spectrometry imaging', *Analytical and bioanalytical chemistry*, 407(8), pp. 2035–2045.

doi:10.1007/s00216-014-8322-6.

McLafferty, F.W. (1981) 'Tandem mass spectrometry', *Science*, 214(4518), pp. 280–287. doi:10.1126/science.7280693.

Meier, F. *et al.* (2015) 'Parallel Accumulation-Serial Fragmentation (PASEF): Multiplying Sequencing Speed and Sensitivity by Synchronized Scans in a Trapped Ion Mobility Device', *Journal of proteome research*, 14(12), pp. 5378–5387. doi:10.1021/acs.jproteome.5b00932.

Meier, F. *et al.* (2020) 'diaPASEF: parallel accumulation-serial fragmentation combined with data-independent acquisition', *Nature methods*, 17(12), pp. 1229–1236. doi:10.1038/s41592-020-00998-0.

Mesa Sanchez, D. *et al.* (2020) 'Ion Mobility-Mass Spectrometry Imaging Workflow', *Journal of the American Society for Mass Spectrometry*, 31(12), pp. 2437–2442. doi:10.1021/jasms.0c00142.

Minaee, S. *et al.* (2021) 'Image Segmentation Using Deep Learning: A Survey', *IEEE transactions on pattern analysis and machine intelligence*, PP. doi:10.1109/TPAMI.2021.3059968.

Mounfield, W.P., 3rd and Garrett, T.J. (2012) 'Automated MALDI matrix coating system for multiple tissue samples for imaging mass spectrometry', *Journal of the American Society for Mass Spectrometry*, 23(3), pp. 563–569. doi:10.1007/s13361-011-0324-4.

Nakamura, J. *et al.* (2017) 'Spatially resolved metabolic distribution for unraveling the physiological change and responses in tomato fruit using matrix-assisted laser desorption/ionization-mass spectrometry imaging (MALDI-MSI)', *Analytical and bioanalytical chemistry*, 409(6), pp. 1697–1706. doi:10.1007/s00216-016-0118-4.

Nazari, M. *et al.* (2018) 'Quantitative mass spectrometry imaging of glutathione in healthy and cancerous hen ovarian tissue sections by infrared matrix-assisted laser desorption electrospray ionization (IR-MALDESI)', *The Analyst*, pp. 654–661. doi:10.1039/c7an01828b.

Neumann, N.K.N. *et al.* (2014) 'Automated LC-HRMS(/MS) approach for the annotation of fragment ions derived from stable isotope labeling-assisted untargeted metabolomics', *Analytical chemistry*, 86(15), pp. 7320–7327. doi:10.1021/ac501358z.

Nguyen, S.N. *et al.* (2018) 'Towards High-Resolution Tissue Imaging Using Nanospray Desorption Electrospray Ionization Mass Spectrometry Coupled to Shear Force Microscopy', *Journal of the American Society for Mass Spectrometry*, 29(2), pp. 316–322. doi:10.1007/s13361-017-1750-8.

Niedermeyer, T.H.J. and Strohalm, M. (2012) 'mMass as a software tool for the annotation of cyclic peptide tandem mass spectra', *PloS one*, 7(9), p. e44913. doi:10.1371/journal.pone.0044913.

Niehaus, M. *et al.* (2019) 'Transmission-mode MALDI-2 mass spectrometry imaging of cells and tissues at subcellular resolution', *Nature methods*, 16(9), pp. 925–931. doi:10.1038/s41592-019-0536-2.

Nizioł, J. and Ruman, T. (2013) 'Surface-transfer mass spectrometry imaging on a monoisotopic silver nanoparticle enhanced target', *Analytical chemistry*, 85(24), pp. 12070–12076. doi:10.1021/ac4031658.

Norris, J.L. *et al.* (2007) 'Processing MALDI Mass Spectra to Improve Mass Spectral Direct Tissue Analysis', *International journal of mass spectrometry*, 260(2-3), pp. 212–221. doi:10.1016/j.ijms.2006.10.005.

Norris, J.L. and Caprioli, R.M. (2013) 'Analysis of tissue specimens by matrix-assisted laser desorption/ionization imaging mass spectrometry in biological and clinical research', *Chemical reviews*, 113(4), pp. 2309–2342. doi:10.1021/cr3004295.

Nothias, L.-F. *et al.* (2020) 'Feature-based molecular networking in the GNPS analysis environment', *Nature methods*, 17(9), pp. 905–908. doi:10.1038/s41592-020-0933-6.

Novák, J., Škríba, A. and Havlíček, V. (2020) 'CycloBranch 2: Molecular Formula Annotations Applied to imzML Data Sets in Bimodal Fusion and LC-MS Data Files', *Analytical chemistry*, 92(10), pp. 6844–6849. doi:10.1021/acs.analchem.0c00170.

Ntshangase, S. *et al.* (2019) 'Spatial distribution of elvitegravir and tenofovir in rat brain tissue: Application of matrix-assisted laser desorption/ionization mass spectrometry imaging and liquid chromatography/tandem mass spectrometry', *Rapid communications in mass spectrometry: RCM*, 33(21), pp. 1643–1651. doi:10.1002/rcm.8510.

Ogrinc Potočnik, N. *et al.* (2014) 'Gold sputtered fiducial markers for combined secondary ion mass spectrometry and MALDI imaging of tissue samples', *Analytical chemistry*, 86(14), pp. 6781–6785. doi:10.1021/ac500308s.

Otter, D.W., Medina, J.R. and Kalita, J.K. (2021) 'A Survey of the Usages of Deep Learning for Natural Language Processing', *IEEE transactions on neural networks and learning systems*, 32(2), pp. 604–624. doi:10.1109/TNNLS.2020.2979670.

Ovchinnikova, K., Stuart, L., *et al.* (2020) 'ColocML: machine learning quantifies co-localization between mass spectrometry images', *Bioinformatics* , 36(10), pp. 3215–3224. doi:10.1093/bioinformatics/btaa085.

Ovchinnikova, K., Kovalev, V., *et al.* (2020) 'OffsampleAI: artificial intelligence approach to recognize off-sample mass spectrometry images', *BMC bioinformatics*, 21(1), p. 129. doi:10.1186/s12859-020-3425-x.

Paine, M.R.L. *et al.* (2019) 'Three-Dimensional Mass Spectrometry Imaging Identifies Lipid Markers of Medulloblastoma Metastasis', *Scientific reports*, 9(1), p. 2205. doi:10.1038/s41598-018-38257-0.

Palmer, A. *et al.* (2015) 'Using collective expert judgements to evaluate quality measures of mass spectrometry images', in *Bioinformatics*, pp. i375–i384. doi:10.1093/bioinformatics/btv266.

Palmer, A. *et al.* (2016) 'FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry', *Nature methods*, 14(1), pp. 57–60. doi:10.1038/nmeth.4072.

Patti, G.J., Yanes, O. and Siuzdak, G. (2012) 'Innovation: Metabolomics: the apogee of the omics trilogy', *Nature reviews. Molecular cell biology*, 13(4), pp. 263–269. doi:10.1038/nrm3314.

Perdian, D.C. and Lee, Y.J. (2010) 'Imaging MS Methodology for More Chemical Information in Less Data Acquisition Time Utilizing a Hybrid Linear Ion Trap−Orbitrap Mass Spectrometer', *Analytical Chemistry*, pp. 9393–9400. doi:10.1021/ac102017q.

Phan, N.T.N. *et al.* (2016) 'Laser Desorption Ionization Mass Spectrometry Imaging of Drosophila Brain Using Matrix Sublimation versus Modification with Nanoparticles', *Analytical chemistry*, 88(3), pp. 1734–1741. doi:10.1021/acs.analchem.5b03942.

Phillips, L., Gill, A.J. and Baxter, R.C. (2019) 'Novel Prognostic Markers in Triple-Negative Breast Cancer Discovered by MALDI-Mass Spectrometry Imaging', *Frontiers in oncology*, 0. doi:10.3389/fonc.2019.00379.

Picache, J.A. *et al.* (2019) 'Collision cross section compendium to annotate and predict multi-omic compound identities', *Chemical science* , 10(4), pp. 983–993. doi:10.1039/c8sc04396e.

Pietrowska, M. *et al.* (2016) 'Tissue fixed with formalin and processed without paraffin embedding is suitable for imaging of both peptides and lipids by MALDI-IMS', *Proteomics*, 16(11-12), pp. 1670–1677. doi:10.1002/pmic.201500424.

Pirman, D.A. *et al.* (2013) 'Identifying tissue-specific signal variation in MALDI mass spectrometric imaging by use of an internal standard', *Analytical chemistry*, 85(2), pp. 1090–1096. doi:10.1021/ac3029618.

Pitt, J.J. (2009) 'Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry', *The Clinical biochemist. Reviews / Australian Association of Clinical Biochemists*, 30(1), pp. 19–34. Available at: https://www.ncbi.nlm.nih.gov/pubmed/19224008.

Polanska, J. *et al.* (2012) 'Gaussian mixture decomposition in the analysis of MALDI-TOF spectra', *Expert Systems*, 29(3), pp. 216–231. doi:10.1111/j.1468-0394.2011.00582.x.

Pornwilard *et al.* (2013) 'Bioimaging of copper deposition in Wilson's diseases mouse liver by laser ablation inductively coupled plasma mass spectrometry imaging (LA-ICP-MSI)', *International journal of mass spectrometry*, 354-355, pp. 281–287. doi:10.1016/j.ijms.2013.07.006.

Porta Siegel, T. *et al.* (2018) 'Mass Spectrometry Imaging and Integration with Other Imaging Modalities for Greater Molecular Understanding of Biological Tissues', *Molecular imaging and biology: MIB: the official publication of the Academy of Molecular Imaging*, 20(6), pp. 888–901. doi:10.1007/s11307-018-1267-y.

Race, A.M. *et al.* (2021) 'Deep Learning-Based Annotation Transfer between Molecular Imaging Modalities: An Automated Workflow for Multimodal Data Integration', *Cite This: Anal. Chem*, 93, pp. 3061–3071. doi:10.1021/acs.analchem.0c02726.

Ràfols, P., Vilalta, D., Torres, S., *et al.* (2018) 'Assessing the potential of sputtered gold nanolayers in mass spectrometry imaging for metabolomics applications', *PloS one*,

13(12), p. e0208908. doi:10.1371/journal.pone.0208908.

Ràfols, P., Castillo, E. del, *et al.* (2018) 'Novel automated workflow for spectral alignment and mass calibration in MS imaging using a sputtered Ag nanolayer', *Analytica chimica acta*, 1022, pp. 61–69. doi:10.1016/j.aca.2018.03.031.

Ràfols, P., Vilalta, D., Brezmes, J., *et al.* (2018) 'Signal preprocessing, multivariate analysis and software tools for MA(LDI)-TOF mass spectrometry imaging for biological applications', *Mass Spectrometry Reviews*, pp. 281–306. doi:10.1002/mas.21527.

Ràfols, P. *et al.* (2020) 'RMSIproc: An R package for mass spectrometry imaging data processing', *Bioinformatics* , 36(11), pp. 3618–3619. doi:10.1093/bioinformatics/btaa142.

Ren, J.-L. *et al.* (2018) 'Advances in mass spectrometry-based metabolomics for investigation of metabolites', *RSC advances*, 8(40), pp. 22335–22350. doi:10.1039/C8RA01574K.

Ruttkies, C. *et al.* (2016) 'MetFrag relaunched: incorporating strategies beyond in silico fragmentation', *Journal of cheminformatics*, 8, p. 3. doi:10.1186/s13321-016-0115-9.

Ryan, D.J., Spraggins, J.M. and Caprioli, R.M. (2019) 'Protein identification strategies in MALDI imaging mass spectrometry: a brief review', *Current opinion in chemical biology*, 48, pp. 64–72. doi:10.1016/j.cbpa.2018.10.023.

Rzagalinski, I. and Volmer, D.A. (2017) 'Quantification of low molecular weight compounds by MALDI imaging mass spectrometry - A tutorial review', *Biochimica et Biophysica Acta: Proteins and Proteomics*, 1865(7), pp. 726–739. doi:10.1016/j.bbapap.2016.12.011.

Sabine Becker, J. (2013) 'Imaging of metals in biological tissue by laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS): state of the art and future developments', *Journal of mass spectrometry: JMS*, 48(2), pp. 255–268. doi:10.1002/jms.3168.

Salek, R. (2019) 'Data Sharing and Standards', in *Metabolomics*. Chapman and Hall/CRC, pp. 235–252. Available at: https://www.taylorfrancis.com/chapters/edit/10.1201/9781315370583-10/data-sharing-standards-reza-salek.

Salek, R.M. *et al.* (2013) 'The role of reporting standards for metabolite annotation and identification in metabolomic studies', *GigaScience*. doi:10.1186/2047-217x-2-13.

Sans, M., Feider, C.L. and Eberlin, L.S. (2018) 'Advances in mass spectrometry imaging coupled to ion mobility spectrometry for enhanced imaging of biological tissues', *Current opinion in chemical biology*, 42, pp. 138–146. doi:10.1016/j.cbpa.2017.12.005.

Schramm, T. *et al.* (2012) 'ImzML - A common data format for the flexible exchange and processing of mass spectrometry imaging data', *Journal of proteomics*, 75(16), pp. 5106–5110. doi:10.1016/j.jprot.2012.07.026.

Schrimpe-Rutledge, A.C. *et al.* (2016) 'Untargeted Metabolomics Strategies—

Challenges and Emerging Directions', *Journal of the American Society for Mass Spectrometry*, 27(12), pp. 1897–1905. doi:10.1007/s13361-016-1469-y.

Schulz, S. *et al.* (2019) 'Advanced MALDI mass spectrometry imaging in pharmaceutical research and drug development', *Current opinion in biotechnology*, 55, pp. 51–59. doi:10.1016/j.copbio.2018.08.003.

Schymanski, E.L. *et al.* (2014) 'Identifying small molecules via high resolution mass spectrometry: communicating confidence', *Environmental science & technology*, 48(4), pp. 2097–2098. doi:10.1021/es5002105.

Sementé, L. *et al.* (2021) 'rMSIannotation: A peak annotation tool for mass spectrometry imaging based on the analysis of isotopic intensity ratios', *Analytica chimica acta*, 1171, p. 338669. doi:10.1016/j.aca.2021.338669.

Shariatgorji, M. *et al.* (2012) 'Deuterated matrix-assisted laser desorption ionization matrix uncovers masked mass spectrometry imaging signals of small molecules', *Analytical chemistry*, 84(16), pp. 7152–7157. doi:10.1021/ac301498m.

Shariatgorji, M. *et al.* (2014) 'Direct targeted quantitative molecular imaging of neurotransmitters in brain tissue sections', *Neuron*, 84(4), pp. 697–707. doi:10.1016/j.neuron.2014.10.011.

Shariatgorji, M. *et al.* (2015) 'Pyrylium Salts as Reactive Matrices for MALDI-MS Imaging of Biologically Active Primary Amines', *Journal of the American Society for Mass Spectrometry*, 26(6), pp. 934–939. doi:10.1007/s13361-015-1119-9.

Shariatgorji, R. *et al.* (2020) 'Bromopyrylium Derivatization Facilitates Identification by Mass Spectrometry Imaging of Monoamine Neurotransmitters and Small Molecule Neuroactive Compounds', *Journal of the American Society for Mass Spectrometry*, 31(12), pp. 2553–2557. doi:10.1021/jasms.0c00166.

Shobo, A. *et al.* (2016) 'MALDI MSI and LC-MS/MS: Towards preclinical determination of the neurotoxic potential of fluoroquinolones', *Drug testing and analysis*, 8(8), pp. 832–838. doi:10.1002/dta.1862.

Signor, L. *et al.* (2007) 'Analysis of erlotinib and its metabolites in rat tissue sections by MALDI quadrupole time-of-flight mass spectrometry', *Journal of mass spectrometry: JMS*, 42(7), pp. 900–909. doi:10.1002/jms.1225.

Smets, T. *et al.* (2019) 'Evaluation of Distance Metrics and Spatial Autocorrelation in Uniform Manifold Approximation and Projection Applied to Mass Spectrometry Imaging Data', *Analytical chemistry*, 91(9), pp. 5706–5714. doi:10.1021/acs.analchem.8b05827.

Smith, C.A. *et al.* (2005) 'METLIN: a metabolite mass spectral database', *Therapeutic drug monitoring*, 27(6), pp. 747–751. doi:10.1097/01.ftd.0000179845.53213.39.

Soltwisch, J. *et al.* (2015) 'Mass spectrometry imaging with laser-induced postionization', *Science*, 348(6231), pp. 211–215. doi:10.1126/science.aaa1051.

Spengler, B. (2015) 'Mass spectrometry imaging of biomolecular information', *Analytical chemistry*, 87(1), pp. 64–82. doi:10.1021/ac504543v.

Steven, R.T. *et al.* (2019) 'Construction and testing of an atmospheric-pressure

transmission-mode matrix assisted laser desorption ionisation mass spectrometry imaging ion source with plasma ionisation enhancement', *Analytica chimica acta*, 1051, pp. 110–119. doi:10.1016/j.aca.2018.11.003.

Sud, M. *et al.* (2007) 'LMSD: LIPID MAPS structure database', *Nucleic acids research*, 35(Database issue), pp. D527–32. doi:10.1093/nar/gkl838.

Sud, M. *et al.* (2016) 'Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools', *Nucleic acids research*, 44(D1), pp. D463–70. doi:10.1093/nar/gkv1042.

Sumner, L.W. *et al.* (2007) 'Proposed minimum reporting standards for chemical analysis', *Metabolomics: Official journal of the Metabolomic Society*, 3(3), pp. 211–221. doi:10.1007/s11306-007-0082-2.

Sünderhauf, N. *et al.* (2018) 'The limits and potentials of deep learning for robotics', *The International journal of robotics research*, 37(4-5), pp. 405–420. doi:10.1177/0278364918770733.

Sun, N. *et al.* (2018) 'High-Resolution Tissue Mass Spectrometry Imaging Reveals a Refined Functional Anatomy of the Human Adult Adrenal Gland', *Endocrinology*, 159(3), pp. 1511–1524. doi:10.1210/en.2018-00064.

Thomas, A. *et al.* (2012) 'Sublimation of new matrix candidates for high spatial resolution imaging mass spectrometry of lipids: enhanced information in both positive and negative polarities after 1,5-diaminonapthalene deposition', *Analytical chemistry*, 84(4), pp. 2048–2054. doi:10.1021/ac2033547.

Tortorella, S. *et al.* (2020) 'LipostarMSI: Comprehensive, Vendor-Neutral Software for Visualization, Data Analysis, and Automated Molecular Identification in Mass Spectrometry Imaging', *Journal of the American Society for Mass Spectrometry*, 31(1), pp. 155–163. doi:10.1021/jasms.9b00034.

Touboul, D. and Brunelle, A. (2016) 'What more can TOF-SIMS bring than other MS imaging methods?', *Bioanalysis*, 8(5), pp. 367–369. doi:10.4155/bio.16.11.

Towers, M.W. *et al.* (2018) 'Optimised Desorption Electrospray Ionisation Mass Spectrometry Imaging (DESI-MSI) for the Analysis of Proteins/Peptides Directly from Tissue Sections on a Travelling Wave Ion Mobility Q-ToF', *Journal of the American Society for Mass Spectrometry*, pp. 2456–2466. doi:10.1007/s13361-018-2049-0.

Trede, D. *et al.* (2012) 'O5. SCiLS Lab: software for analysis and interpretation of large MALDI-IMS datasets', *OurCon 2012*, p. 50. Available at: https://orbi.uliege.be/bitstream/2268/131796/1/Book%20of%20abstractsOurCon2012_v1.3%20with%20covers.pdf#page=51.

Trimpin, S. *et al.* (2009) 'Field-free transmission geometry atmospheric pressure matrix-assisted laser desorption/ionization for rapid analysis of unadulterated tissue samples', *Rapid communications in mass spectrometry: RCM*, 23(18), pp. 3023–3027. doi:10.1002/rcm.4213.

Tsugawa, H. *et al.* (2015) 'MS-DIAL: data-independent MS/MS deconvolution for

comprehensive metabolome analysis', *Nature methods*, 12(6), pp. 523–526. doi:10.1038/nmeth.3393.

Tuck, M. *et al.* (2021) 'Multimodal Imaging Based on Vibrational Spectroscopies and Mass Spectrometry Imaging Applied to Biological Tissue: A Multiscale and Multiomics Review', *Analytical chemistry*, 93(1), pp. 445–477. doi:10.1021/acs.analchem.0c04595.

'Ultra-high resolution MALDI-FTICR-MSI analysis of intact proteins in mouse and human pancreas tissue' (2019) *International journal of mass spectrometry*, 437, pp. 10–16. doi:10.1016/j.ijms.2017.11.001.

Unsihuay, D., Mesa Sanchez, D. and Laskin, J. (2021) 'Quantitative Mass Spectrometry Imaging of Biological Systems', *Annual review of physical chemistry*, 72, pp. 307–329. doi:10.1146/annurev-physchem-061020-053416.

Uslu, A. *et al.* (2017) 'Imidazole/benzimidazole-modified cyclotriphosphazenes as highly selective fluorescent probes for Cu 2+: synthesis, configurational isomers, and crystal structures', *Dalton transactions: a journal of inorganic chemistry*, 46(28), pp. 9140–9156. Available at: https://pubs.rsc.org/en/content/articlehtml/2017/dt/c7dt01134b.

Vaysse, P.-M. *et al.* (2017) 'Mass spectrometry imaging for clinical research – latest developments, applications, and current limitations', *The Analyst*, pp. 2690–2712. doi:10.1039/c7an00565b.

Vos, D.R.N. *et al.* (2019) 'Class-specific depletion of lipid ion signals in tissues upon formalin fixation', *International journal of mass spectrometry*, 446, p. 116212. doi:10.1016/j.ijms.2019.116212.

Wäldchen, F. *et al.* (2020) 'Multifunctional Reactive MALDI Matrix Enabling High-Lateral Resolution Dual Polarity MS Imaging and Lipid C═C Position-Resolved MS Imaging', *Analytical chemistry*, 92(20), pp. 14130–14138. doi:10.1021/acs.analchem.0c03150.

Walzer, M. *et al.* (2013) 'The mzquantml data standard for mass spectrometry--based quantitative studies in proteomics', *Molecular & cellular proteomics: MCP*, 12(8), pp. 2332–2340. Available at: https://www.mcponline.org/article/S1535-9476(20)32541-X/abstract.

Wang, L. *et al.* (2019) 'Peak Annotation and Verification Engine for Untargeted LC-MS Metabolomics', *Analytical chemistry*, 91(3), pp. 1838–1846. doi:10.1021/acs.analchem.8b03132.

'Why the metabolism field risks missing out on the AI revolution' (2019) *Nature metabolism*, 1(10), pp. 929–930. doi:10.1038/s42255-019-0133-9.

Wishart, D.S. *et al.* (2018) 'HMDB 4.0: the human metabolome database for 2018', *Nucleic acids research*, 46(D1), pp. D608–D617. doi:10.1093/nar/gkx1089.

Wishart, D.S. (2019) 'Metabolomics for Investigating Physiological and Pathophysiological Processes', *Physiological reviews*, 99(4), pp. 1819–1875. doi:10.1152/physrev.00035.2018.

Wisztorski, M. *et al.* (2010) 'MALDI direct analysis and imaging of frozen versus FFPE tissues: what strategy for which sample?', *Methods in molecular biology* , 656, pp. 303–322. doi:10.1007/978-1-60761-746-4_18.

Xian, F., Hendrickson, C.L. and Marshall, A.G. (2012) 'High resolution mass spectrometry', *Analytical chemistry*, 84(2), pp. 708–719. doi:10.1021/ac203191t.

Xue, J. *et al.* (2020) 'Enhanced in-Source Fragmentation Annotation Enables Novel Data Independent Acquisition and Autonomous METLIN Molecular Identification', *Analytical chemistry*, 92(8), pp. 6051–6059. doi:10.1021/acs.analchem.0c00409.

Xu, J. *et al.* (2019) 'Integrated UPLC-Q/TOF-MS Technique and MALDI-MS to Study of the Efficacy of YiXinshu Capsules Against Heart Failure in a Rat Model', *Frontiers in pharmacology*, 10, p. 1474. doi:10.3389/fphar.2019.01474.

Yagnik, G.B., Korte, A.R. and Lee, Y.J. (2013) 'Multiplex mass spectrometry imaging for latent fingerprints', *Journal of mass spectrometry: JMS*, 48(1), pp. 100–104. doi:10.1002/jms.3134.

Yang, J. and Caprioli, R.M. (2011) 'Matrix Sublimation/Recrystallization for Imaging Proteins by Mass Spectrometry at High Spatial Resolution', *Analytical chemistry*, 83(14), pp. 5728–5734. doi:10.1021/ac200998a.

Ye, H. *et al.* (2013) 'MALDI mass spectrometry-assisted molecular imaging of metabolites during nitrogen fixation in the Medicago truncatula-Sinorhizobium meliloti symbiosis', *The Plant journal: for cell and molecular biology*, 75(1), pp. 130–145. doi:10.1111/tpj.12191.

Ye, H. *et al.* (2014) 'Top-down proteomics with mass spectrometry imaging: a pilot study towards discovery of biomarkers for neurodevelopmental disorders', *PloS one*, 9(4), p. e92831. doi:10.1371/journal.pone.0092831.

Yoon, S. and Lee, T.G. (2018) 'Biological tissue sample preparation for time-of-flight secondary ion mass spectrometry (ToF–SIMS) imaging', *Nano Convergence*. doi:10.1186/s40580-018-0157-y.

Zavalin, A. *et al.* (2012) 'Direct imaging of single cells and tissue at sub-cellular spatial resolution using transmission geometry MALDI MS', *Journal of mass spectrometry: JMS*, 47(11), p. i. doi:10.1002/jms.3132.

Zavalin, A. *et al.* (2015) 'Tissue protein imaging at 1 μm laser spot diameter for high spatial resolution and high imaging speed using transmission geometry MALDI TOF MS', *Analytical and bioanalytical chemistry*, 407(8), pp. 2337–2342. doi:10.1007/s00216-015-8532-6.

Zhang, G. *et al.* (2020) 'DESI-MSI and METASPACE indicates lipid abnormalities and altered mitochondrial membrane components in diabetic renal proximal tubules', *Metabolomics: Official journal of the Metabolomic Society*, 16(1), p. 11. doi:10.1007/s11306-020-1637-8.

Zhang, W. *et al.* (2021) 'Spatially aware clustering of ion images in mass spectrometry imaging data using deep learning', *Analytical and Bioanalytical Chemistry*, pp. 2803–2819. doi:10.1007/s00216-021-03179-w.

Zhang, Z., Kuang, J. and Li, L. (2013) 'Liquid chromatography-matrix-assisted laser desorption/ionization mass spectrometric imaging with sprayed matrix for improved sensitivity, reproducibility and quantitation', *The Analyst*, 138(21), pp. 6600–6606. doi:10.1039/c3an01225e.

Zhan, L. *et al.* (2021) 'MALDI-TOF/TOF tandem mass spectrometry imaging reveals non-uniform distribution of disaccharide isomers in plant tissues', *Food chemistry*, 338, p. 127984. doi:10.1016/j.foodchem.2020.127984.

Zhou, Q., Fülöp, A. and Hopf, C. (2021) 'Recent developments of novel matrices and on-tissue chemical derivatization reagents for MALDI-MSI', *Analytical and bioanalytical chemistry*, 413(10), pp. 2599–2617. doi:10.1007/s00216-020-03023-7.

Zhou, Z. *et al.* (2020) 'Ion mobility collision cross-section atlas for known and unknown metabolite annotation in untargeted metabolomics', *Nature communications*, 11(1), p. 4334. doi:10.1038/s41467-020-18171-8.

Zubarev, R.A. and Makarov, A. (2013) 'Orbitrap mass spectrometry', *Analytical chemistry*, 85(11), pp. 5288–5296. doi:10.1021/ac4001223.

Züllig, T. and Köfeler, H.C. (2021) 'HIGH RESOLUTION MASS SPECTROMETRY IN LIPIDOMICS', *Mass spectrometry reviews*, 40(3), pp. 162–176. doi:10.1002/mas.21627.

9. Supplementary Materials

Table S1. Use of the terms annotation and identification Mass Spectrometry Imaging publications in the last 5 years. The criteria for correct use of the term "identification" has been relaxed to include not only reference standard matched (Level 1) but also library matched (Levels 2 and 3) orthogonal measurements. Exact mass matching (Levels 4 and 5) is considered "annotation". Only those papers with correct distinction between "annotation" and "identification" in all instances of the word have been labeled as compliant.

Paper		**Use referring to molecular annotation/identification**			**Use in other contexts (spatial, tissue, peak picking…)**	**Correct distinction**
Reference	**Type**	**Annotation**	**Identification**	**Putative/ Tentative**		
(Hermann et al. 2020)	Experimental Protocol	1	4	0	4	NO
(Prentice et al. 2019)	Experimental Protocol	0	15	0	0	YES
(Dexter et al. 2019)	Experimental Protocol	0	4	1	0	YES
(Zubair et al. 2016)	Experimental Protocol	0	17	2	0	YES

(Cillero-Pastor and Heeren 2014)	Experimental Protocol	0	48	0	0	YES
(Zhou et al. 2021)	Experimental Protocol	0	6	0	0	YES
(Diehl et al. 2015)	Experimental Protocol	0	12	0	3	YES
(Harkin et al. 2021)	Experimental Protocol	0	10	0	8	YES
(Tobias and Hummon 2020)	Experimental Protocol	1	1	0	1	NO
(Dueñas et al. 2019)	Experimental Protocol	0	5	1	0	YES
(Prentice et al. 2019)	Experimental Protocol	0	10	0	7	NO
(Ly et al. 2019)	Experimental Protocol	0	2	0	2	YES
(Hansen and Lee 2017)	Experimental Protocol	0	5	0	0	YES
(Erich et al. 2017)	Experimental Protocol	0	1	0	2	YES
(Dong et al. 2016)	Experimental Protocol	0	8	0	0	YES
(Chughtai et al. 2012)	Experimental Protocol	0	24	0	0	NO
(Cerruti et al. 2011)	Experimental Protocol	0	4	0	0	YES
(Novák et al. 2020)	Software	9	2	0	0	YES
(Falcetta et al. 2018)	Software	0	14	0	15	NO
(Baquer et al. 2020)	Software	37	8	2	3	NO
(Tortorella et al. 2020)	Software	6	75	4	11	YES
(Bond et al. 2017)	Software	11	2	3	0	NO
(Ellis et al. 2018)	Software	6	56	0	1	NO

(Palmer et al. 2017)	Software	64	4	1	0	YES
(Ovchinnikova et al. 2020)	Software	14	2	0	0	YES
(Eriksson et al. 2019)	Software	1	5	0	3	NO
(Kune et al. 2019)	Software	0	12	1	2	NO
(C Silva et al. 2018)	Software	101	4	0	4	NO
(Hohenester et al. 2020)	Instrumental	14	2	0	0	NO
(Nagy et al. 2019)	Instrumental	4	7	0	1	NO
(Eliuk and Makarov 2015)	Instrumental	1	20	0	2	NO
(Perry et al. 2008)	Instrumental	0	24	0	0	NO
(Züllig and Köfeler 2021)	Application	16	12	0	0	YES
(Conceição et al. 2021)	Application	0	15	0	1	YES
(DeLaney and Li 2020)	Application	0	33	0	0	YES
(Bowman et al. 2020)	Application	2	38	1	0	NO
(Spraggins et al. 2015)	Application	0	42	2	0	NO
(Fu et al. 2020)	Application	0	4	0	0	YES
(Groseclose et al. 2007)	Application	0	28	1	0	YES
(Ly et al. 2016)	Application	7	21	0	6	NO
(Pepi et al. 2021)	Application	1	8	0	2	YES
(Züllig et al. 2020)	Application	23	25	0	0	NO

(Han et al. 2019)	Application	0	33	0	7	YES
(Bruinen et al. 2018)	Application	2	47	2	2	YES
(Vaysse et al. 2017)	Application	1	41	0	5	NO
(Moreno-Gordaliza et al. 2017)	Application	0	23	8	2	NO
(Ong et al. 2016)	Application	0	28	0	5	YES
(Maier et al. 2013)	Application	3	93	0	1	YES
(Garrett and Yost 2010)	Application	0	52	0	0	NO
(Tuck et al. 2021)	Future Perspectives	5	42	1	13	NO
(Alexandrov 2020)	Future Perspectives	20	21	1	1	NO
(Gilmore et al. 2019)	Future Perspectives	5	6	0	1	NO
(Buchberger et al. 2018)	Future Perspectives	2	21	0	11	NO
(Müller et al. 2021)	Other	1	15	2	3	NO
(Kirchberger-Tolstik et al. 2021)	Other	3	14	0	24	NO
(Zang et al. 2021)	Other	0	5	2	3	NO
(Pinsky et al. 2021)	Other	2	16	0	9	YES
(Truong et al. 2021)	Other	1	14	1	1	NO

CHAPTER 3:

rMSIfragment: Automated lipid in-source fragmentation and adduct annotation of lipids in MALDI-MSI

Gerard Baquer [1], Lluc Sementé [1], Pere Ràfols [1,2,3,*], Lucía Martín-Saiz [4], Christoph Bookmeyer [1], José A. Fernández [4], Xavier Correig [1,2,3] & María García-Altares [1,2]

1 Department of Electronic Engineering, University Rovira I Virgili, Tarragona, Spain

2 Spanish Biomedical Research Center in Diabetes and Associated Metabolic Disorders (CIBERDEM), Madrid, Spain

3 Institut D'Investigacio Sanitaria Pere Virgili, Tarragona, Spain

4 Department of Physical Chemistry, Faculty of Science and Technology, University of the Basque Country (UPV/EHU), Leioa, Spain

*Correspondence:

Pere Ràfols, Department of Electronic Engineering, University Rovira i Virgili, Tarragona, Spain. Email: pere.rafols@urv.cat

Abstract

Matrix-Assisted Laser Desorption Ionization Mass Spectrometry Imaging (MALDI-MSI) spatially resolves the chemical composition of tissues. Its use to study lipids is of particular interest, as they play a key role in understanding key biological processes in health and disease. However, the identification of lipids in MALDI-MSI remains a challenge due to the lack of chromatographic separation or untargeted tandem mass spectrometry. In particular, fragments originating in the ion source (in-source fragments) clutter the spectra and hamper identification. This is particularly challenging in lipidomics as certain lipid families fragment into others. Recent studies have proposed the use of MALDI in-source fragmentation to infer structural information and aid identification. Here we present rMSIfragment, an open-source R package that exploits known adducts and fragmentation pathways to confidently annotate lipids in MALDI-MSI. The annotations are ranked using a novel score that demonstrates an area under the curve of 0.7 in ROC curves using HPLC and Target-Decoy validations. rMSIfragment is applicable to multiple MALDI-MSI workflows as it demonstrates comparable performance across sample types and experimental setups. Finally, we demonstrate that annotation tools that overlook in-source fragmentation incorrectly annotate a substantial fraction of in-source fragments as endogenous lipids. Annotation tools should consider in-source to increase annotation confidence and reduce the number of false positives.

Code Availability: rMSIfragment is available under the GPLv3 at www.github.com/gbaquer/rMSIfragment

Keywords: In-source fragmentation, In-source decay, lipids, annotation, Mass Spectrometry Imaging

1. Introduction

Matrix-Assisted Laser Desorption Ionization Mass Spectrometry Imaging (MALDI-MSI) is an analytical technique used in biochemical and clinical studies to reveal the chemical composition and spatial information of organic tissues (Israr et al. 2020; Nishidate et al. 2019; He et al. 2021; Baijnath 2022). It provides valuable information in many applications, including the understanding and diagnosis of complex diseases such as cancer (Berghmans et al. 2020; Janßen et al. 2022; Denti et al. 2021; Ma and Fernández 2022; Nascentes Melo et al. 2022; Coy et al. 2022; Notarangelo et al. 2022), diabetes (Miyamoto et al. 2016; Z. Wang et al. 2021; Harkin et al. 2022), Alzheimer's (Kaya et al. 2018, 2020; Ikegawa et al. 2019) and infectious diseases (Tans et al. 2021; Nguyen et al. 2022). In particular, the study of lipids is pivotal, as they play important roles in different pathways in health and disease (Muro, Atilla-Gokcumen, and Eggert 2014).

Despite the surge of MALDI-MSI's popularity, associating each mass-to-charge (m/z) signal with univocal molecular identifications remains challenging: (1) samples include thousands of molecules; (2) each molecule is responsible for several MS signals (isotopes, adducts, in-source fragments, multiple charges…); and (3) isomers and isobars cannot be resolved using only MS1 (Baquer et al. 2022).

Traditional mass spectrometry techniques rely on chromatographic separation (LC-MS, GC-MS) for sample simplification (Zhou et al. 2012). However, MALDI-MSI does not include such separation steps. Complementary, tandem mass spectrometry can

augment the depth of the chemical analysis by providing fragmentation information on single molecules. Many MALDI-MSI instruments are equipped with tandem-MS capabilities (Bruker's ultrafleXtreme, Thermo Scientific's MALDI LTQ Orbitrap XL, or Waters' MALDI SYNAPT G2-Si) but untargeted imaging MS/MS is not routinely feasible due to (1) prohibitive running times, (2) limited parental ion selectivity and intensity, and (3) increased data size and complexity. In fact, the use of tandem-MS is not supported by the standard format for MSI data (.imZML) (Schramm et al. 2012). For all these reasons, untargeted fragmentation of all ions in a sample is only possible using highly specialized instrumental setups (Ellis et al. 2018; Heijs et al. 2020). Other studies choose to fragment only a subset of target ions (Fu et al. 2018; Takeo et al. 2019; Zhan et al. 2021), reducing the depth of the chemical analysis.

Traditionally, In-Source Decay (ISD) or In-Source Fragmentation (ISF) (i.e. the natural and unavoidable generation of fragments inside the MALDI ion source) has been considered an undesired artifact and thus minimized (Hu et al. 2022). ISD depends mainly on the chemical structure of the analyte and ionization conditions such as ionization temperature or voltage (Hu et al. 2022) and can be problematic in the study of lipids, as several fragmentation pathways lead to isobaric lipid species. As an example, phosphatidic acid (PA) fragments can be produced in-source from their phosphatidylserine (PS) counterparts, and phosphatidylcholine (PC) in-source fragments are isobaric to endogenous phosphatidylethanolamine (PE) species (Hu et al. 2022). These known lipid fragmentation pathways result in falsely low concentrations of lipids suffering from ISD and falsely high concentrations of lipids overlapping with isobaric in-source fragments. Additionally, if not properly annotated and removed, in-source fragments can yield an increased number of incorrect annotations using common MALDI-MSI annotation tools such as LipoStar, METASPACE, and rMSIannotation (Tortorella et al. 2020; Palmer et al. 2016; Sementé et al. 2021).

Nevertheless, recent studies have advocated for the use of well-characterized MALDI-ISD as a fast way of obtaining complementary fragmentation information to assist identification in the analysis of large molecules (van der Burgt et al. 2019; Nicolardi et al. 2022; Antone et al. 2019) and even lipids (H.-Y. J. Wang and Hsu 2022, 2020). Some of these studies use automated protein ISD annotation tools like ProteinProspector (Baker, P.R. and Clauser, K.R. http://prospector.ucsf.edu) or DataAnalysis (Bruker) but heavily rely on manual annotation. The use of ISD to strengthen identification has also been applied in MALDI-MSI in the field of top-down proteomics. Debois et al. demonstrated the use of ISD for in-situ de novo sequencing of several proteins on a porcine eye lens and a mouse brain slice (Debois et al. 2010; Zimmerman et al. 2011). Similarly, Ait-Belkacem et al. (Ait-Belkacem et al. 2014) used ISD to identify several tumorigenic proteins in glioblastoma mouse brain tissue with MALDI-MSI. Franceschi et al. proposed a semi-automated workflow for in-source fragmentation annotation based on initial manual annotation of parental metabolites followed by an Intensity Correlation Analysis to find ions with a high spatial correlation and thus assumed to be in-source fragments. They used this approach to image flavonols and dihydrochalcones in golden apple samples (Franceschi et al. 2012).

Recently, Garate et al. (Garate et al. 2020), reported the in-source fragmentation pathways and adduct formation of the 17 main lipid classes in MALDI-MSI (Supplementary Figure S1 & Table S1). Here we propose rMSIfragment, a software solution that exploits these known in-source fragmentation pathways to increase confidence in lipid annotations. Our novel ranking score combines the times a given lipid

has been found in the dataset (adducts and in-source fragments) and their spatial correlation to filter out unlikely lipids. After validation with HPLC and 2 different Target-Decoy approaches, rMSIfragment demonstrates an Area Under the Curve (AUC) of over 0.7 on multiple sample types and experimental conditions. We also find that ISD-agnostic annotation tools like METASPACE can falsely annotate in-source fragments as endogenous lipids.

2. Algorithm Description

2.1. Input and output format

MSI datasets are provided as rMSIproc (Ràfols et al. 2020) peak matrix ([# pixels] x [# m/z] intensity matrix). Refer to the public repository (https://github.com/prafols/rMSIproc) for instructions on how to convert profile and centroid mode .imZML files to the peak matrix format. The theoretical fragmentation pathways and adducts for each lipid class (Garate et al. 2020) (Supplementary Figure S1 & Table S1) are already included in rMSIcleanup. To use alternative fragmentations and adducts: (1) export the theoretical ones with exportFragsAdds(), (2) modify the .csv file, and (3) use that path in the function call. Additionally, The ppm tolerance for exact mass searches against the database is provided by the user.

The algorithm produces a table of annotations (lipid annotation, number of carbons: number of double chains, and chemical formula) for each of the monoisotopic masses found in the data set. Each annotation is associated with a likelihood score to assess and an FDR score to assist manual curation. The resulting table can be exported as a .csv file (Supplementary Table S3). Additionally, the annotation results can be used to generate a molecular network graph.

2.2. Database search & likelihood score

As reported by Garate et al. (Garate et al. 2020) ISD can produce lipid fragments that overlap with endogenous lipids (Fig 1A). rMSIfragment estimates the likelihood of each lipid by computing two novel metrics: (1) the number of m/z features in the dataset associated with the lipid (considering adducts and in-source fragments) (LO), and (2) the spatial correlation between all of the m/z features associated with the lipid (C). The final ranking score is given by $S = LO * C$.

Figure 1B summarizes the complete workflow. Initially, rMSIannotation (Sementé et al. 2021) is used to perform deisotoping (removal of m/z features considered to be isotopes). Later, all remaining m/z features are matched against LIPID MAPS considering all theoretically feasible adducts and fragments. Each lipid class has a different list of theoretical adducts and in-source fragmentation pathways. The results of the annotation are stored in an R data.frame that can easily be exported to .csv for manual inspection.

During validation, rMSIcleanup uses two alternative decoy libraries (Fig 1C) to estimate the False Discovery Rate (FDR) and performance (ROC AUC) of the annotation. The first decoy library is formed by generating highly unlikely adducts and fragmentation pathways. The second decoy library replaces LIPIDMAPS with a list of compounds found

in non-animal specimens and thus highly unlikely to be found in animals. Both approaches will be described more in detail in section 3.2.

Although by default, LipidMAPS (O'Donnell et al. 2019) is used to perform database searches, the user can adjust the software to use any publicly available database (HMDB, MoNA, METLIN, NIST ...) or an in-house compound database.

3. Results

3.1. rMSIfragment matches HPLC annotations in human nevi samples.

As initial validation, we challenged rMSIfragment with the annotation of human nevi samples G1-G15, manually and validated with HPLC by Garate et al. (Garate et al. 2020). These annotations were used to estimate the performance of our automatic annotation tool.

In the samples acquired in negative ion mode (G9-G15) (Figure 2A) rMSIfragment retrieves 91.81% of the HPLC-validated annotations reported in the original publication. To retrieve at least 50% of the HPLC annotations we need to retain the top 9 hits with the highest S score in each m/z feature. When only keeping the top 5 annotations rMSIfragment returns 35.92% of the HPLC-validated annotations. Figure 2B shows that the HPLC-validated annotations obtain a significantly higher S score than the annotations not validated by HPLC. Additionally, Figure 2C shows a ROC area under the curve (AUC) of 0.7 when using the HPLC annotations as validation.

In positive ion mode (G1-G8) (Figure 2D) rMSIfragment retrieves 56.18% of all the HPLC-validated annotations. In spite of the lower performance when compared to the negative ion mode samples, rMSIfragment is still capable of retrieving 18.56% of the annotations when focusing on the top 5 annotations for each MS feature. HPLC-validated annotations still report a significantly higher S score (Figure 2E) but present a slightly worse performance at 0.63 AUC.

An alternative representation of the same results is shown in Figure S2, where a global threshold for the likelihood score (S) is used instead of selecting the top N annotations per m/z feature.

These results suggest that rMSIfragment has a strong classification power when annotating lipid in-source fragments.

3.2. rMSIfragment shows high performs in a target-decoy validation

The validation in the previous section focuses exclusively on HPLC-validated ions. It assumes that all HPLC-validated lipids should be found in the samples and the rest should not. HPLC and MALDI-MSI are fundamentally different analytical techniques (Awad, Khamis, and El-Aneed 2015) and it is possible that lipids found with one technique are not found in the other. We consider that HPLC annotations can accurately

estimate the number of true positives (present and matched by rMSIfragment) and false negatives (present but not matched by rMSIfragment); but may fail to estimate true negatives (not present and not matched by rMSIfragment) and false positives (not present but matched by rMSIfragment).

In an attempt to gauge and overcome the limitations of the HPLC validation, we propose a second validation based on a target-decoy search strategy, a commonly used approach in MSI (Palmer et al. 2016; Guo et al. 2021). This strategy runs rMSIfragment on the same MSI data using our target library (LIPID-MAPS) and a decoy library containing compounds that should not be found in the sample. The decoy library matches the size and distribution of masses of the target library, to ensure that randomly generated masses are equally likely to hit either of the two databases. The rate of matches in the decoy can then be used to estimate measures such as true positives, true negatives, false positives, false negatives, and False Discovery Rate (FDR) (Elias and Gygi 2007).

Figure 3 shows the results of the target-decoy validation on the human nevi datasets (G1-G15) using a decoy library composed of highly unlikely adducts and fragmentation pathways, an approach adapted from pySM (Palmer et al. 2016). The classification performance obtained a value of 0.72 AUC for the negative ion mode datasets (Figure 3A) and 0.6 AUC for the positive ion mode datasets (Figure 3C). These results are consistent with the performances obtained in HPLC validation. When retaining the top 10 matches for each MS feature the FDR is estimated to be 14.93% in negative ion mode (Figure 3B) and 34.24% in positive ion mode (Figure 3D). Similarly, when only retaining the top 5 matches per MS feature the FDR is 4.5% and 17.95% respectively.

As extra validation, we created a second decoy library, a subset of ChEBI (Hastings et al. 2016) containing only metabolites found in non-animal specimens (plants, algae, fungi, and bacteria) and xenobiotics. We assume these compounds are highly unlikely to be found in human nevi samples. The results are summarized in Figure S3. The performance on the negative ion mode samples is consistent with the one estimated in previous approaches (0.73 AUC). In the positive ion mode samples, the performance is estimated to be higher than in previous validations (0.75 AUC). In both cases, the FDR is estimated to be under 10% when retaining the top 10 matches.

The two decoy libraries provided comparable estimations of performance and FDR. Nevertheless, the definition of a decoy library based on highly-unlikely compounds (Figure S3) depends highly on the sample being annotated. The definition of highly-unlikely adducts and fragments, on the other hand, has already been discussed (Palmer et al. 2016) and can be assumed to be much more sample-independent. Additionally, its performance estimates are more conservative and closely match the results obtained in the HPLC validation. For all these reasons, further validations use highly-unlikely adducts and fragments as a decoy library.

These results further confirm that rMSIfragment can confidently annotate lipids and their fragments.

3.3. rMSIfragment is applicable to different experimental conditions

In order to determine its applicability to other experimental conditions we challenged rMSIfragment with the annotation of 12 different publicly available datasets from METASPACE (Datasets M1-M12) (Alexandrov et al. 2019). Since the real lipid composition of these datasets is unknown, we use the target and decoy approach based on highly-unlikely adducts and fragments to estimate the performance of rMSIfragment in each dataset. Figure 4 summarizes the results. The performances ranged between 0.65 AUC and 0.84 AUC (μ=0.74 AUC, σ= 7.5%) (Figure 4A). These performances are comparable to the performance on human nevi samples validated with HPLC.

Additionally, when grouping the datasets according to different experimental parameters (ionization mode, matrix, tissue, analyzer, specimen, or mass range) the differences in performance were not found to be significant (p-value > 0.1). This demonstrates that rMSIfragment is applicable to a wide range of experimental conditions.

3.4. Annotation software must consider in-source fragmentation

Major automatic annotation tools such as pySM (Palmer et al. 2016), LipostarMSI (Tortorella et al. 2020) or rMSIannotation (Sementé et al. 2021) completely overlook in-source fragmentation and almost exclusively focus on protonated and alkali ions. Ignoring in-source fragmentation during annotation could potentially lead to false annotations (Baquer et al. 2022). This is a particular concern in lipidomics where several lipids fragment in-source and become isobaric to other lipids (Garate et al. 2020).

To assess the impact of in-source fragmentation on lipid annotation we annotate datasets M1-M12 with METASPACE (Alexandrov et al. 2019), only considering traditional MALDI adducts (M+H, M+Na, M+K in positive ion mode and M-H, M+Cl in negative mode). We then annotate the same datasets using rMSIfragment, considering all adducts and fragmentation pathways specified in Supplementary Figure 1 and Table 1. LipidMAPS is used with both tools.

Figure 5A summarizes the overall comparison of annotations between the 2 tools. On average, 48.6% of the annotations returned by METASPACE are also found with rMSIcleanup. Interestingly, crossing the annotations also allow us to determine that, on average, 54.21% of METASPACE annotations are overlapped with at least one in-source fragment found by rMSIfragment. Supplementary Figure S4 shows the same results color-coded based on the different sample and experimental parameters.

To exemplify this overlap we highlight the annotation of m/z 887.57 from a human lung cancer biopsy prepared with NEDC and analyzed in negative mode with an FTICR (Dataset M5) (Figure 5B). Both tools reliably annotate this m/z feature as PI 38:3 (M-H). In the same dataset, rMSIfragment also finds 3 adducts ($M-OH$, $M-CH_3$, $M+Na-2H$) and 3 in-source fragments ($M-CH_3-NH_2$, $M-H-C_4H_{10}O_5$, $M-2H_2O-NH_2$) with high spatial correlation to the parental ion. Two of these in-source fragments ($M-H-C_4H_{10}O_5$, $M-2H_2O-NH_2$) are overlapped with 2 METASPACE annotations (PA 40:5 M-H, PS 40:5 M-H).

These results are not a comparison between tools but rather a quantification of the negative impact that in-source fragmentation has on automatic annotation tools. It is

clear that annotation tools in MSI need to take into account in-source fragmentation. Due to the fundamental limitations of MS1, software tools are unable to resolve an endogenous PA from an isobaric PA originating from the in-source fragmentation of its PS counterpart. However, rMSIfragment mitigates the issue by (1) making the user aware of the potential overlap and (2) giving a higher score to the lipid found forming other adducts and in-source fragments.

3.5. rMSIfragment provides a molecular network to visually interpret the results

Figure 6 shows an example exploration of the annotation results using rMSIfragment GUI. The top 3 annotations for m/z 744.55 are shown as individual molecular networks including adducts (purple) and in-source fragments (yellow) (Figure 6A). The number of lipid occurrences (LC) and spatial correlation (C) are shown as the main metrics to filter out unlikely lipids. When selecting the desired molecular network the user can view the spatial distribution of all adducts and in-source fragments.

4. Discussion and Conclusions

We have demonstrated the performance of rMSIfragment on 15 human nevi datasets with two orthogonal approaches: (1) matching its annotations to HPLC and (2) using a target-decoy approach. Both approaches yield similar performance estimations (0.7 AUC and 0.6 AUC for the samples acquired in negative and positive mode respectively).

As a next step, we deployed rMSIfragment to annotate lipids and their fragments in 12 publicly available samples covering a wide combination of samples and experimental setups. The performances obtained are comparable and often better than the ones obtained on the human nevi datasets. Additionally, rMSIcleanup shows a high lipidome coverage overlap when compared to available annotation tools like METASPACE (Alexandrov et al. 2019). Collectively, these results indicate that rMSIfragment is capable of reliably annotating lipids and their in-source fragments across multiple experimental conditions

One key highlight of our study is the importance of considering in-source fragmentation pathways when performing molecular annotation. We have found that overlooking ISD pathways, can lead to up to 75% of the reported lipid annotations to be overlapped with at least one in-source fragment. rMSIfragment mitigates this issue through two mechanisms: (1) unlikely lipids with low occurrences (number of adducts and in-source fragments) and poor spatial correlation are filtered out, and (2) the user is aware of the overlap, allowing them to be cautious with their interpretation of the automated annotations.

We envision three new avenues to further improve the automatic annotation of lipids and their in-source fragments: (1) exploiting known ion suppression effects between different lipid classes, (2) exploiting MS/MS libraries, and (3) compiling MALDI-ISD or MALDI-MS/MS libraries.

Ion suppression effects strongly favor certain classes of lipids, difficulting the analysis of suppressed species (Boskamp and Soltwisch 2020). In positive mode, for instance, phosphatidylcholines (PC) display a strong signal in detriment to lower signals of phosphatidylethanolamines (PE) or phosphatidylserines (PS), phosphatidylglycerol (PG)

or phosphatidylinositol (PI). In negative mode the effect is reversed, PC species show lower signals in samples containing other lipid species in favor of other lipid species. These interactions have been characterized in the past (Boskamp and Soltwisch 2020) and could be considered to define a new ranking score to filter out unlikely lipid annotations with intensity values that contradict them.

Initially, we approached the annotation of MALDI-MSI in-source fragmentation by modeling and exploiting similarities of MALDI-MSI spectra to publicly available MS/MS libraries. These approaches were inspired by recent advancements in the LC-MS community (Xue et al. 2020), where the in-source fragmentation in an ESI source is enhanced to yield fragmentation patterns similar to those present in METLIN to aid the quick identification of metabolites. We compared individual MALDI-MSI spectra to MS/MS fragmentation spectra available in MS2ID (https://github.com/jmbadia/MS2ID) and in-silico fragmentation algorithms such as MetFrag (Ruttkies et al. 2016), CFM-ID (F. Wang et al. 2021) and Sirius (Dührkop et al. 2019). The preliminary results suggested that the use of MS/MS fragmentation spectra available in databases and in-silico tools was not sufficient to annotate in-source fragmentation in MALDI-MSI confidently. The two main limitations are (1) the different ion sources used (MALDI vs ESI) and (2) the underrepresentation of common MALDI adducts such as M+Na and M+K (< 10%) in MS/MS libraries. This is of particular interest given that different adducts can yield different fragmentation patterns (Fuchs et al. 2007).

To overcome these limitations, one interesting avenue would be the compilation of MALDI-ISD libraries. This community-wide effort would help better characterize MALDI-ISD in a wide range of biomolecules. A MALDI-MS/MS library, perhaps a more pressing interest of the MALDI community, could already provide a lot of information due to the common ionization mechanisms. These two libraries would be invaluable tools to foster the development of the next generation of ISD annotation algorithms in MALDI-MSI.

In conclusion, neglecting in-source fragmentation leads to an increased number of false lipid annotations. rMSIfragment mitigates this effect by prioritizing annotations of lipids found forming multiple adducts and in-source fragments.

Abbreviations

2,5-Dihydroxybenzoic acid (DHB), 2,5-Dihydroxybenzoic acid with 13C-labeled aromatic ring (*13C*^{6}-DHB)), alpha-Cyano-4-hydroxycinnamic acid (CHCA), 9-Aminoacridine (9AA), 1,5-Diaminonaphthalene (DAN), N-(1-naphthyl) ethylenediamine dihydrochloride (NEDC), Machine Learning (ML), Artificial Intelligence (AI), Human Metabolome Database (HMDB), Standard Isotope Labeled (SIL), False Discovery Rate (FDR), Receiver Operating Characteristic (ROC), Nuclear Magnetic Resonance (NMR), Matrix Assister Laser Desorption Ionization (MALDI), Mass Spectrometry Imaging (MSI), Total Ion Current (TIC), Uniform Manifold Approximation and Projection (UMAP), Competitive Fragmentation Modeling for Metabolite Identification (CFM-ID), Fourier-transform ion cyclotron resonance (FTICR), Time Of Flight (TOF), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Indium tin oxide (ITO), Area Under the Curve (AUC)

Acknowledgments

We acknowledge Angelos Rigopoulos (EMBL), Sergio Heli Triana (EMBL), Elisa Ruhland (IBMP), Jessica Lukowski (PNNL), Patricia Thomsen (University of Copenhagen), Anne Mette Handler (University of Copenhagen), David Rudd (Murdoch University), and respective colleagues as the original contributors of the METASPACE datasets used for validation. We acknowledge Jone Garate, Sergio Lage, Arantza Perez-Valle, Begoña Ochoa, M. Dolores Boyano, and Roberto Fernandez for their characterization of lipid in-source fragmentation (Garate et al. 2020) used in this study.

Authors contributions

GB, PR, XC, and MG developed the concept for the study. LM, CB, and JAF designed and performed mass spectrometry imaging experiments. GB developed the computational workflow in collaboration with LS, PR, XC, and MG. GB and MG wrote the original manuscript with substantial edits and contributions from all authors. JAF, and XC provided supervision, project administration, and funding.

Funding

The authors acknowledge the financial support of the Spanish Ministry of Economy and Competitivity through projects TEC2015-69076-P and RTI2018096061-B-100. GB acknowledges the financial support of the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 713679 and the Universitat Rovira i Virgili (URV). LS acknowledges the financial support of Universitat Rovira i Virgili through the pre-doctoral grant 2017PMF-PIPF-60. MGA acknowledges the financial support from the Agency for Management of University and Research Grants of the Generalitat de Catalunya (AGAUR) through the postdoctoral grant 2018 BP 00188.

Availability of data and materials

The platform-independent R package rMSIfragment presented in this publication is freely available under the terms of the GNU General Public License v3.0 at https://github.com/gbaquer/rMSIfragment. The datasets supporting the conclusions of this article are available in the Mendeley Data repository. Datasets M1-M12 are available at https://metaspace2020.eu/ (References provided in Table S1)

5. Materials and Methods

A total of 20 different datasets were used to validate this study. Human nevi samples acquired in positive (G1-G8) and negative (G9-G15) ion mode from a previous study (Garate et al. 2020) were used to perform HPLC validation and determine the best target-decoy strategy for further validation. Publically available METASPACE (Alexandrov et al. 2019) datasets M1-M12 were used to demonstrate the applicability of rMSIfragment to different sample types and experimental conditions. Finally, we demonstrated an application of rMSIfragment with 2 newly acquired mouse brain datasets (B1-B2) and described next. Table 1 summarizes the main processing parameters for each of the 20 datasets.

5.1. Materials

Indium tin oxide (ITO)-coated glass slides were obtained from Bruker Daltonics (Bremen, Germany).

5.2. Sample preparation

All mice brain samples were provided by the animal facility at the Faculty of Medicine and Health Sciences of the University Rovira i Virgili. All tissues were snap-frozen at −80 °C after collection and kept at this temperature during shipping and storing until MSI acquisition. The tissues were sectioned with a Leica CM-1950 cryostat (Leica Biosystems Nussloch GmbH) located at the Centre for Omics Sciences (COS) of the University Rovira i Virgili into 10 µm sections. Tissue sections were thaw-mounted onto ITO-coated glass slides.

The sputtering system ATC Orion 8-HV (AJA International, N. Scituate, MA, USA) was used to deposit a gold nanolayer onto each tissue section. An argon atmosphere with a pressure of 30 mTorr was used to create the plasma in the gun. The working distance of the plate was set to 35 mm. The sputtering conditions were ambient temperature using DC mode at 100 W for 10 s. With these parameters, an Au layer thickness of roughly 5 nm was obtained. The deposition times were short to prevent the substrate temperature from increasing excessively and, consequently, degrading metabolites.

5.3. MALDI-MS acquisition

A MALDI TOF/TOF ultrafleXtreme instrument with SmartBeam II Nd:YAG/355 nm laser from Bruker Daltonics available at COS was used for MSI acquisition. Acquisitions were carried out by operating the laser at 2 kHz and collecting a total of 500 shots per pixel.

The TOF spectrometer was operated in positive ion, reflectron mode, in m/z ranges according to Table 1. The spectrometer was calibrated prior to MSI data acquisition using Au cluster peaks as internal reference masses.

5.4. MSI data processing

Datasets B1-B2 and G1-G15, originally in .RAW format (Thermo Fischer), were exported to .mzML using ProteoWizard msConvert (Adusumilli and Mallick 2017), and later converted to .imzML (Schramm et al. 2012) using imzMLConverter (Race, Styles, and Bunch 2012). The software rMSIproc (Ràfols et al. 2020) was used to process the data and generate a peak matrix in centroid mode. The default processing parameters were used. The Signal-to-Noise Ratio (SNR) threshold was set to 5 and the Savitzky– Golay smoothing had a kernel size of 7. Peaks appearing in less than 5% of the pixels were filtered out. Peaks within a window of 6 data points or scans were binned together as the same mass peak.

Datasets M1-M12, already in centroid mode .imZML, were imported using rMSIproc (Ràfols et al. 2020).

Deisotoping was performed using rMSIannotation (Sementé et al. 2021). No data normalization was performed. Data were visualized and explored using rMSI (Ràfols et al. 2017).

6. Figures

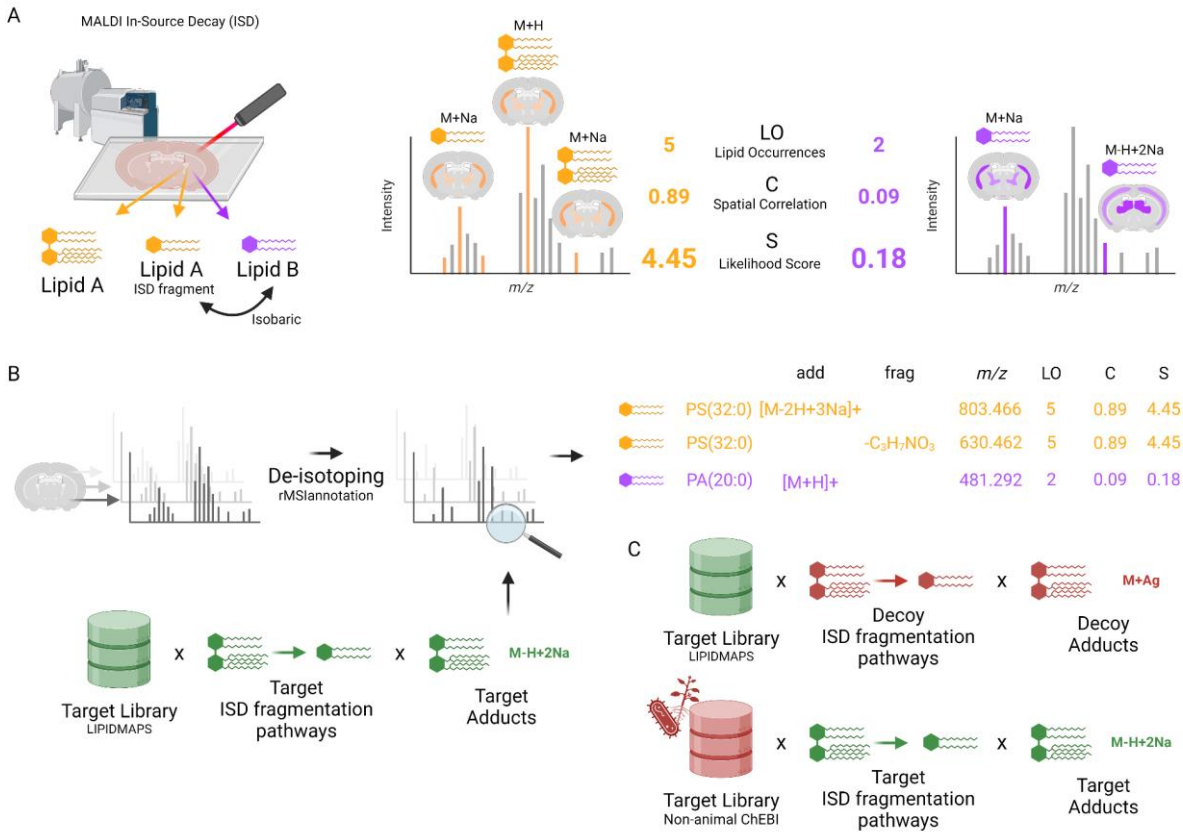


Figure 1. Main algorithmic foundations for the annotation of in-source fragments. **(A)** ISD can generate in-source fragments that overlap with endogenous lipids. The use of Lipid Occurrences (LO) and Spatial Correlation (C) allows rMSIfragment to rank the likelihood of isobaric lipid annotations. **(B)** General rMSIfragment flux diagram. **(C)** Two alternative decoy libraries based on highly unlikely adducts and fragmentation pathways (top) and highly unlikely compounds (bottom).

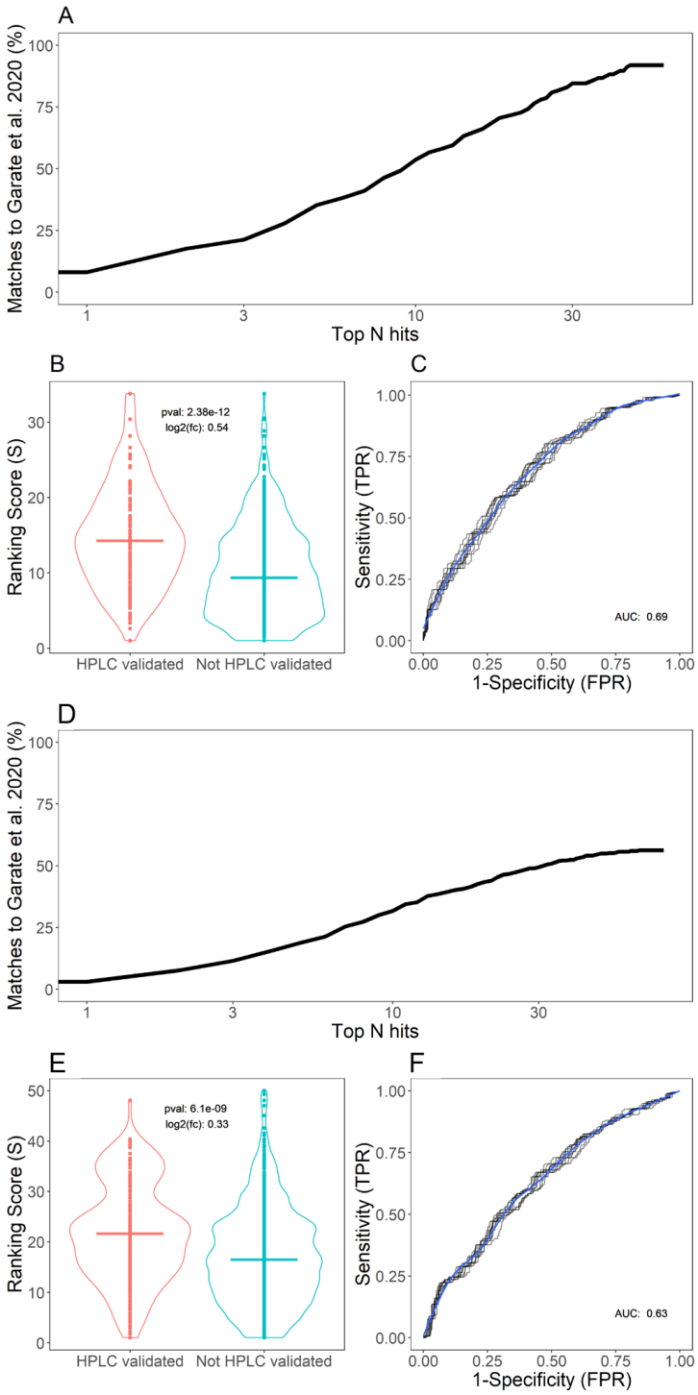


Figure 2. Validation of rMSIfragment against HPLC validated annotations reported by Garate et al. **(A)** Percentage of HPLC matches, **(B)** S score distribution, and **(C)** ROC curve for the negative mode datasets (G9-G15). **(D)** Percentage of HPLC matches, **(E)** S score distribution, and **(F)** ROC curve for the positive mode datasets (G1-G8).

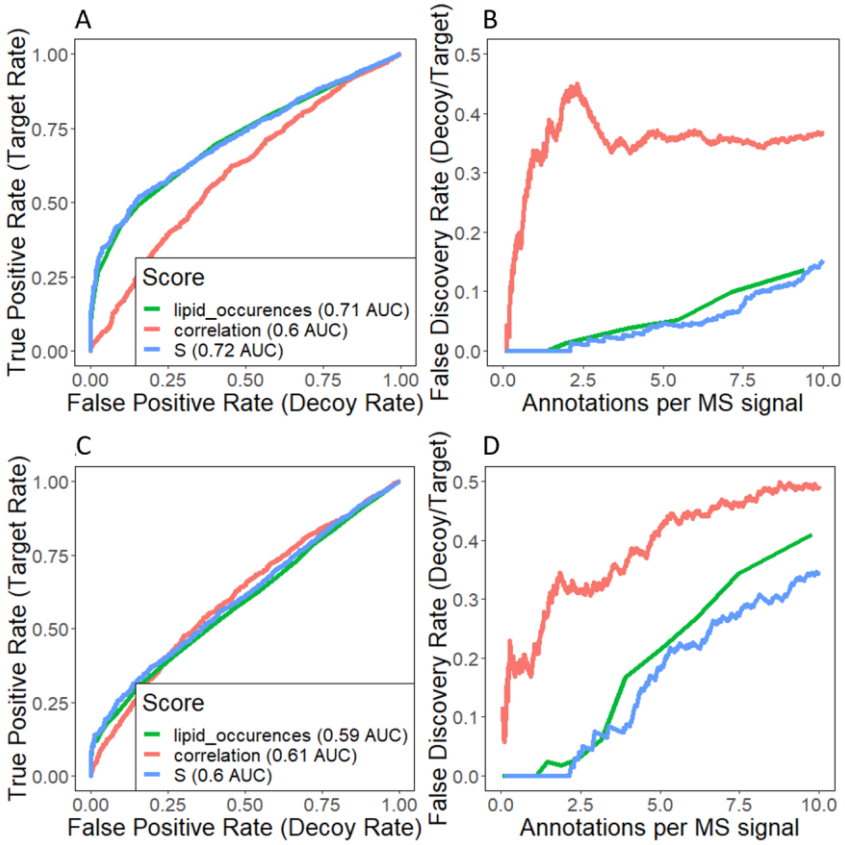


Figure 3. Performance estimation of the ranking scores proposed using a target-decoy validation approach. The decoy database is composed of adducts and fragmentation pathways highly unlikely to be found in the lipid classes considered. Lipid occurrences (LO): number of times a given lipid is found (including parental adducts and in-source fragments); Spatial correlation (C): The weighted mean Pearson's correlation of all m/z features annotated as the same lipid. Final ranking score (S): LO * C. **(A)** ROC curve and **(B)** FDR estimation for the negative mode datasets (G9-G15) **(C)** ROC curve and **(D)** FDR estimation for the positive mode datasets (G1-G8)

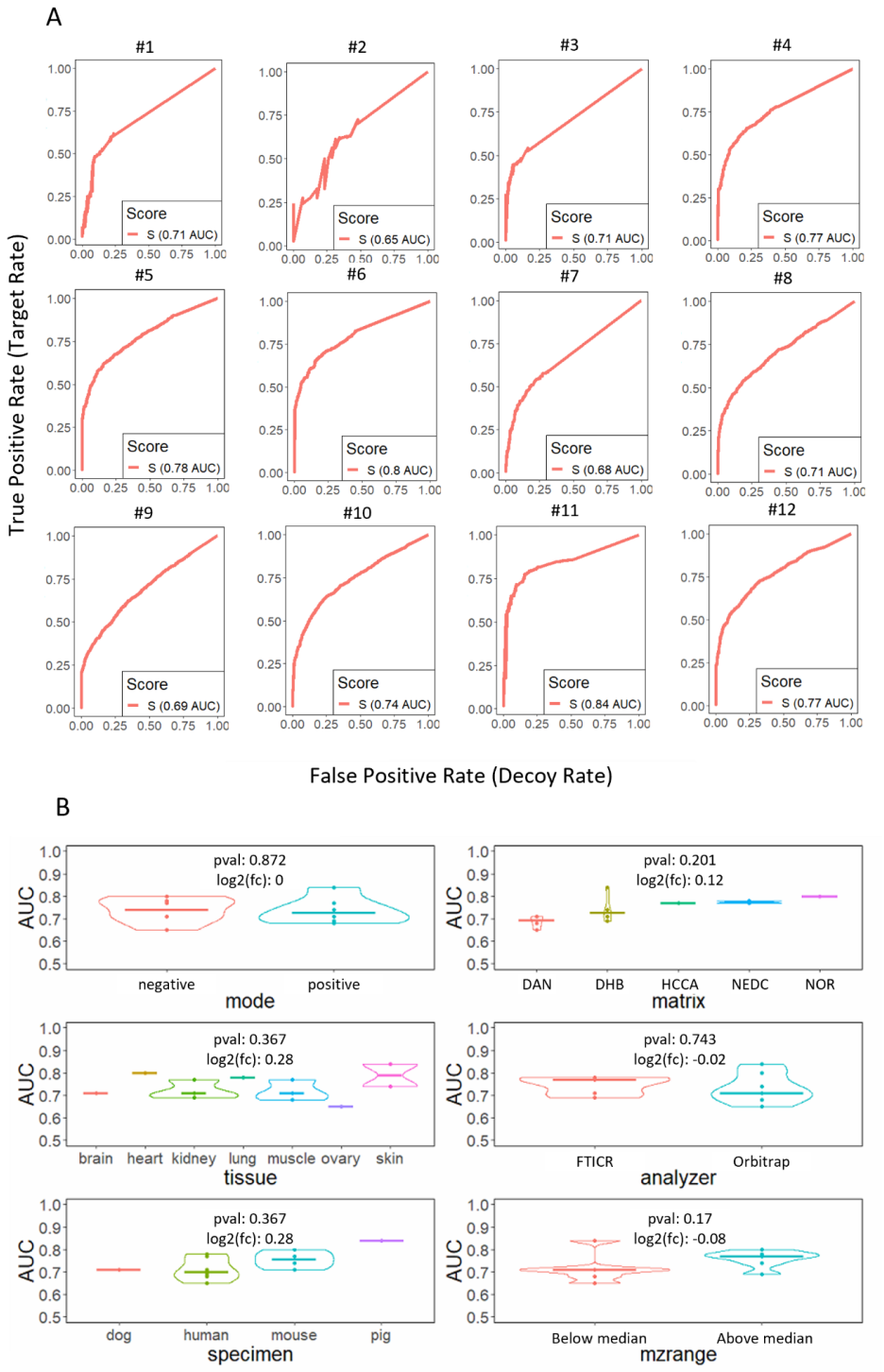


Figure 4. Target and decoy validation on 12 datasets publicly available on METASPACE (Alexandrov et al. 2019). **(A)** ROC curves and Area Under the Curve (AUC) for each dataset M1-M12. **(B)** Statistical significance tests of performance (AUC) differences across different biological (tissue and specimen), sample preparation (matrix), and instrumental (ion mode, analyzer, m/z range) parameters.

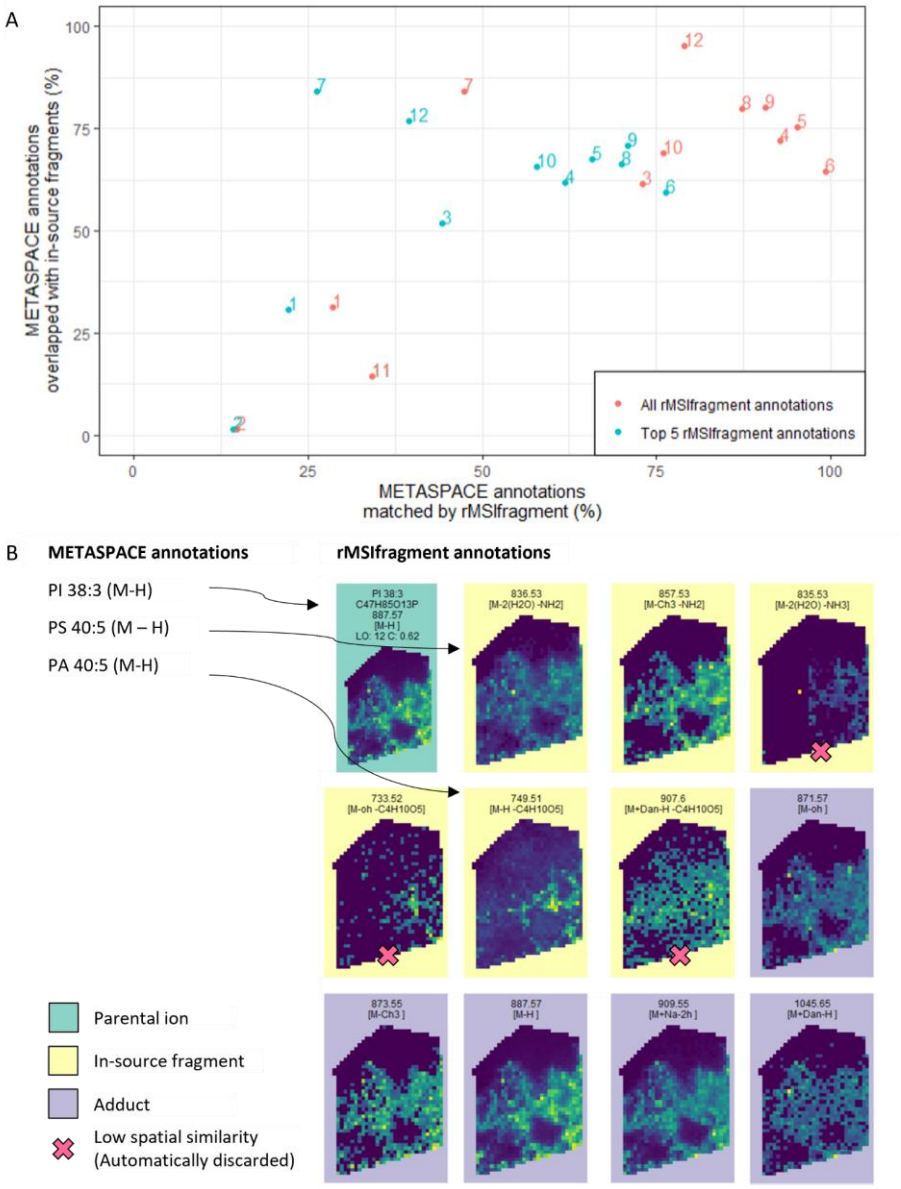


Figure 5. Comparison of annotation results between rMSIfragment and METASPACE **(A)** Bulk comparison using 12 datasets publicly available in METASPACE. The horizontal axis shows the percentage of METASPACE annotations that are matched by rMSIfragment. The vertical axis indicates the percentage of METASPACE annotations that are overlapped with at least one in-source fragment annotated by rMSIfragment. The standard FDR threshold of 0.2 was used for METASPACE annotations. rMSIfragment annotations without any threshold (red) and retaining the top 5 annotations per MS feature (blue). **(B)** Example comparison for a human lung biopsy (Dataset M5) where m/z 887.57 is annotated by both tools as PI 38:3 (M+H) (C47H85O13P).

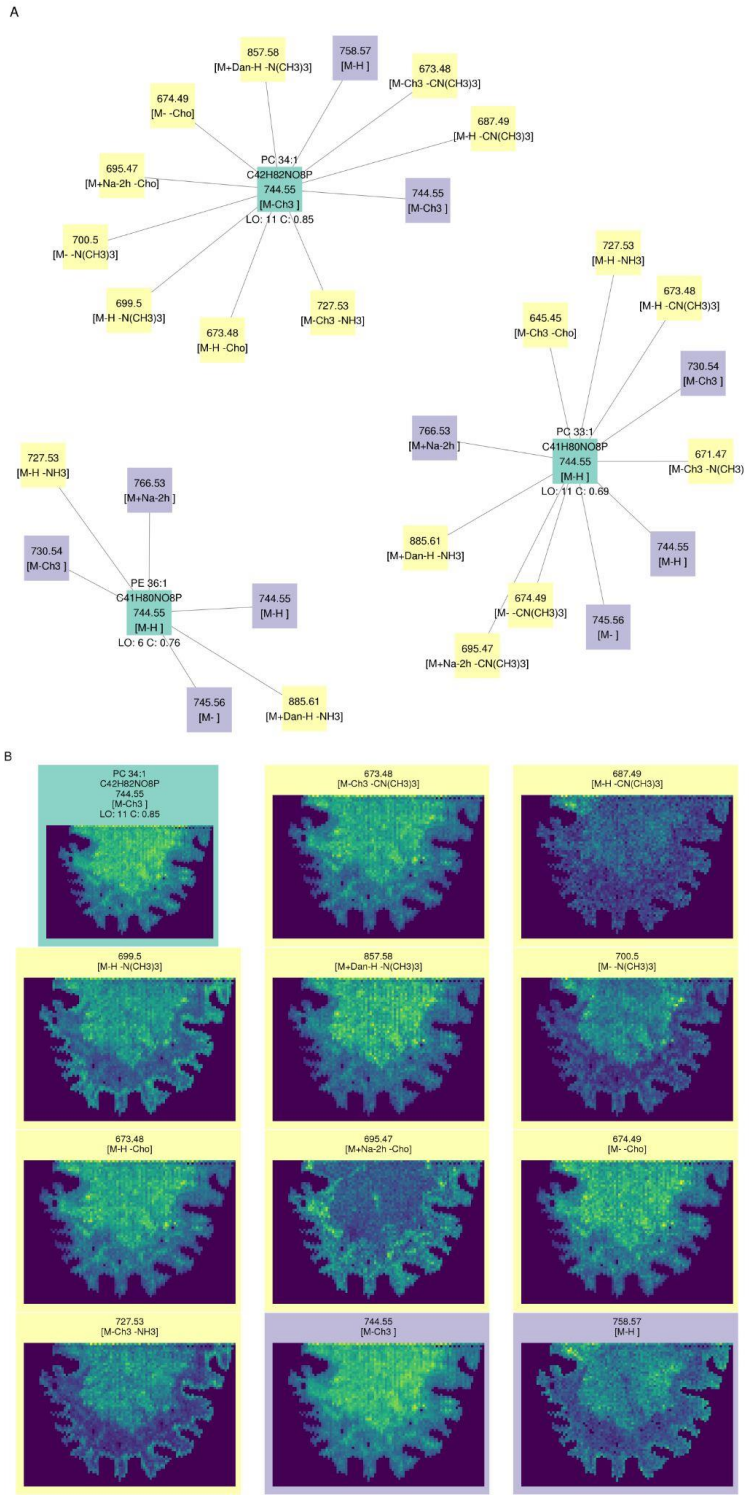


Figure 6. Example application and interpretation of rMSIfragment results on Dataset G9. **(A)** Top 3 annotations for m/z 744.55. Each parental annotation (green) is the center of a network including all adducts (purple) and in-source fragments (yellow) annotated in the sample. **(B)** Spatial representation of the top annotation (PC 34:1 [M-CH3])

7. References

Adusumilli, Ravali, and Parag Mallick. 2017. “Data Conversion with ProteoWizard msConvert.” *Methods in Molecular Biology* 1550: 339–68.

Ait-Belkacem, Rima, Caroline Berenguer, Claude Villard, L ’houcine Ouafik, Dominique Figarella-Branger, Olivier Chinot, and Daniel Lafitte. 2014. “MALDI Imaging and in-Source Decay for Top-down Characterization of Glioblastoma.” *Proteomics* 14 (10): 1290–1301.

Alexandrov, Theodore, Katja Ovchinnikova, Andrew Palmer, Vitaly Kovalev, Artem Tarasov, Lachlan Stuart, Renat Nigmetzianov, Dominik Fay, and Key Metaspace Contributors. 2019. “METASPACE: A Community-Populated Knowledge Base of Spatial Metabolomes in Health and Disease.” *bioRxiv*. https://doi.org/10.1101/539478.

Antone, Angelo J., Qiaoli Liang, Jennifer A. Sherwood, James C. Weiss, Joseph M. Wilson, Sanghamitra Deb, Carolyn J. Cassady, and Yuping Bao. 2019. “Surface Effects of Iron Oxide Nanoparticles on the MALDI In-Source Decay Analysis of Glycans and Peptides.” *ACS Applied Nano Materials* 2 (6): 3999–4008.

Awad, Hanan, Mona M. Khamis, and Anas El-Aneed. 2015. “Mass Spectrometry, Review of the Basics: Ionization.” *Applied Spectroscopy Reviews* 50 (2): 158–75.

Baijnath, Sooraj. 2022. “Mass Spectrometry Imaging: The Future Is Now.” *Bioanalysis* 14 (7): 383–86.

Baquer, Gerard, Lluc Sementé, Toufik Mahamdi, Xavier Correig, Pere Ràfols, and María García-Altares. 2022. “What Are We Imaging? Software Tools and Experimental Strategies for Annotation and Identification of Small Molecules in Mass Spectrometry Imaging.” *Mass Spectrometry Reviews*, July, e21794.

Berghmans, Eline, Kurt Boonen, Evelyne Maes, Inge Mertens, Patrick Pauwels, and Geert Baggerman. 2020. “Implementation of MALDI Mass Spectrometry Imaging in Cancer Proteomics Research: Applications and Challenges.” *Journal of Personalized Medicine* 10 (2). https://doi.org/10.3390/jpm10020054.

Boskamp, Marcel S., and Jens Soltwisch. 2020. “Charge Distribution between Different Classes of Glycerophospolipids in MALDI-MS Imaging.” *Analytical Chemistry* 92 (7): 5222–30.

Burgt, Yuri E. M. van der, David P. A. Kilgour, Yury O. Tsybin, Kristina Srzentić, Luca Fornelli, Alain Beck, Manfred Wuhrer, and Simone Nicolardi. 2019. “Structural Analysis of Monoclonal Antibodies by Ultrahigh Resolution MALDI In-Source Decay FT-ICR Mass Spectrometry.” *Analytical Chemistry* 91 (3): 2079–85.

Coy, Shannon, Shu Wang, Sylwia A. Stopka, Jia-Ren Lin, Clarence Yapp, Cecily C. Ritch, Lisa Salhi, et al. 2022. “Single Cell Spatial Analysis Reveals the Topology of Immunomodulatory Purinergic Signaling in Glioblastoma.” *Nature Communications* 13 (1): 4814.

Debois, Delphine, Virginie Bertrand, Loïc Quinton, Marie-Claire De Pauw-Gillet, and Edwin De Pauw. 2010. “MALDI-in Source Decay Applied to Mass Spectrometry Imaging: A New Tool for Protein Identification.” *Analytical Chemistry* 82 (10): 4036–45.

Denti, Vanna, Maria K. Andersen, Andrew Smith, Anna Mary Bofin, Anna Nordborg, Fulvio Magni, Siver Andreas Moestue, and Marco Giampà. 2021. “Reproducible

Lipid Alterations in Patient-Derived Breast Cancer Xenograft FFPE Tissue Identified with MALDI MSI for Pre-Clinical and Clinical Application." *Metabolites* 11 (9). https://doi.org/10.3390/metabo11090577.

Dührkop, Kai, Markus Fleischauer, Marcus Ludwig, Alexander A. Aksenov, Alexey V. Melnik, Marvin Meusel, Pieter C. Dorrestein, Juho Rousu, and Sebastian Böcker. 2019. "SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information." *Nature Methods* 16 (4): 299–302.

Elias, Joshua E., and Steven P. Gygi. 2007. "Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry." *Nature Methods* 4 (3): 207–14.

Ellis, Shane R., Martin R. L. Paine, Gert B. Eijkel, Josch K. Pauling, Peter Husen, Mark W. Jervelund, Martin Hermansson, Christer S. Ejsing, and Ron M. A. Heeren. 2018. "Automated, Parallel Mass Spectrometry Imaging and Structural Identification of Lipids." *Nature Methods* 15 (7): 515–18.

Franceschi, Pietro, Yonghui Dong, Kerstin Strupat, Urska Vrhovsek, and Fulvio Mattivi. 2012. "Combining Intensity Correlation Analysis and MALDI Imaging to Study the Distribution of Flavonols and Dihydrochalcones in Golden Delicious Apples." *Journal of Experimental Botany* 63 (3): 1123–33.

Fuchs, Beate, Celestina Schober, Grit Richter, Rosmarie Süss, and Jürgen Schiller. 2007. "MALDI-TOF MS of Phosphatidylethanolamines: Different Adducts Cause Different Post Source Decay (PSD) Fragment Ion Spectra." *Journal of Biochemical and Biophysical Methods* 70 (4): 689–92.

Fu, Tingting, David Touboul, Serge Della-Negra, Emeline Houël, Nadine Amusant, Christophe Duplais, Gregory L. Fisher, and Alain Brunelle. 2018. "Tandem Mass Spectrometry Imaging and in Situ Characterization of Bioactive Wood Metabolites in Amazonian Tree Species Sextonia Rubra." *Analytical Chemistry* 90 (12): 7535–43.

Garate, Jone, Sergio Lage, Lucía Martín-Saiz, Arantza Perez-Valle, Begoña Ochoa, M. Dolores Boyano, Roberto Fernández, and José A. Fernández. 2020. "Influence of Lipid Fragmentation in the Data Analysis of Imaging Mass Spectrometry Experiments." *Journal of the American Society for Mass Spectrometry* 31 (3): 517–26.

Guo, G., M. Papanicolaou, N. J. Demarais, Z. Wang, K. L. Schey, P. Timpson, T. R. Cox, and A. C. Grey. 2021. "Automated Annotation and Visualisation of High-Resolution Spatial Proteomic Mass Spectrometry Imaging Data Using HIT-MAP." *Nature Communications* 12 (1): 3241.

Harkin, Carla, Karl W. Smith, C. Logan MacKay, Tara Moore, Simon Brockbank, Mark Ruddock, and Diego F. Cobice. 2022. "Spatial Localization of β-Unsaturated Aldehyde Markers in Murine Diabetic Kidney Tissue by Mass Spectrometry Imaging." *Analytical and Bioanalytical Chemistry* 414 (22): 6657–70.

Hastings, Janna, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. 2016. "ChEBI in 2016: Improved Services and an Expanding Collection of Metabolites." *Nucleic Acids Research* 44 (D1): D1214–19.

Heijs, Bram, Alexander Potthoff, Jens Soltwisch, and Klaus Dreisewerd. 2020. "MALDI-2 for the Enhanced Analysis of N-Linked Glycans by Mass Spectrometry Imaging." *Analytical Chemistry* 92 (20): 13904–11.

He, Quan, Cuirong Sun, Jian Liu, and Yuanjiang Pan. 2021. “MALDI-MSI Analysis of Cancer Drugs: Significance, Advances, and Applications.” *Trends in Analytical Chemistry: TRAC* 136 (March): 116183.

Hu, Changfeng, Wenqing Luo, Jie Xu, and Xianlin Han. 2022. “Recognition and Avoidance of Ion Source-Generated Artifacts in Lipidomics Analysis.” *Mass Spectrometry Reviews* 41 (1): 15–31.

Ikegawa, Masaya, Takashi Nirasawa, Nobuto Kakuda, Tomohiro Miyasaka, Yuki Kuzuhara, Shigeo Murayama, and Yasuo Ihara. 2019. “Visualization of Amyloid β Deposits in the Human Brain with Matrix-Assisted Laser Desorption/Ionization Imaging Mass Spectrometry.” *Journal of Visualized Experiments: JoVE*, no. 145 (March). https://doi.org/10.3791/57645.

Israr, Muhammad Zubair, Dennis Bernieh, Andrea Salzano, Shabana Cassambai, Yoshiyuki Yazaki, and Toru Suzuki. 2020. “Matrix-Assisted Laser Desorption Ionisation (MALDI) Mass Spectrometry (MS): Basics and Clinical Applications.” *Clinical Chemistry and Laboratory Medicine: CCLM / FESCC* 58 (6): 883–96.

Janßen, Charlotte, Tobias Boskamp, Lena Hauberg-Lotte, Jens Behrmann, Sören-Oliver Deininger, Mark Kriegsmann, Katharina Kriegsmann, et al. 2022. “Robust Subtyping of Non-Small Cell Lung Cancer Whole Sections through MALDI Mass Spectrometry Imaging.” *PROTEOMICS--Clinical Applications*, 2100068.

Kaya, Ibrahim, Eva Jennische, Stefan Lange, Ahmet Tarik Baykal, Per Malmberg, and John S. Fletcher. 2020. “Brain Region-Specific Amyloid Plaque-Associated Myelin Lipid Loss, APOE Deposition and Disruption of the Myelin Sheath in Familial Alzheimer’s Disease Mice.” *Journal of Neurochemistry* 154 (1): 84–98.

Kaya, Ibrahim, Henrik Zetterberg, Kaj Blennow, and Jörg Hanrieder. 2018. “Shedding Light on the Molecular Pathology of Amyloid Plaques in Transgenic Alzheimer’s Disease Mice Using Multimodal MALDI Imaging Mass Spectrometry.” *ACS Chemical Neuroscience* 9 (7): 1802–17.

Ma, Xin, and Facundo M. Fernández. 2022. “Advances in Mass Spectrometry Imaging for Spatial Cancer Metabolomics.” *Mass Spectrometry Reviews*, September, e21804.

Miyamoto, Satoshi, Cheng-Chih Hsu, Gregory Hamm, Manjula Darshi, Maggie Diamond-Stanic, Anne-Emilie Declèves, Larkin Slater, et al. 2016. “Mass Spectrometry Imaging Reveals Elevated Glomerular ATP/AMP in Diabetes/obesity and Identifies Sphingomyelin as a Possible Mediator.” *EBioMedicine* 7 (May): 121–34.

Muro, Eleonora, G. Ekin Atilla-Gokcumen, and Ulrike S. Eggert. 2014. “Lipids in Cell Biology: How Can We Understand Them Better?” *Molecular Biology of the Cell* 25 (12): 1819–23.

Nascentes Melo, Luiza Martins, Nicholas P. Lesner, Marie Sabatier, Jessalyn M. Ubellacker, and Alpaslan Tasdogan. 2022. “Emerging Metabolomic Tools to Study Cancer Metastasis.” *Trends in Cancer Research*, July. https://doi.org/10.1016/j.trecan.2022.07.003.

Nguyen, Tra D., Yunpeng Lan, Shelley S. Kane, Jacob J. Haffner, Renmeng Liu, Laura-Isobel McCall, and Zhibo Yang. 2022. “Single-Cell Mass Spectrometry Enables Insight into Heterogeneity in Infectious Disease.” *Analytical Chemistry* 94 (30): 10567–72.

Nicolardi, Simone, Renzo Danuser, Viktoria Dotz, Elena Domínguez-Vega, Ali Al

Kaabi, Michel Beurret, Chakkumkal Anish, and Manfred Wuhrer. 2022. “Glycan and Protein Analysis of Glycoengineered Bacterial E. Coli Vaccines by MALDI-in-Source Decay FT-ICR Mass Spectrometry.” *Analytical Chemistry* 94 (12): 4979–87.

Nishidate, Masanobu, Mitsuhiro Hayashi, Hiroaki Aikawa, Kouji Tanaka, Naoyuki Nakada, Shin-Ichi Miura, Shoraku Ryu, et al. 2019. “Applications of MALDI Mass Spectrometry Imaging for Pharmacokinetic Studies during Drug Development.” *Drug Metabolism and Pharmacokinetics* 34 (4): 209–16.

Notarangelo, Giulia, Jessica B. Spinelli, Elizabeth M. Perez, Gregory J. Baker, Kiran Kurmi, Ilaria Elia, Sylwia A. Stopka, et al. 2022. “Oncometabolite D-2HG Alters T Cell Metabolism to Impair CD8+ T Cell Function.” *Science* 377 (6614): 1519–29.

O’Donnell, Valerie B., Edward A. Dennis, Michael J. O. Wakelam, and Shankar Subramaniam. 2019. “LIPID MAPS: Serving the next Generation of Lipid Researchers with Tools, Resources, Data, and Training.” *Science Signaling* 12 (563). https://doi.org/10.1126/scisignal.aaw2964.

Palmer, Andrew, Prasad Phapale, Ilya Chernyavsky, Regis Lavigne, Dominik Fay, Artem Tarasov, Vitaly Kovalev, et al. 2016. “FDR-Controlled Metabolite Annotation for High-Resolution Imaging Mass Spectrometry.” *Nature Methods* 14 (1): 57–60.

Race, Alan M., Iain B. Styles, and Josephine Bunch. 2012. “Inclusive Sharing of Mass Spectrometry Imaging Data Requires a Converter for All.” *Journal of Proteomics* 75 (16): 5111–12.

Ràfols, Pere, Bram Heijs, Esteban Del Castillo, Oscar Yanes, Liam A. McDonnell, Jesús Brezmes, Iara Pérez-Taboada, Mario Vallejo, María García-Altares, and Xavier Correig. 2020. “RMSIproc: An R Package for Mass Spectrometry Imaging Data Processing.” *Bioinformatics* 36 (11): 3618–19.

Ràfols, Pere, Sònia Torres, Noelia Ramírez, Esteban Del Castillo, Oscar Yanes, Jesús Brezmes, and Xavier Correig. 2017. “RMSI: An R Package for MS Imaging Data Handling and Visualization.” *Bioinformatics* 33 (15): 2427–28.

Ruttkies, Christoph, Emma L. Schymanski, Sebastian Wolf, Juliane Hollender, and Steffen Neumann. 2016. “MetFrag Relaunched: Incorporating Strategies beyond in Silico Fragmentation.” *Journal of Cheminformatics* 8: 3.

Schramm, Thorsten, Alfons Hester, Ivo Klinkert, Jean Pierre Both, Ron M. A. Heeren, Alain Brunelle, Olivier Laprévote, et al. 2012. “ImzML - A Common Data Format for the Flexible Exchange and Processing of Mass Spectrometry Imaging Data.” *Journal of Proteomics* 75 (16): 5106–10.

Sementé, Lluc, Gerard Baquer, María García-Altares, Xavier Correig-Blanchar, and Pere Ràfols. 2021. “rMSIannotation: A Peak Annotation Tool for Mass Spectrometry Imaging Based on the Analysis of Isotopic Intensity Ratios.” *Analytica Chimica Acta* 1171 (August): 338669.

Takeo, Emi, Yuki Sugiura, Tatsuki Uemura, Koshiro Nishimoto, Masanori Yasuda, Eiji Sugiyama, Sumio Ohtsuki, et al. 2019. “Tandem Mass Spectrometry Imaging Reveals Distinct Accumulation Patterns of Steroid Structural Isomers in Human Adrenal Glands.” *Analytical Chemistry* 91 (14): 8918–25.

Tans, Roel, Shoumit Dey, Nidhi Sharma Dey, Grant Calder, Peter O’Toole, Paul M. Kaye, and Ron M. A. Heeren. 2021. “Spatially Resolved Immunometabolism to Understand Infectious Disease Progression.” *Frontiers in Microbiology* 12 (August): 709728.

Tortorella, Sara, Paolo Tiberi, Andrew P. Bowman, Britt S. R. Claes, Klára Ščupáková, Ron M. A. Heeren, Shane R. Ellis, and Gabriele Cruciani. 2020. “LipostarMSI: Comprehensive, Vendor-Neutral Software for Visualization, Data Analysis, and Automated Molecular Identification in Mass Spectrometry Imaging.” *Journal of the American Society for Mass Spectrometry* 31 (1): 155–63.

Wang, Fei, Jaanus Liigand, Siyang Tian, David Arndt, Russell Greiner, and David S. Wishart. 2021. “CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification.” *Analytical Chemistry* 93 (34): 11692–700.

Wang, Hay-Yan J., and Fong-Fu Hsu. 2020. “Revelation of Acyl Double Bond Positions on Fatty Acyl Coenzyme A Esters by MALDI/TOF Mass Spectrometry.” *Journal of the American Society for Mass Spectrometry* 31 (5): 1047–57.

———. 2022. “Structural Characterization of Phospholipids and Sphingolipids by in-Source Fragmentation MALDI/TOF Mass Spectrometry.” *Analytical and Bioanalytical Chemistry* 414 (6): 2089–2102.

Wang, Zhonghua, Wenqing Fu, Meiling Huo, Bingshu He, Yaqi Liu, Lu Tian, Wanfang Li, et al. 2021. “Spatial-Resolved Metabolomics Reveals Tissue-Specific Metabolic Reprogramming in Diabetic Nephropathy by Using Mass Spectrometry Imaging.” *Acta Pharmaceutica Sinica. B* 11 (11): 3665–77.

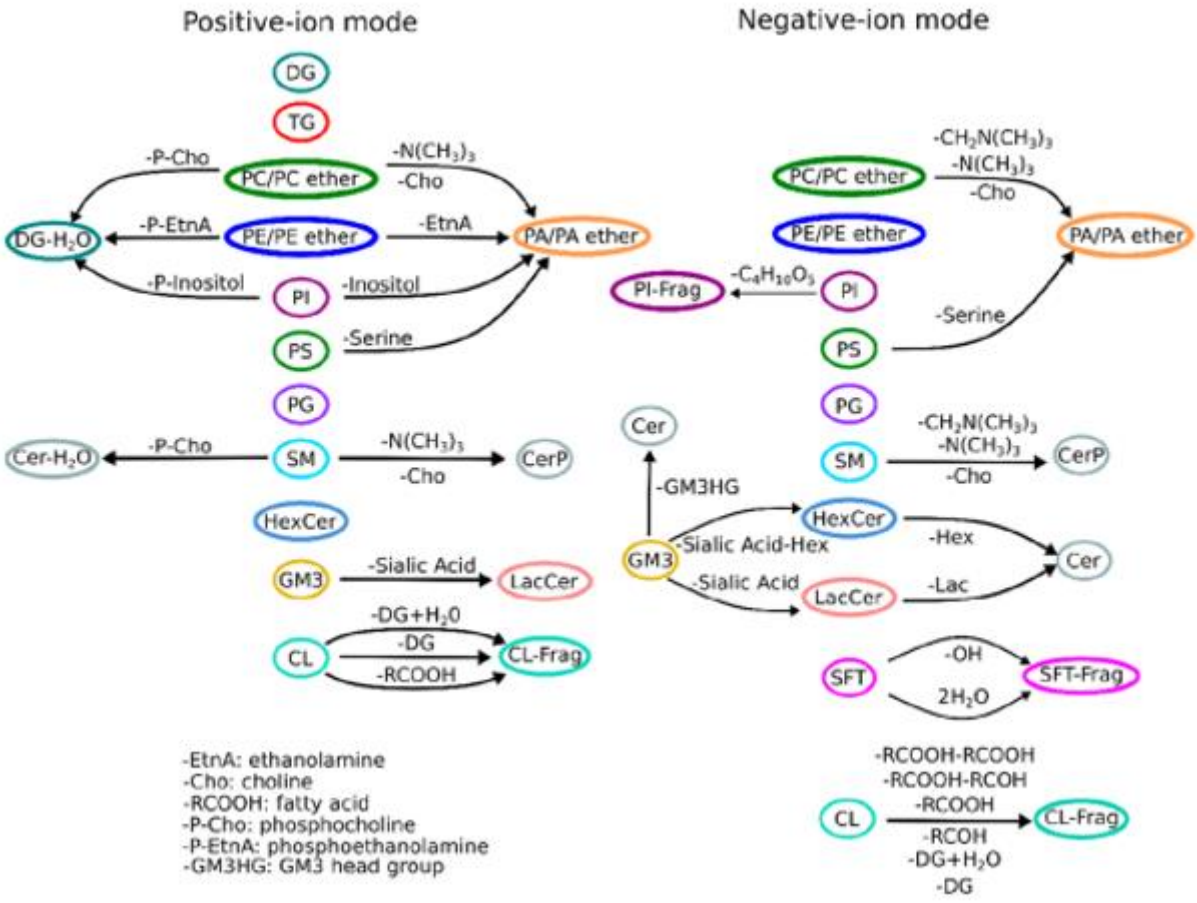
Xue, Jingchuan, Xavier Domingo-Almenara, Carlos Guijas, Amelia Palermo, Markus M. Rinschen, John Isbell, H. Paul Benton, and Gary Siuzdak. 2020. “Enhanced in-Source Fragmentation Annotation Enables Novel Data Independent Acquisition and Autonomous METLIN Molecular Identification.” *Analytical Chemistry* 92 (8): 6051–59.

Zhan, Lingpeng, Xi Huang, Jinjuan Xue, Huihui Liu, Caiqiao Xiong, Jiyun Wang, and Zongxiu Nie. 2021. “MALDI-TOF/TOF Tandem Mass Spectrometry Imaging Reveals Non-Uniform Distribution of Disaccharide Isomers in Plant Tissues.” *Food Chemistry* 338 (February): 127984.

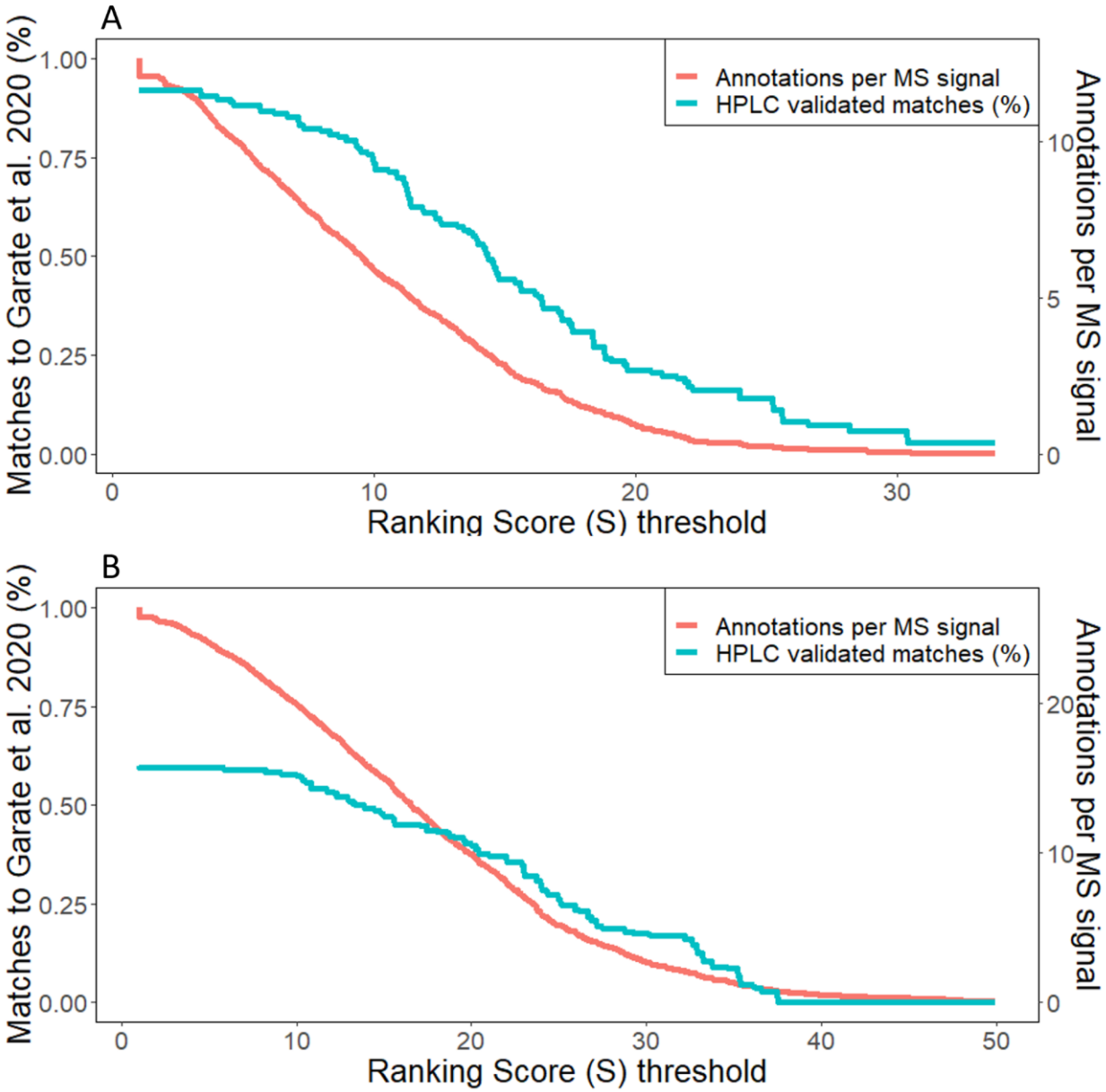
Zhou, Bin, Jun Feng Xiao, Leepika Tuli, and Habtom W. Ressom. 2012. “LC-MS-Based Metabolomics.” *Molecular bioSystems* 8 (2): 470–81.

Zimmerman, Tyler A., Delphine Debois, Gabriel Mazzucchelli, Virginie Bertrand, Marie-Claire De Pauw-Gillet, and Edwin De Pauw. 2011. “An Analytical Pipeline for MALDI in-Source Decay Mass Spectrometry Imaging.” *Analytical Chemistry* 83 (15): 6090–97.

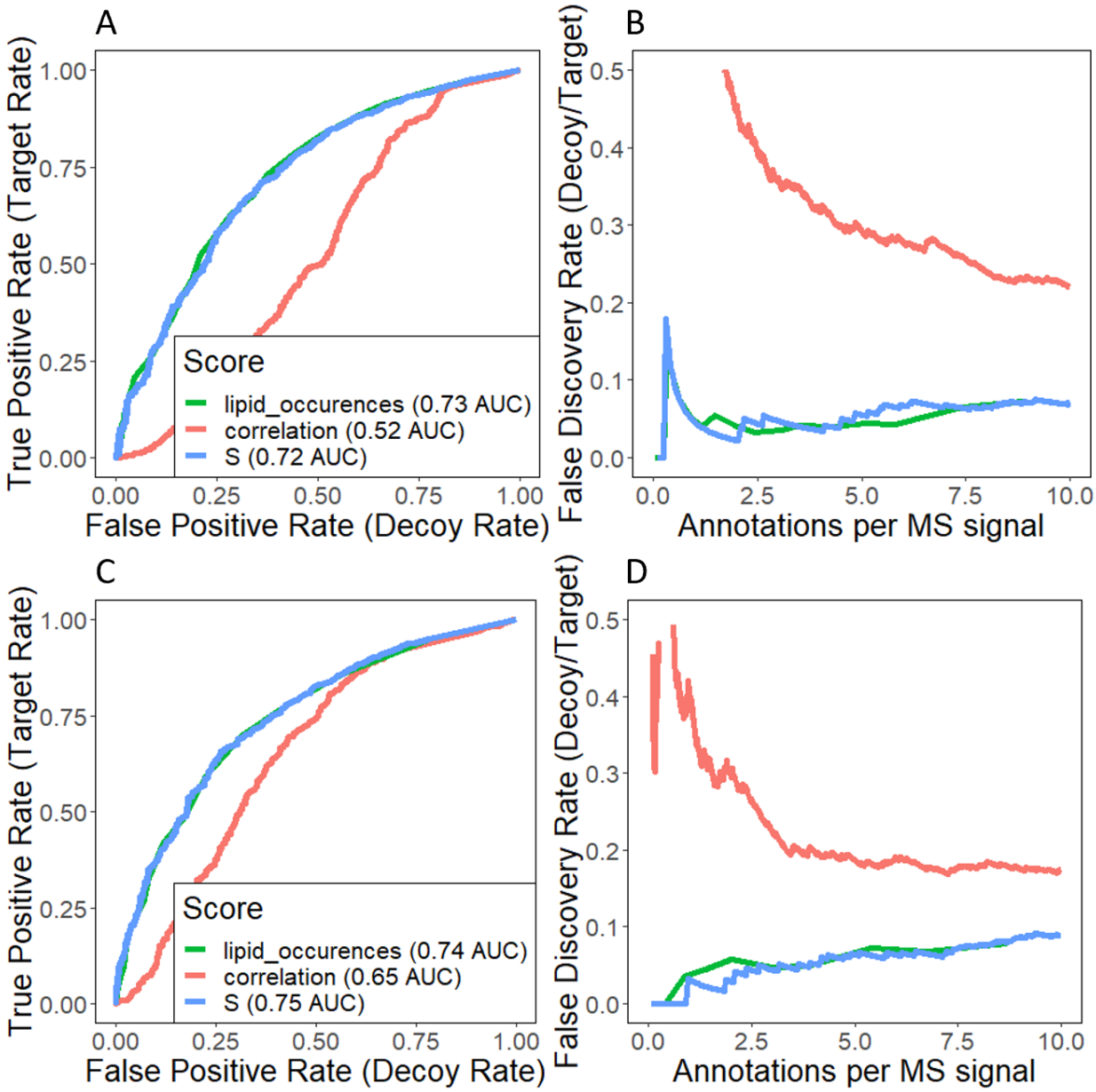
8. Supplementary Materials



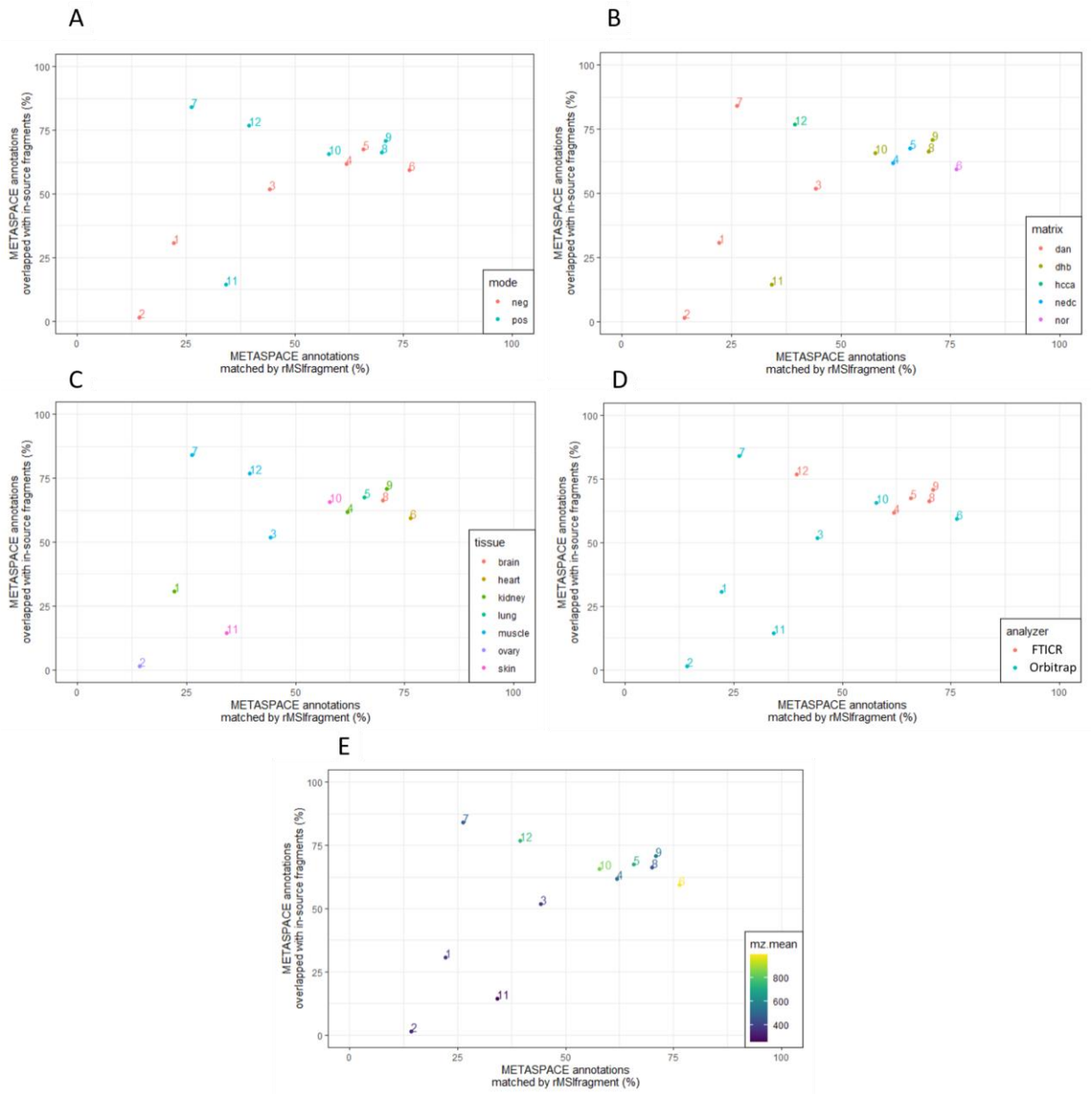
Supplementary Figure 1. Lipid fragmentation pathways. Adapted with the permission of Garate et al. 2020.



Supplementary Figure 2. Automatic annotation with rMSIfragment validated with HPLC in human nevi samples (Garate et al. 2020). Percentage of HPLC validated matches (Garate et al. 2020) against increasing ranking score (S) threshold (blue). **(A)** Samples G9-G15 (negative ion mode). **(B)** Samples G1-G8 (positive ion mode).



Supplementary Figure 3. Performance estimation of the Ranking scores proposed using a Target Decoy Validation approach. The Decoy database is composed of metabolites and lipids unlikely to be found in animal tissues (plants, bacterial, fungus…). (A) ROC and (B) FDR estimation on samples G9-G15 (negative ion mode). (C) ROC and (D) FDR estimation on samples G1-G8 (positive ion mode).



Supplementary Figure 4. METASPACE annotations overlapped with in-source fragments vs METASPACE annotations matched by rMSIfragment color-coded based on: **(A)** Ion mode **(B)** MALDI matrix **(C)** Tissue type **(D)** Analyzer **(E)** Mean *m/z*

Supplementary Table 1. Lipid adduct formation. Adapted with the permission of Garate et al. 2020

adduct	Δmass	DG	TG	PA	PE & PE P	PC & PC P	PG	PI	PS	CL	Cer	CerP	SM	HexCer	SFT	GM3
$[M+Na-COOH]^+$	−21.0006															*
$[M-H_2O+H]^+$	−14.9876	*									**					
$[M + H]^+$	1.0073			*	*	**					*		**			
$[M-H_2O+Na]^+$	6.9943	*									*					
$[M-H_2O+K]^{+b}$	22.9682	*									*					
$[M + Na]^+$	22.9892	***	***	*	**	***	*	*	*	*	***	*	***	***		*
$[M+K]^{+b}$	38.9632	*	*	*	*	*	*	*	*	*	*	*	*	**		*
$[M-H+2Na]^+$	44.9712			***	***		***	***	*	**		***			***	***
$[M-H+Na+K]^{+b}$	60.9451			*	*		*	*	*	*		*				*
$[M-2H+3Na]^{+b}$	66.9531			*					***	***						
$[M-H+2K]^{+b}$	76.9190			*	*		*	*	*	*		*			*	*
$[M-2H+2Na+K]^{+b}$	82.9271			*					*	*						
$[M-2H+Na+2K]^{+b}$	98.9010			*					*	*						
$[M-2H+3K]^{+b}$	114.8749			*					*	*						

[a]***Strongest and **Second strongest adduct for a given lipid class. *Other adducts found for a given lipid class. [b]Potassium adducts were observed in the tissue, but not in pure standards, because no potassium salt was added. Thus, the intensity of the K^+ adduct will depend on the tissue's K^+ concentration.

Supplementary Table 2. List of the 29 MALDI MSI datasets used for validation. Sample type, sample preparation and MALDI-MSI acquisition parameters.

No.	Species	Tissue type	Matrix deposition	Lateral Res. (um)	*m/z* range	Mass spectrometer	Acq. Mode	Notes	Ref.
G1-G8	*Homo sapiens sapiens*	Nevus	MBT, Ace Glass 8023 Glass Sublimator, 10 min	25	480-1100	ThermoFisher™ LTQ-Orbitrap XL	Positive / Profile	7 replicates	(Garate et al. 2020)
G9-G15	*Homo sapiens sapiens*	Nevus	DAN, Ace Glass 8023 Glass Sublimator, 10 min	25	550−1200	ThermoFisher™ LTQ-Orbitrap XL	Negative / Profile	7 replicates	(Garate et al. 2020)
M1	*Canis familiaris*	Kidney	DAN, TM sprayer	Not Specified	200 - 915	Orbitrap	Negative / Centroid		(Alexandrov et al. 2019)
M2	*Homo sapiens sapiens*	Ovary	DAN, TM sprayer	Not Specified	200 - 645	Orbitrap	Negative / Centroid		(Alexandrov et al. 2019)
M3	*Mus musculus*	Muscle	DAN, TM sprayer	Not Specified	200 - 890	Orbitrap	Negative / Centroid		(Alexandrov et al. 2019)
M4	*Homo sapiens sapiens*	Kidney	NEDC, TM sprayer	50	300 - 1500	FTICR	Negative / Centroid		(Alexandrov et al. 2019)
M5	*Homo sapiens sapiens*	Lung	NEDC, HTX M5 sprayer	40	400 - 1500	FTICR	Negative / Centroid		(Alexandrov et al. 2019)

M6	*Mus musculus*	Heart	Norharmane, HTX TM sprayer	30	400 - 1945	Orbitrap	Negative / Centroid		(Alexandrov et al. 2019)
M7	*Homo sapiens sapiens & Mus musculus*	Cervix & Muscle Coculture	DAN, TM sprayer	Not Specified	200 - 1080	Orbitrap	Positive / Centroid		(Alexandrov et al. 2019)
M8	*Homo sapiens sapiens*	Brain	DHA, Spray robot	Not Specified	100 - 1465	FTICR	Positive / Centroid		(Alexandrov et al. 2019)
M9	*Homo sapiens sapiens*	Kidney	DHB, TM sprayer	50	250 - 1300	FTICR	Positive / Centroid		(Alexandrov et al. 2019)
M10	*Mus musculus*	Skin	DHB, Airbrush	60	400 - 1600	Orbitrap	Positive / Centroid		(Alexandrov et al. 2019)
M11	*Sus domesticus*	Skin	DHB, Airbrush	30	70 - 400	Orbitrap	Positive / Centroid		(Alexandrov et al. 2019)
M12	*Mus musculus*	Left upper arm	CHCA, HTX	50	155 - 1970	FTICR	Positive / Centroid		(Alexandrov et al. 2019)
B1-B2	*Mus musculus*	Brain	2,5-DHB, Custom-made sublimation chamber	50	100-1200	Bruker ultrafleXtreme™ MALDI-TOF/TOF	Positive / Profile	2 replicates	-

Supplementary Table 3. Example rMSIfragment output

abbreviation	formula	adduct	fragmentation	experimental_mz	ppm_error	lipid_occurences	correlation
PC 32:1	C40H78NO8P	M-Ch3	-CN(CH3)3	645.4466	4.780602228	8	0.6745115
PC 32:1	C40H78NO8P	M-H	-N(CH3)3	671.4651	0.807263839	8	0.6745115
PC 32:1	C40H78NO8P	M-	-N(CH3)3	672.4694	4.943208786	8	0.6745115
PC 32:1	C40H78NO8P	M-H	-Cho	645.4466	4.712171206	8	0.6745115
PC 32:1	C40H78NO8P	M-Ch3	-NH2	700.5013	4.867592105	8	0.6745115
PC 32:1	C40H78NO8P	M-Ch3	-NH3	699.4970	0.013934849	8	0.6745115
PC 32:1	C40H78NO8P	M-Ch3		716.5210	3.617014969	8	0.6745115
PC 32:1	C40H78NO8P	M-H		730.5418	3.546428453	8	0.6745115
PC 35:1	C43H84NO8P	M-Ch3	-CN(CH3)3	687.4945	3.298120830	8	0.7343465
PC 35:1	C43H84NO8P	M-H	-CN(CH3)3	701.5120	0.850782529	8	0.7343465
PC 35:1	C43H84NO8P	M-Ch3	-N(CH3)3	699.4970	0.012931844	8	0.7343465
PC 35:1	C43H84NO8P	M-Ch3	-Cho	673.4834	2.596329578	8	0.7343465
PC 35:1	C43H84NO8P	M-H	-Cho	687.4945	3.233409416	8	0.7343465
PC 35:1	C43H84NO8P	M-Ch3		758.5689	2.197780959	8	0.7343465
PC 35:1	C43H84NO8P	M-H		772.5848	1.725601783	8	0.7343465
PC 35:1	C43H84NO8P	M+Na-2h		794.5667	1.843665677	8	0.7343465
PS 36:1	C42H80NO10P	M-Ch3	-Serine	687.4945	2.537782689	8	0.6414475
PS 36:1	C42H80NO10P	M-H	-Serine	701.5120	0.139913844	8	0.6414475
PS 36:1	C42H80NO10P	M-	-NH2	773.5306	3.378992080	8	0.6414475
PS 36:1	C42H80NO10P	M-H	-NH3	771.5172	1.224394457	8	0.6414475

CHAPTER 4:

rMSIcleanup: An open-source tool for matrix-related peak annotation in mass spectrometry imaging and its application to silver-assisted laser desorption/ionization

Gerard Baquer,[1] LLuc Sementé,[1] María García-Altares,[1,2 *] Young Jin Lee,[4] Pierre Chaurand,[5] Xavier Correig,[1,2,3] Pere Ràfols[1,2,3]

[1]Department of Electronic Engineering, Rovira i Virgili University, Tarragona, Spain
[2]Spanish Biomedical Research Centre in Diabetes and Associated Metabolic Disorders (CIBERDEM), 28029, Madrid, Spain
[3]Institut d'Investigació Sanitària Pere Virgili, Tarragona, Spain
[4] Department of Chemistry, Iowa State University, Ames, IA 50011, USA
[5] Department of Chemistry, Université de Montréal, Montreal, Quebec, H3C 3J7, Canada

*Correspondance to:
María García-Altares; e-mail: maria.garcia-altares@urv.cat

J Cheminform 12, 45 (2020)

https://doi.org/10.1186/s13321-020-00449-0

Abstract

Mass spectrometry imaging (MSI) has become a mature, widespread analytical technique to perform non-targeted spatial metabolomics. However, the compounds used to promote desorption and ionization of the analyte during acquisition cause spectral interferences in the low mass range that hinder downstream data processing in metabolomics applications. Thus, it is advisable to annotate and remove matrix-related peaks to reduce the number of redundant and non-biologically-relevant variables in the dataset. We have developed rMSIcleanup, an open-source R package to annotate and remove signals from the matrix, according to the matrix chemical composition and the spatial distribution of its ions. To validate the annotation method, rMSIcleanup was challenged with several images acquired using silver-assisted laser desorption ionization MSI (AgLDI MSI). The algorithm was able to correctly classify *m/z* signals related to silver clusters. Visual exploration of the data using Principal Component Analysis (PCA) demonstrated that annotation and removal of matrix-related signals improved spectral data post-processing. The results highlight the need for including matrix-related peak annotation tools such as rMSIcleanup in MSI workflows.

Keywords: mass spectrometry imaging; spatial metabolomics; matrix annotation; overlapping-signal detection; silver-assisted laser/desorption ionization; spectral processing

1. Introduction

Mass spectrometry imaging (MSI) is a label-free technology that allows to obtain molecular and spatial information from intact tissue sections [1]. MSI has been gradually adopted for spatial-resolved metabolomics and it has been regarded as a potential tool for understanding the mechanisms underlying complex diseases such as cancer or diabetes [2]. However, the conventional organic matrices used in Matrix-Assisted Laser Desorption Ionization (MALDI) produce spectral signals that interfere in the low *m/z* range. This is an issue particularly in metabolomics which analyses low molecular weight compounds, so mass spectrometers are set to acquire within the *m/z* range where MALDI matrices exhibit most MS signals. This seriously hampers downstream metabolomics data processing [3, 4], as the matrix introduces noise, redundant variables, and variables with no biological meaning into the complex MSI datasets.

Several alternatives to the common organic matrices have been proposed to deal with exogenous contamination caused by matrix ion signals. Nanomaterials or metal layer deposition methods, for instance, dramatically reduce the number of signals related to the LDI promoting material in the low *m/z* range. Some examples are graphene oxide, silicon or metals such as gold, platinum or silver [5–8]. Nevertheless, even when these alternatives are used and the number of peaks related to the LDI promoting material is reduced, there is still a need to annotate them in order to reduce spectral complexity and distinguish exogenous from endogenous compounds, especially in untargeted applications.

To tackle the issue of annotating MS signals related to the LDI-promoting material several software-based solutions have been proposed. A simple approach consists of

acquiring a reference area outside the sample during the MSI experiment. Under the assumption that only matrix-related peaks will be recorded, the peaks found in the outside area are then subtracted from the tissue spectrum. Given its simplicity, some variation of this procedure has been adopted by many researchers in their workflows. Expanding on this idea, Fonville et al. [9] presented a method that relies on the hypothesis that matrix-related peaks will correlate positively to a set of reference peaks outside the tissue region while endogenous peaks will correlate negatively. However, this approach has three main limitations. Firstly, due to ion suppression [10] and the formation of matrix adducts with endogenous compounds, the matrix-related peaks outside and inside the tissue region might differ. Additionally, endogenous molecules that are delocalized during the matrix application process can be misclassified as matrix-related. Finally, the method cannot distinguish a given matrix-related MS peak from an isobaric or overlapping endogenous MS peak. Thus, simplified approaches to annotate matrix-related signals are not suitable for untargeted applications such as spatial metabolomics. Recent work by Ovchinnikova et al. [11] takes a more comprehensive approach in defining three automated algorithms for off-sample ion classification. Their methods have proved to perform well when trained and validated against a “gold standard set” of ion images manually annotated by experts. However, their focus is not specifically on matrix-related peaks, but on the annotation of signals that exhibit a spatial distribution with high concentrations outside of the tissue region. For this reason, these methods focus on classifying each ion image separately as “on-sample” or “off-sample” and do not exploit relevant information such as the identity of the ion, adduct type, matrix type, etc. Additionally, since they are based in machine and deep learning methods they inherently suffer from the black box problem given that annotation results cannot be traced back and easily justified.

To solve these limitations we propose a new algorithm that relies not only on the ion images but also on the chemical information of the LDI promoting material used. The algorithm also incorporates an overlapping peak detection feature to prevent misclassification of overlapped or isobaric ions. The presented algorithm is implemented in an open-source R package freely available to facilitate its use. Additionally, the package generates a visual report to transparently justify each annotation.

In order to validate and optimize the proposed method, we opted for a well-understood LDI promoting material such as silver. The use of silver nanolayers for MSI (AgLDI MSI) has been steadily growing in recent years [6, 12–17]. The characteristic isotopic pattern of silver (^{107}Ag and ^{109}Ag, 51.84% and 48.16% abundance, respectively), as well as its well-known ionization and adduct formation allow to define a list of possible and not-possible silver-related peaks of a typical AgLDI MSI experiment. This set of possible and not-possible peaks is used as a validation list to assess the performance of the classification algorithm. A total of 14 MSI datasets acquired with an Ag-sputtered nanolayer from three different laboratories, were used for validation.

2. Materials & Methods

Table 1 summarizes the main processing parameters for each of the 14 datasets used in this study. Datasets 1-10 were acquired in our lab and the materials, sample preparation and MSI acquisition parameters are described here. In order to overcome lab-specific bias in our study, four additional datasets were provided by collaborating laboratories. For further details about the materials, sample preparation and MSI acquisition of these datasets, refer to the original publications of Dataset 11 [18], Dataset 12 [14] and Datasets 13 and 14 [6].

2.1. Materials

For the samples acquired by our group, indium tin oxide (ITO)-coated glass slides were obtained from Bruker Daltonics (Bremen, Germany). The silver-target (purity grade > 99.99%) used for sputtering was acquired from Kurt J. Lesker Company (Hastings, England).

2.2. Sample preparation

All the samples acquired by our group were obtained from mice and provided by the animal facility at the Faculty of Medicine and Health Sciences of the University Rovira i Virgili. All tissues were snap-frozen at -80ºC after collection and kept at this temperature during shipping and storing until MSI acquisition.

The tissues were sectioned with a Leica CM-1950 cryostat (Leica Biosystems Nussloch GmbH) located at the Centre for Omics Sciences (COS) of the University Rovira i Virgili into 10 μm sections. Tissue sections were mounted on ITO coated slides by directly placing the glass slide at ambient temperature onto the section.

The sputtering system ATC Orion 8-HV (AJA International, N. Scituate, MA, USA) was used to deposit a silver nanolayer onto each tissue section. An argon atmosphere with a pressure of 30 mTorr was used to create the plasma in the gun. The working distance of the plate was set to 35 mm. The sputtering conditions were ambient temperature using DC mode at 100W for 10s. With these parameters, an Ag layer thickness of roughly 5nm was obtained. The deposition times were short to prevent the substrate temperature from increasing excessively and, consequently, degrading metabolites.

2.3. LDI-MS acquisition

A MALDI TOF/TOF ultrafleXtreme instrument with SmartBeam II Nd:YAG/355 nm laser from Bruker Daltonics available at COS was used for MSI acquisition. Acquisitions were carried out by operating the laser at 2 kHz and collecting a total of 500 shots per pixel.

The TOF spectrometer was operated in positive ion, reflectron mode, in *m/z* ranges according to Table 1. The spectrometer was calibrated prior to MSI data acquisition using $[Ag]_n^+$ cluster peaks as internal reference masses.

2.4. MSI data processing

The raw spectral data of each MSI dataset was exported to the imzML data format [19] in profile mode. The software rMSIproc [20] was used to process the data and generate a peak matrix in centroid mode. The default processing parameters were used. The Signal-to-Noise Ratio (SNR) threshold was set to 5 and the Savitzky-Golay smoothing had a kernel size of 7. Peaks appearing in less than 5% of the pixels were filtered out. Peaks within a window of 6 data-points or scans were binned together as the same mass peak. Mass spectra were re-calibrated using the Ag reference peaks as reference masses [21].

Datasets 13 and 14 were acquired in centroid mode with an Orbitrap mass spectrometer. These datasets were directly submitted to the binning process of rMSIproc [21] to conform to the peak matrix format.

No data normalization was performed. Data were visualized and explored using rMSI [22].

3. Algorithm description

3.1. Input and output format

The matrix-related annotation algorithm takes the peak matrix in centroid mode and the processed spectral data in profile mode as input. The user must also provide the chemical formulae of the matrix applied and a list of possible adducts and neutral losses to consider. The choice of adducts and neutral losses to consider is purely application dependent and is therefore left to the user.

The algorithm produces a vector containing the similarity scores that indicate the likelihood of each mass in the input image being a matrix-related ion. The package also provides an informative visual report for the user to understand the justification behind the classification. Supplementary Figures S1-S4 show examples of the visual report.

3.2. In-silico cluster & adduct calculation

The theoretical mass and relative isotopic pattern intensities of all possible matrix-related silver clusters ("Ag_n^+ theoretical cluster" in this work) are calculated using the open-source package enviPat [23], a fast and memory-efficient algorithm to compute theoretical isotope patterns.

For each Ag_n^+ theoretical cluster t_i its experimental counterpart e_i is obtained from the mean spectra of the dataset. The experimental masses closest to the theoretical ones within a given tolerance specified by the user are used. The Ag_n^+ theoretical clusters will then be matched against their experimental counterparts and their presence in the experimental dataset assessed using two similarity metrics. In the event of finding more than one experimental match within the tolerance, all possible options will be evaluated. The experimental cluster e_i with the highest similarity metrics is selected. From our experience, this is an unlikely event mostly associated with an exceedingly high tolerance threshold.

3.3. Similarity metrics

The similarity between each theoretical matrix-related cluster and experimental clusters is assessed using two similarity scores according to equation 1.

$$S = S_1 \cdot S_2 \tag{1}$$

where S is the total similarity score, S_1 is the cluster spectral similarity (i.e. similarity between the experimental and theoretical isotopic intensity patterns) and S_2 is the intra-cluster morphological similarity (i.e. similarity between the spatial distribution of the experimental ions). Both similarity scores range from 0 to 1.

The cluster spectral similarity score $S_{1,i}$ for theoretical cluster t_i determines the degree of similarity between the scaled intensity vectors of intensities I_{t_i} and I_{e_i} and it is computed according to equation 2.

$$S_{1,i} = e^{-dist\left(\frac{I_{t_i}}{(I_{t_i})}, \frac{I_{e_i}}{(I_{e_i})}\right)} \tag{2}$$

where $dist(a,b)$ is the distance function chosen by the user (Euclidean distance by default), I_{t_i} is the vector of intensities of the theoretical cluster t_i and I_{e_i} is the vector of intensities of experimental cluster e_i. Experimental cluster e_i is determined by accessing the element in the peak matrix with a mass corresponding to t_i within a given tolerance. In plain terms, S_1 is a decaying exponential function of the distance between the intensity scaled intensity vectors I_{t_i} and I_{e_i} .

The intra-cluster morphological similarity $S_{2,i}$ returns the degree of similarity between the spatial distributions of the ions conforming the experimental cluster e_i. Ions with a high spatial correlation are more likely to belong to the same cluster. This metric is computed using equation 3.

$$S_{2,i} = \frac{I_{t_i}^{'} \cdot I_{t_i} \cdot correl(Images_{e_i})}{(\sum I_{t_i})^2} \tag{3}$$

where I_{t_i} is the intensity vector of the theoretical cluster t_i, $correl(A)$ is the correlation function specified by the user (Pearson correlation by default) and $Images_{e_i}$ is the set of images corresponding to each ion in the experimental cluster e_i. In plain terms, S_2 is the weighted mean across both directions of the correlation matrix between each ion image in e_i.

3.4. Overlapping peak detection

Insufficient resolving power leads to overlapped MS signals, which can be a severe problem in matrix-related peak annotation as they can lead to a greater number of misclassified peaks. This is a particularly limiting issue in lower resolution spectrometers such as some TOFs in contrast to higher resolution analysers such as Orbitrap or FTICR [24]. An additional problem with the same effect is the intrinsic inability of mass spectrometry to distinguish between isobaric species. In order to cope with these issues, we propose an overlapping detection algorithm capable of determining if a given MS signal corresponds to more than one overlapped ion peaks.

The overlapping detection algorithm is only executed in those clusters that report S1 and S2 scores under a threshold specified by the user. Before concluding that the cluster is not present, the algorithm determines whether the low similarity metrics could be attributed to the presence of overlapped signals.

The algorithm is based on the operating principle of bisecting k-means [25]. All the ions in an experimental cluster e_i are split into two subgroups ($e_{i:1}$ and $e_{i:2}$) based on the correlation of their spatial distributions using k-means. For each subgroup of ions the similarity metrics S1 and S2 are recomputed. If the S1 and S2 scores of a given subgroup surpass the specified threshold, all ions in the subgroup are tagged as matrix-related. The remaining ions in e_i are tagged as matrix-related but suffering from overlapping, and the overlapping detection algorithm terminates. If instead, none of the subgroups obtains an S1 and S2 above the threshold, the process of splitting into two subgroups by k-means and recomputing the similarity scores is repeated for both $e_{i:1}$ and $e_{i:2}$. This bisection of the ions in e_i is repeated iteratively until a subgroup obtains S1 and S2 scores above the threshold. To prevent overfitting, the iterative process will also stop when the number of peaks contained by the biggest subgroup becomes smaller than half the amount of peaks in e_i. In such event, it is concluded that there are no overlapped peaks and all ions in the experimental cluster e_i are tagged as not-matrix-related. To sum, overlapped MS signals will be detected and distinguished from the rest of the ions in the cluster based on the dissimilarity of their spatial distributions.

4. Results

4.1. Algorithm validation with AgLDI MSI

In order to validate and optimize the algorithm, we opted to use sample tissues covered by silver nanoparticles, a well-defined and understood LDI promoting material. A total of 14 datasets, from 3 different laboratories, were used. The datasets included several animal tissues, plant tissues and human fingermarks.

The algorithm was challenged with the task of classifying a list of silver-containing compounds and adducts for each dataset. The list includes a “positive class” formed by clusters that should be present in all samples used in this study and a “negative class” containing clusters that should not be present in any of them. This list is referred to as “validation list” and allowed us to assess the performance of the algorithm. An algorithm with a perfect performance should classify all clusters in the “positive class” as matrix-related signals and all clusters in the negative class as not present and thus not-matrix related. This is a common approach in bioinformatics for validating and assessing the performance of a classifier algorithm [26]. Table 2 shows the complete validation list.

Silver clusters containing up to 60 atoms have been reported to form during silver sputtering [27]. The “positive class” expected to be found in all datasets is therefore formed by all silver clusters within the acquired mass range. For most of the datasets, this includes clusters from Ag_1^+ to Ag_{10}^+. Given the high heterogeneity in adduct formation of the samples used (i.e. the possibility of biological compounds from the tissue to form adducts with silver cations), no silver adducts were included in the “positive class”.

The “negative class” consists of silver compounds or adducts that should not be present in any of the samples used in this study. Firstly, this list includes various silver neutral salts which cannot be measured using LDI MSI, and some synthetic compounds that are not expected to be present in animal or plant samples [28]. It also includes compounds found in aerial parts of plants, wax and insects (not found in mammal tissues nor in corn root) that have been reported to form adducts with silver in AgLDI MSI applications [29]. For each of these molecules, we also included all clusters within the acquired mass range. These particular molecules and their clusters were selected in an attempt to have a “negative class” covering the full mass range.

4.2. Performance of similarity scores

Using the validation list described in section 4.1, we assessed the performance of the similarity scores as a classifier to annotate Ag_n^+-related peaks in AgLDI MSI datasets.

Figure 1 shows the similarity scores obtained for each cluster in Table 2 when searched in all 14 datasets from Table 1. The blue points represent the “positive class” (clusters that should be present) while the red points represent the negative class (clusters that should not be present). The tolerance threshold was set to 4 data-points or scans.

Figure 1A represents the spectral similarity score ($S1$) against the intra-cluster similarity score (S2) of each of these clusters. The “positive class” is clearly separated on the top right corner (high $S1$ and high $S2$).

To evaluate the classifying performance of the two similarity metrics we use the Precision vs. Recall (PR) curve [26]. The precision is defined as the ratio between the number of clusters in the “positive class” classified as matrix-related (i.e. true positives) and the total number of clusters classified as matrix-related (i.e. true positives + false positives). The recall, on the other hand, is the ratio between the number of clusters in the “positive class” classified as matrix-related (i.e. true positives) and the total number of clusters in the “positive class” (i.e. true positives + false negatives). Figure 1B shows the PR curves for each of the similarity metrics proposed. The areas under the curve (AUC) of 0.97 and 0.91, respectively, show that the spectral similarity score S1 is the best classifier followed by the intra-cluster morphology similarity score $S2$. The product of $S1 \cdot S2$ had the same classifying skill as $S1$ with an AUC of 0.97. These results prove that Ag_n^+-related peaks can be well classified by these two metrics.

$S1$ performs much better than $S2$ as a classifier, and the product of $S1 \cdot S2$ matches but does not improve the performance of $S1$ alone. Nevertheless, we still decided to use the product of $S1 \cdot S2$ as a classifier in rMSIcleanup instead of using $S1$ alone due to three main reasons. Firstly, the overlapping detection algorithm strongly relies on the morphological similarity of ions and thus depends on $S2$. Moreover, even though we did not find a single instance of a cluster with a high $S1$ score and a low $S2$ score (matching isotopic patterns but unmatching spatial distributions) in any of the samples, we still consider that $S2$ should be present to allow for correct classification should this occur. Finally, $S2$ can be a strong asset in applications other than AgLDI MSI where, due to less distinctive isotopic ratios, the performance of $S1$ as a classifier is diminished.

Figure 1C shows the similarity score S1·S2 obtained by each cluster in all datasets. Clusters are arranged in decreasing order of mean similarity score. Supplementary Table S1 maps the cluster numbers to cluster chemical formula. A clear gap between an S of 0.5 and 0.7 separates the "positive class" from the negative one.

Only three false positives (i.e. clusters that should not be present but have a high S value) were reported for adduct $[C_{28}H_{58}O + Ag]^+$. An example is shown for Dataset 4 in Supplementary Figure S5. Identification by MS/MS is required to assess if the compound is indeed present in the sample. Nevertheless, the mass error between experimental and theoretical isotopic patterns for this compound was 154 ppm, an error much higher than the expected for this dataset (acquired with a TOF MS analyzer). Therefore, we inferred that the experimental pattern detected is not related to adduct $[C_{28}H_{58}O + Ag]^+$ and this is, in fact, a false positive. In order to reduce the number of false positives, the mass tolerance of the algorithm can be decreased, however, a too strict mass tolerance increases the number of false negatives.

A total of six false negatives (i.e. clusters that should be present but have a low S value) were reported for some datasets for clusters Ag_3, Ag_6 and Ag_{10}. False negatives correspond to clusters for which the majority of peaks in their isotopic pattern were under the SNR threshold, and thus were excluded during pre-processing. In these cases, the few included peaks were not sufficient to reliably annotate the cluster. Supplementary Figure S6 shows the only exception, the Ag_6 cluster in Dataset 12, whose misclassification is not due to intensity problems. In this case, the fingerprint analysed showed highly homogeneous ion images, which impedes the proper operation of the overlapping algorithm and leads to misclassification. Representative examples of correct annotations are shown in Supplementary Figures S7-S8.

As an additional validation, the results were matched against the published annotations of the datasets provided by external laboratories. Dataset 12 contains 60 identifications by MS/MS [14]. Dataset 13 contains 4 metabolites identified by MS/MS and a total of 10 tentatively identified formulae based on exact mass [6]. Dataset 14 contains 10 metabolites identified by MS/MS and 6 tentatively identified formulae based on exact mass [6]. None of these endogenous signals was misclassified as Ag_n^+-related by our algorithm.

4.3. Overlapping peak detection performance

Figure 2 shows a case example where the overlapping peak detection algorithm successfully identified overlapping ions when searching for the Ag_6 cluster in Dataset 1. Figure 2A depicts the experimental mean profile spectrum in the mass range of interest along with the calculated profile of the Ag_6 cluster. While most peaks follow the calculated isotopic distribution, experimental peaks at *m/z* 641.43, *m/z* 643.43 and *m/z* 653.43 are considerably more intense than in the predicted pattern. This generates a mismatch between the experimental and calculated peaks that leads to a low $S1$ score. Figure 2B shows the spatial distributions of each of the ions in the Ag_6 cluster. The correlation map in Figure 2D clearly indicates that peaks at *m/z* 641.43 and *m/z* 643.43 have a spatial distribution that is unlike that of the rest of the ions in the cluster. The peak at *m/z* 653.43 also shows a considerably different spatial correlation to the rest. These low correlations lead to a lower $S2$ score. Figure 2C is a zoom-in of the peaks at *m/z*

641.43 and *m/z* 643.43 showing that the silver ion peaks are clearly overlapped with Ag-unrelated signals.

Initially, given the low S1 and S2 scores, all peaks in the Ag_6 cluster were misclassified as not Ag_n^+-related. Using the overlapping detection algorithm, the peaks at *m/z* 645.43, *m/z* 647.43, *m/z* 649.43 and *m/z* 651.43 were correctly tagged as belonging to Ag_6. Peaks at *m/z* 641.43, *m/z* 643.43 and *m/z* 653.43 were tagged as related to Ag_6 but with overlapping.

Supplementary Figure S9 explores the effects of overlapping peak detection on overall performance. Two main differences can be appreciated. Firstly, there is an overall increase in the $S_1 \cdot S_2$ score obtained by the "positive class" which leads to a bigger gap between the "positive class" and the "negative class" making the thresholding classification more robust. This is due to the identification of some overlapping peaks in the Ag_n^+ clusters. Additionally, there is a clear improvement in the scores obtained by the Ag_6 cluster. The Ag_6 cluster suffers from overlapping in most of the datasets and is, therefore, the cluster most benefitted from the overlapping detection algorithm. It is also important to note that the overlapping peak detection algorithm does not add any false positives as the $S_1 \cdot S_2$ remains unchanged for the "negative class". This proves that overlapping detection leads to less misclassification of Ag_n^+-related peaks.

4.4. Matrix-related peak annotation improves the post-processing

In order to explore the influence of the annotation and removal of matrix-related peaks in the post-processing workflows, we carried out a multivariate statistical exploratory analysis. The widely used linear algorithm Principal Component Analysis (PCA) [30] was performed on all 14 datasets before and after removal of the Ag_n^+ peaks. Given that the features in an MSI experiment have a direct physical relationship [33, 34], prior to PCA the data was centred and no scaling was performed. We then compared the quality of the spatial representation of the first three principal components. Given the lack of a standard quantitative metric to compare the quality of two images in MSI, we followed the trend established by recent work [11, 31, 32] and performed a qualitative visual comparison.

Figure 3 shows the results of this exploratory analysis on Dataset 2 and Dataset 11. In the pancreatic tissue represented in Figure 3A (Dataset 2), PC1 did not change significantly after matrix removal, while PC2 and PC3 showed a wider variety of morphologies on the tissue after the Ag_n^+ interference was removed. In the brain tissue shown in Figure 3B (Dataset 11) the contrast enhancement is even clearer in the three PCs. Before the Ag_n^+ peaks were removed, PC1 and PC3 did not capture any substantial morphology but afterwards, they did and PC2, which already showed morphological information, did so with increased contrast. To convey the three principal components in a single picture we encoded each of them as a colour in the Red Green Blue colour model (RGB). The RGB picture became richer and more informative after the Ag_n^+ peaks were removed. Similar results were obtained in the remaining 12 datasets and their corresponding images can be accessed in Supplementary Figures S10-S13.

As shown in Supplementary Table S2 and Supplementary Figure S14 the removed Ag_n^+ peaks can represent a substantial fraction of the total number of features. Figure S15 shows the same PCA analysis compensating for the lower number of features after removing the Ag_n^+ peaks by reducing the original dataset to the most intense peaks or by random feature selection. The results still show a clear improvement of the morphological contrast after removal of the Ag_n^+ peaks regardless of the feature reduction in the original dataset.

The main conclusion that can be drawn from the visual analysis of these results is that the removal of matrix-related peaks leads to a generalized enhancement in the contrast of morphological structures obtained with the first principal components. This is due to the fact that the variance contribution of the matrix-related signals is not fed to the PCA and therefore the resulting principal components are better focused on the morphology of the tissue. In agreement with previous work on the effects of MSI data reduction [35], these results demonstrate that the removal of matrix-related signals improves post-processing, especially when using linear algorithms such as the widely used PCA.

4.5. Performance comparison to blank subtraction

In order to quantify the improvement of rMSIcleanup over previous alternatives, we compared its performance to the widely used “blank subtraction”. Datasets 9 and 10 were used to perform such a comparison. In both cases, a Region Of Interest (ROI) outside of the tissue region was defined. The basic principle of “blank subtraction” relies on discarding signals found in these off-sample regions under the assumption that they are matrix-related. Figure S16 shows the two ROIS defined. In the case of Dataset 9, an ROI close to the tissue with apparent signs of metabolite delocalization was selected. The ROI for Dataset 10, on the other hand, was selected in a region that was far enough from the tissue and clean. Figure S17 compares the mean spectra of these ROIS to the mean spectrum of their respective datasets. It can be appreciated that overall the intensities are noticeably lower in the off-sample ROIs, especially in the clean ROI defined for Dataset 10. In this case, due to the thin layers of silver used in Ag-LDI, the spectrum is orders of magnitude less intense.

Three different metrics were evaluated to determine the presence of a peak in the off-sample ROI and thus label it as Ag-related by the “blank subtraction” algorithm. The three standard metrics used were: intensity fold-change between in and off-sample, off-sample intensity and the off-sample SNR. Figure S18 shows the precision vs recall curves obtained using “blank subtraction” based on the three metrics. The highest AUC of 0.61 reported for Dataset 10 using intensity as the classification metric was well below the reported AUC of 0.97 for rMSIcleanup.

As an example, analysed blank subtraction with a threshold of 10% of the maximum intensity (the top-performing metric). This resulted in 9 signals correctly classified as Ag-related (true positives TP), 141 correctly classified as not Ag-related (true negatives TN), 5 misclassified not Ag-related signals (false positives FP) and 19 misclassified Ag-related signals (false negatives FN). These results are associated with a poor false discovery rate (FP/(FP+TP)) of 35.7% and false omission rate (FN/(FN+TN)) of 11.88%. The blank subtraction method misclassified as matrix-related three metabolites with potentially

relevant biological information: choline (C5H14NO; as $[M + H - H_2O]^+$ *m/z* 86.09); cholesterol (C27H46O; as $\left[M + {}^{107}Ag\right]^+$ and $\left[M + {}^{109}Ag\right]^+$ *m/z* 493.24 and *m/z* 495.24 respectively) and an unidentified compound ($\left[M + {}^{107}Ag\right]^+$ and $\left[M + {}^{109}Ag\right]^+$, *m/z* 538.49 and *m/z* 540.49 respectively). On the other hand, several $[Ag_n]^+$ clusters (from n=4 to n=9) were overlooked by the blank subtraction method but were properly classified by rMSIcleanup.

5. Discussion and Conclusion

The goal of this study was to develop, optimize and validate a new algorithm to annotate signals attributed to the LDI promoting material in MSI. The developed algorithm is packaged and released as rMSIcleanup, an open-source R package freely available for the scientific community and fully integrated with rMSIproc [20], a stand-alone package for the visualization, pre-processing and analysis of MSI datasets.

As demonstrated, the widely used “blank subtraction” approach is outperformed by rMSIcleanup in the annotation Ag-related signals. In comparison to the top-performing alternatives for matrix-related peak annotation which are based on machine and deep learning [11], rMSIcleanup has the main advantage of using two intuitive scores (accounting for the isotopic ratios of clusters and the spatial distribution of their ions) and providing a visual justification of each annotation. This is a key contribution as it helps overcome the black-box problem, increases the user’s confidence in the annotation and can help researchers optimize experimental workflows (for instance, choosing LDI promoters that minimize interferences in the *m/z* range of interest). Another merit of our work is that, to our knowledge, it is the first matrix signal annotation algorithm to explicitly detect and deal with overlapping MS signals, which successfully prevents overlapped peaks from being misclassified. Given that we follow a targeted analytical approach, our classification is focused only on matrix-related signals while the algorithms presented by Ovchinnikova et al. [11] have a broader scope and also classify as off-sample other exogenous compounds. In the era of big data, these two apparently opposite approaches (namely our analytical approach based on chemical similarity scores and their untargeted approach based on machine learning) must not only coexist but also complement each other following the trend already initiated in other fields [36]. This reality urges the MSI community to develop annotation algorithms capable of, not only exploiting the knowledge in the increasingly large amounts of MSI datasets available, but also incorporating metrics that take into account the chemical context of the sample to aid transparent justification.

AgLDI MSI was chosen to validate the algorithm, due to the well-understood ionization of silver. A “validation list” was compiled from the literature, which included silver clusters that should be present in all samples and silver adducts or compounds that should not be present in any of them. Given the heterogeneity of the samples used in this study, the described validation list was adapted to each dataset. For each dataset, those clusters in the validation list for which the experimental data contained none of their theoretical masses were excluded. These adjustments in the validation list prevented an overestimation of the performance of the algorithm attributed to a high number of correctly classified “negative class” clusters (i.e. true negatives) located in mass ranges with no signal. We propose this validation strategy as a novel alternative to more

common validation approaches such as chemical standards [6] or expert annotation [11, 32]. This study adds to previous work [6, 14, 17, 29, 37] and further demonstrates the potentiality of AgLDI MS imaging, a thriving technology known for its reduced background signals in spatial metabolomics that is strongly complemented by our annotation algorithm as it further removes the influence of the matrix.

In agreement with previous work on the effects of MSI data reduction [35], we have demonstrated that the annotation and removal of signals related to the LDI promoting material used can further enhance post-processing, due to the elimination of variables attributed to exogenous compounds that do not reflect the morphology nor chemical composition of the sample. These results highlight the need to include software annotation tools such as rMSIcleanup in MSI workflows before exploring the datasets with classical data analysis techniques used in metabolomics. Here we would like to emphasize the need for a standardized quantitative metric to assess the quality of MSI images and we acknowledge the relevance of standardization initiatives such as the MALDISTAR project (www.maldistar.org).

We envision two main applications for rMSIcleanup. On the one hand, it can be used in a purely exploratory fashion to better understand ionization and adduct cluster formation in new matrices, tissues and applications. In this case, the user is advised to add a long list of potential adducts or neutral losses to assess their formation. The validation approach followed in this paper is a clear example of this exploratory application of rMSIcleanup. A second application is the automated peak annotation of well-known matrices and tissues. In this case, only the clusters that are known to be formed need to be given to the software. This curated selection increases the data-processing speed. The set of matrix-related annotated peaks can then be eliminated from the dataset prior to performing post-processing workflows such as multivariate statistical analysis. In any case, the choice of adducts and neutral losses to consider (or matrix adducts with endogenous compounds, e.g. fatty acids + Ag) is application dependent and is therefore left to the user. This list must be manually specified as an input parameter to rMSIcleanup.

Finally, the promising results obtained in the annotation of Ag_n^+-related peaks in AgLDI MSI open the door to the extension of this methodology to more widely used matrices such as 2,5-Dihydroxybenzoic acid (DHB), 1,5-Diaminonaphthalene (DAN), and 9-Aminoacridine (9AA) among others. These organic matrices pose greater challenges. Firstly, they lead to increased matrix background due to their greater fragmentation and adduct formation [38–40] and the higher quantities in which they are added [39]. Moreover, they present the problem of “hot spot” formation given their less homogeneous application process [41]. These issues highlight not only the benefits of AgLDI MSI but also that matrix-related peak annotation can benefit data post-processing even further in applications using organic matrices.

List of abbreviations

MSI: Mass spectrometry imaging; LDI: Laser desorption ionization; AgLDI MSI: Silver-assisted laser desorption ionization; PCA: Principal Component Analysis; MALDI: Matrix-Assisted Laser Desorption Ionization; ITO: Indium tin oxide; COS: Centre for Omics Sciences; TOF: Time of flight; FTICR: Fourier-transform ion cyclotron

resonance; AUC: Area under the curve; RGB: Red green blue; DHB: 2,5-Dihydroxybenzoic acid; DAN: 1,5-Diaminonaphthalene; 9AA: 9-Aminoacridine.

Availability of data and materials

The platform-independent R package rMSIcleanup presented in this publication is freely available under the terms of the GNU General Public License v3.0 at https://github.com/gbaquer/rMSIcleanup. The datasets supporting the conclusions of this article are available in the Mendeley Data repository: http://dx.doi.org/10.17632/vsk68tjcgh (Datasets 1 to 10), http://dx.doi.org/10.17632/9564zth9dj.1 (Datasets 11 and 12) and http://dx.doi.org/10.17632/hk32mhxkjh.1 (Datasets 13 and 14).

Competing interests

The authors declare that they have no competing interests.

Funding

The authors acknowledge the financial support of the Spanish Ministry of Economy and Competitivity through projects TEC2015-69076-P and RTI2018-096061-B-100. GB acknowledges the financial support of the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 713679 and the Universitat Rovira i Virgili (URV). LS acknowledges the financial support of Universitat Rovira i Virgili through the pre-doctoral grant 2017PMF-PIPF-60. MGA acknowledges the financial support from the Agency for Management of University and Research Grants of the Generalitat de Catalunya (AGAUR) through the post-doctoral grant 2018 BP 00188. PC acknowledges financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Authors' contributions

Gerard Baquer: Conceptualization, Methodology, Software, Validation, Visualization, Writing - Original Draft. **Lluc Sementé:** Validation, Writing - Review & Editing. **María Garcia-Altares:** Conceptualization, Methodology, Investigation, Writing - Review & Editing, Supervision. **Young Jin Lee:** Resources, Review & Editing. **Pierre Chaurand:** Resources, Review & Editing. **Xavier Correig:** Conceptualization, Methodology, Writing - Review & Editing, Supervision, Project administration, Funding acquisition. **Pere Ràfols:** Conceptualization, Methodology, Investigation, Writing - Review & Editing, Supervision.

Acknowledgements

Not applicable

6. Tables and Figures

No.	Species	Tissue type	Ag deposition system and estimated layer thickness	Lateral Res.	m/z range	Mass spectrometer	Acq. Mode	Ref.
1	Mouse	Pancreas	ATC Orion 8-HV Sputtering system, 5 nm	30	70-1200	Bruker ultrafleXtreme™ MALDI-TOF/TOF	Positive / Profile	-
2	Mouse	Pancreas	ATC Orion 8-HV Sputtering system, 5 nm	30	70-1200	Bruker ultrafleXtreme™ MALDI-TOF/TOF	Positive / Profile	-
3	Mouse	Kidney	ATC Orion 8-HV Sputtering system, 5 nm	100	70-1200	Bruker ultrafleXtreme™ MALDI-TOF/TOF	Positive / Profile	-
4	Mouse	Brain	ATC Orion 8-HV Sputtering system, 5 nm	80	70-1200	Bruker ultrafleXtreme™ MALDI-TOF/TOF	Positive / Profile	-
5	Mouse	Brain	ATC Orion 8-HV Sputtering system, 5 nm	80	70-1200	Bruker ultrafleXtreme™ MALDI-TOF/TOF	Positive / Profile	-
6	Mouse	Brain	ATC Orion 8-HV Sputtering system, 5 nm	80	70-1200	Bruker ultrafleXtreme™ MALDI-TOF/TOF	Positive / Profile	-
7	Mouse	Brain	ATC Orion 8-HV Sputtering system, 5 nm	80	80-1000	Bruker ultrafleXtreme™ MALDI-TOF/TOF	Positive / Profile	-
8	Mouse	Brain	ATC Orion 8-HV Sputtering system, 5 nm	80	80-1000	Bruker ultrafleXtreme™ MALDI-TOF/TOF	Positive / Profile	-
9	Mouse	Brain	ATC Orion 8-HV Sputtering system, 5 nm	80	80-1000	Bruker ultrafleXtreme™ MALDI-TOF/TOF	Positive / Profile	-
10	Mouse	Brain	ATC Orion 8-HV Sputtering system, 5 nm	80	80-1000	Bruker ultrafleXtreme™ MALDI-TOF/TOF	Positive / Profile	-
11	Mouse	Brain	Cressington Sputter Coater, 23 ± 2 nm	75	100-1100	Bruker ultrafleXtreme™ MALDI-TOF/TOF	Positive / Profile	[18]
12	Homo sapiens	Fingermark	Cressington Sputter Coater, 14 ± 2 nm	75	100-1100	Bruker ultrafleXtreme™ MALDI-TOF/TOF	Positive / Profile	[14]
13	B73 inbred	Root	Cressington 108Auto, 5s	10	50-970	Thermo Finnigan™ MALDI-LTQ-Orbitrap Discovery	Positive / Centroid	[6]
14	B73 inbred	Root	Cressington 108Auto, 5s	10	50-900	Thermo Finnigan™ MALDI-LTQ-Orbitrap Discovery	Negative / Centroid	[6]

Table 1. List of the 14 AgLDI MSI datasets used for validation. Sample type, sample preparation and LDI-MSI acquisition parameters. Datasets from 1-10 were acquired in-house. Datasets 11-14 were provided by external laboratories.

Chemical formula	Validation list	Type	Monoisotopic mass (n=1)	PubChem CID	Ref
$[Ag]_n^+$	Positive class	Silver cluster	106.9051	104755	[27]
$[AgF]_n$	Negative class	Neutral salt	125.903	62656	[28]
$[AgCl]_n$			141.8734	24561	
$[AgBr]_n$			185.8229	66199	
$[AgI]_n$			233.809	24563	
$[AgH]_n$		Synthetic compound	107.9124	139654	
$[AgH_2]_n$			108.9202	92028350	
$[AgHe]_n$			110.9072	71348557	
$[AgNO_3]_n$			168.8924	24470	
$[AgTh_2]_n$			570.9807	71351869	
$[AgF_2]_n$			144.9014	82221	
$[AgBF_4]_n$			192.9111	159722	
$[C_{27}H_{56} + Ag]_n^+$		Plants, wax, insects' pheromones	487.3428	-	[29]
$[C_{29}H_{60} + Ag]_n^+$			515.3741	-	
$[C_{31}H_{64} + Ag]_n^+$			543.4054	-	
$[C_{26}H_{54}O + Ag]_n^+$		Plant wax	489.322	-	
$[C_{28}H_{58}O + Ag]_n^+$			517.3533	-	
$[C_{30}H_{62}O + Ag]_n^+$			545.3846	-	
$[C_{26}H_{52}O_2 + Ag]_n^+$		Wax	503.3013	-	
$[C_{30}H_{60}O_2 + Ag]_n^+$			559.3639	-	

Table 2. "Validation list" used for validation. The "positive class" consists of silver clusters. The "negative class" consists of neutral silver salts, synthetic silver compounds and silver adducts that are not expected to be found in animal samples. The index n denotes the number of atoms or molecules inside the cluster. The minimum and maximum value of n depend on the monoisotopic mass of the atom or molecule and the mass range of the dataset.

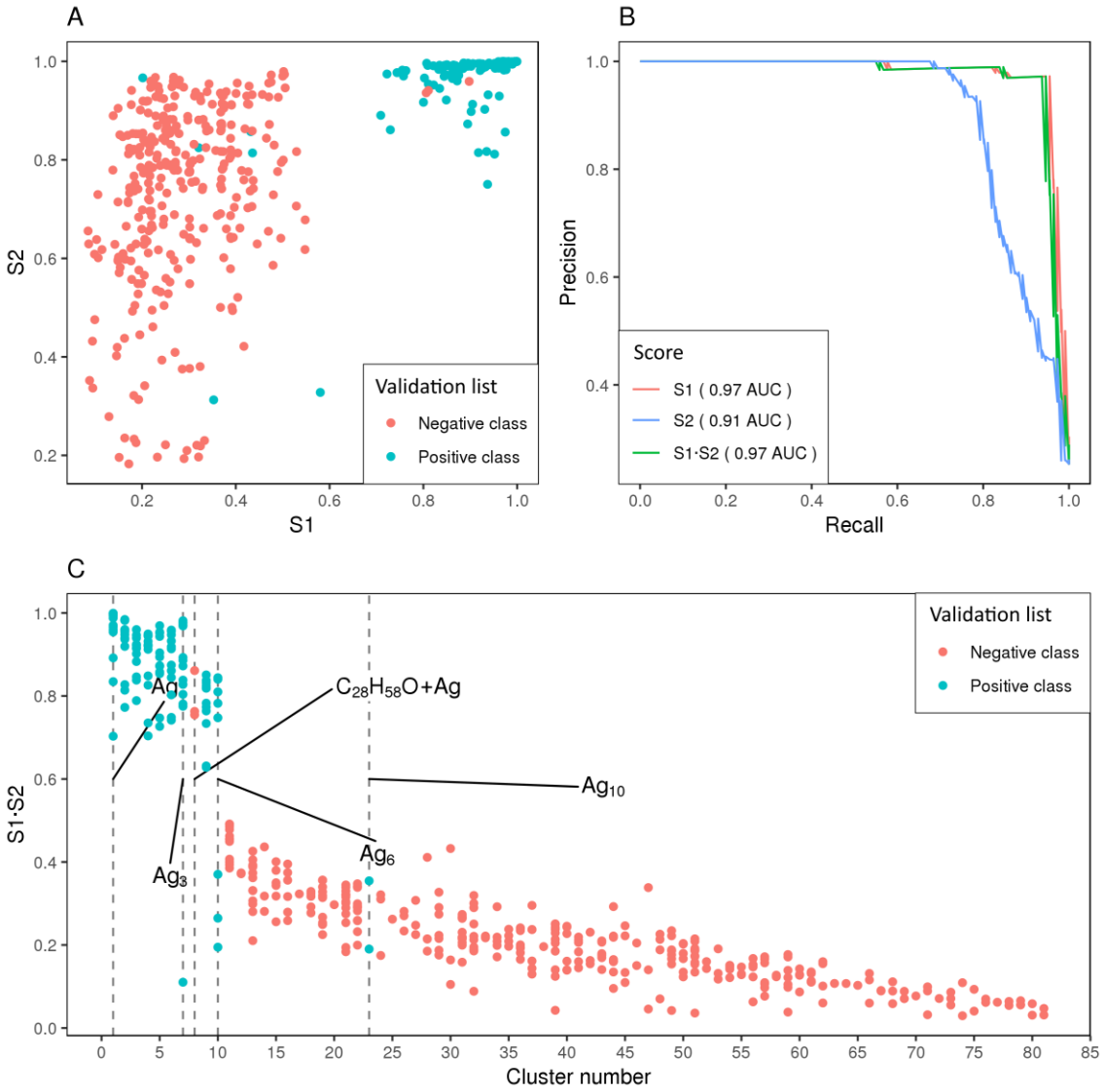


Figure 1. Similarity scores performance (A) Spectral similarity S1 vs. Intra-cluster morphological similarity S2 scatter plot. Each point represents a potential cluster classified by the algorithm. All clusters shown in Table 2 are evaluated for all 14 datasets presented in Table 1. Blue points represent the “positive class” (should be present in the sample) while the red points correspond to the negative class (should not be present in the sample). Most “positive class” points are located in the top right corner well separated from the negative class points. This indicates proper classification power. (B) Precision and recall (PR) curve computed according to Davis et al. 2006 [42]. (C) Similarity score S1·S2 vs. Cluster number. Clusters are arranged in decreasing order of mean similarity score. A clear gap between an S of 0.5 and 0.7 separates the “positive class” from the negative class. Refer to Supplementary Table S1 for a mapping of cluster numbers to cluster chemical formula.

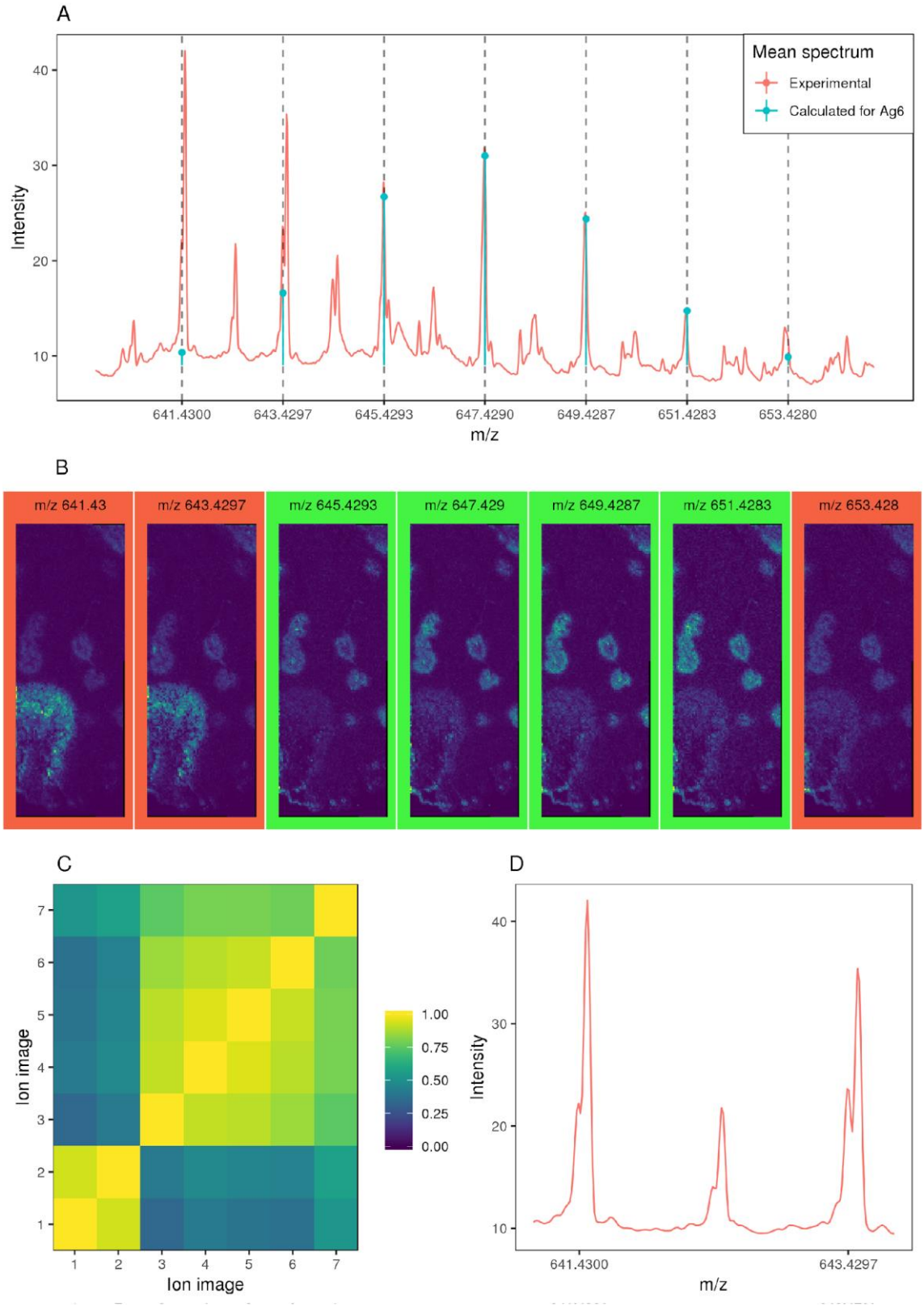


Figure 2. Overlapping detection algorithm performance when searching for the Ag_6 cluster in Dataset 1. (A) Comparison between the mean experimental spectra and the theoretical Ag_6 cluster at the Ag_6 cluster masses within a tolerance of 4 scans. Red and blue represent theoretical and experimental profiles, respectively. As can be seen, while the peaks in the centre of the cluster perfectly match the theoretical ratios, the peaks on

the edges differ considerably. (B) Spatial distributions of the experimental cluster peaks. After performing the overlapping detection only the four ion images in the centre in green are classified as Ag-related. The remaining ion images in red are classified as Ag^{n+} suffering from overlapping. The morphologies of the Ag^{n+} overlapped ions (red) differ from the ones without overlapping (green) due to ion overlapping. (C) Correlation matrix between the experimental ion images of the Ag_6 cluster. The ion image number corresponds to the position of the ion in the isotopic pattern in ascending order of m/z. The first two images are clearly not correlated with the remaining images of the cluster. The last image also shows a considerably lower correlation. (D) Zoom-in of experimental mean spectra. Peaks m/z 641.43 and m/z 643.43 show clear overlapping.

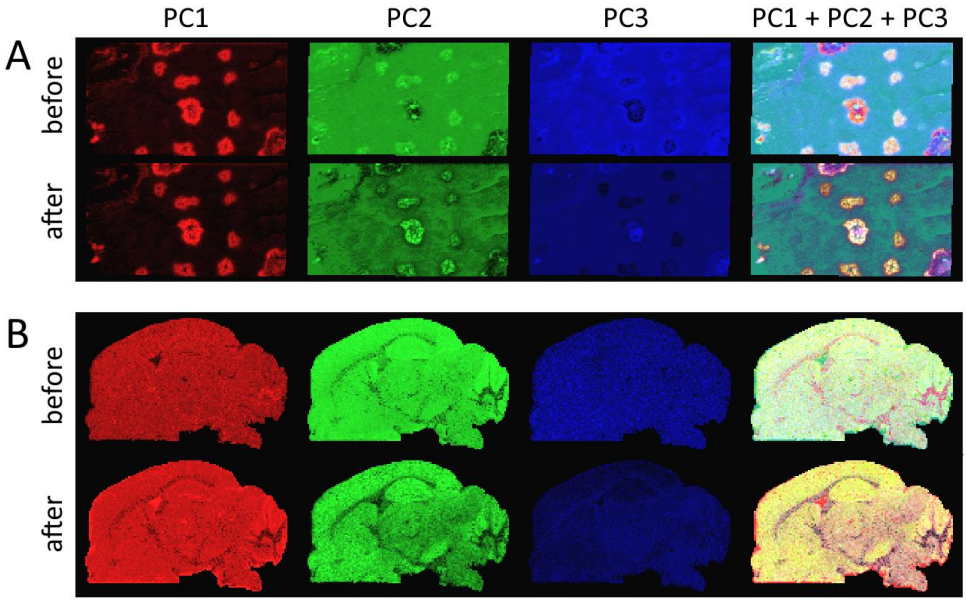


Figure 3. Exploratory analysis with PCA before and after removing matrix-related peaks. Red, green and blue are used to represent the spatial distribution of PC1, PC2 and PC3, respectively. The last column uses the Red Green Blue colour model (RGB) to represent the first three principal components in a single image. The annotation and removal of the matrix-related peaks lead to a generalized improvement in the contrast of morphological structures in all principal components. (A) Pancreas tissue from Dataset 2. (B) Brain tissue from Dataset 11 [18].

7. References

1. Rohner TC, Staab D, Stoeckli M (2005) MALDI mass spectrometric imaging of biological tissue sections. In: Mechanisms of Ageing and Development. pp 177–185

2. Norris JL, Caprioli RM (2013) Imaging mass spectrometry: A new tool for pathology in a molecular age. Proteomics - Clin Appl 7:733–738. https://doi.org/10.1002/prca.201300055

3. Cohen LH, Gusev AI (2002) Small molecule analysis by MALDI mass spectrometry. Anal Bioanal Chem 373:571–586. https://doi.org/10.1007/s00216-002-1321-z

4. Ràfols P, Vilalta D, Brezmes J, et al (2018) Signal preprocessing, multivariate analysis and software tools for MA(LDI)-TOF mass spectrometry imaging for biological applications. Mass Spectrom. Rev. 37:281–306

5. Ràfols P, Vilalta D, Torres S, et al (2018) Assessing the potential of sputtered gold nanolayers in mass spectrometry imaging for metabolomics applications. PLoS One 13:. https://doi.org/10.1371/journal.pone.0208908

6. Hansen RL, Dueñas ME, Lee YJ (2019) Sputter-Coated Metal Screening for Small Molecule Analysis and High-Spatial Resolution Imaging in Laser Desorption Ionization Mass Spectrometry. J Am Soc Mass Spectrom 30:299–308. https://doi.org/10.1007/s13361-018-2081-0

7. Iakab SA, Rafols P, García-Altares M, et al (2019) Silicon-Based Laser Desorption Ionization Mass Spectrometry for the Analysis of Biomolecules: A Progress Report. Adv. Funct. Mater. 29:1903609

8. Yagnik GB, Hansen RL, Korte AR, et al (2016) Large Scale Nanoparticle Screening for Small Molecule Analysis in Laser Desorption Ionization Mass Spectrometry. Anal Chem 88:8926–8930. https://doi.org/10.1021/acs.analchem.6b02732

9. Fonville JM, Carter C, Cloarec O, et al (2012) Robust data processing and normalization strategy for MALDI mass spectrometric imaging. Anal Chem 84:1310–1319. https://doi.org/10.1021/ac201767g

10. Annesley TM (2003) Ion suppression in mass spectrometry. Clin. Chem. 49:1041–1044

11. Ovchinnikova K, Kovalev V, Stuart L, Alexandrov T (2019) Recognizing off-sample mass spectrometry images with machine and deep learning. BioRxiv 518977. https://doi.org/10.1101/518977

12. Guan M, Zhang Z, Li S, et al (2018) Silver nanoparticles as matrix for MALDI FTICR MS profiling and imaging of diverse lipids in brain. Talanta 179:624–631. https://doi.org/10.1016/j.talanta.2017.11.067

13. Moule EC, Guinan TM, Gustafsson OJR, et al (2017) Silver-assisted development and imaging of fingermarks on non-porous and porous surfaces.

Int J Mass Spectrom 422:27–31. https://doi.org/10.1016/j.ijms.2017.08.001

14. Lauzon N, Dufresne M, Chauhan V, Chaurand P (2015) Development of laser desorption imaging mass spectrometry methods to investigate the molecular composition of latent fingermarks. J Am Soc Mass Spectrom 26:878–886. https://doi.org/10.1007/s13361-015-1123-0

15. Lauzon N, Chaurand P (2018) Detection of exogenous substances in latent fingermarks by silver-assisted LDI imaging MS: Perspectives in forensic sciences. Analyst 143:3586–3594. https://doi.org/10.1039/c8an00688a

16. Lauzon N, Dufresne M, Beaudoin A, Chaurand P (2017) Forensic analysis of latent fingermarks by silver-assisted LDI imaging MS on nonconductive surfaces. J Mass Spectrom 52:397–404. https://doi.org/10.1002/jms.3938

17. Dufresne M, Thomas A, Breault-Turcot J, et al (2013) Silver-assisted laser desorption ionization for high spatial resolution imaging mass spectrometry of olefins from thin tissue sections. Anal Chem 85:3318–3324. https://doi.org/10.1021/ac3037415

18. Thomas A, Patterson NH, Dufresne M, Chaurand P (2015) (MA)LDI MS imaging at high specificity and sensitivity. In: Advances in MALDI and Laser-Induced Soft Ionization Mass Spectrometry. Springer International Publishing, Cham, pp 129–147

19. Schramm T, Hester A, Klinkert I, et al (2012) ImzML - A common data format for the flexible exchange and processing of mass spectrometry imaging data. J Proteomics 75:5106–5110. https://doi.org/10.1016/j.jprot.2012.07.026

20. Ràfols P (2019) GitHub - prafols/rMSIproc: An open-source R package for mass spectrometry (MS) imaging data pre-processing. https://github.com/prafols/rMSIproc. Accessed 10 Dec 2019

21. Ràfols P, Castillo E del, Yanes O, et al (2018) Novel automated workflow for spectral alignment and mass calibration in MS imaging using a sputtered Ag nanolayer. Anal Chim Acta 1022:61–69. https://doi.org/10.1016/j.aca.2018.03.031

22. Ràfols P, Torres S, Ramírez N, et al (2017) RMSI: An R package for MS imaging data handling and visualization. Bioinformatics 33:2427–2428. https://doi.org/10.1093/bioinformatics/btx182

23. Loos M, Gerber C, Corona F, et al (2015) Accelerated isotope fine structure calculation using pruned transition trees. Anal Chem 87:5738–5744. https://doi.org/10.1021/acs.analchem.5b00941

24. Römpp A, Spengler B (2013) Mass spectrometry imaging with high resolution in mass and space. Histochem. Cell Biol. 139:759–783

25. Steinbach M, Karypis G, Kumar V, others (2000) A comparison of document clustering techniques, KDD workshop on text mining

26. Irsoy O, Yildiz OT, Alpaydin E (2012) Design and analysis of classifier learning experiments in bioinformatics: Survey and case studies. IEEE/ACM Trans

Comput Biol Bioinforma 9:1663–1675. https://doi.org/10.1109/TCBB.2012.117

27. Staudt C, Heinrich R, Wucher A (2000) Formation of large clusters during sputtering of silver. Nucl Instruments Methods Phys Res Sect B Beam Interact with Mater Atoms 164:677–686. https://doi.org/10.1016/S0168-583X(99)01078-2

28. Kim S, Chen J, Cheng T, et al (2019) PubChem 2019 update: Improved access to chemical data. Nucleic Acids Res 47:D1102–D1109. https://doi.org/10.1093/nar/gky1033

29. Jun JH, Song Z, Liu Z, et al (2010) High-spatial and high-mass resolution imaging of surface metabolites of arabidopsis thaliana by laser desorption-ionization mass spectrometry using colloidal silver. Anal Chem 82:3255–3265. https://doi.org/10.1021/ac902990p

30. Jollife IT, Cadima J (2016) Principal component analysis: A review and recent developments. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 374

31. Ovchinnikova K, Rakhlin A, Stuart L, et al (2019) ColocAI: artificial intelligence approach to quantify co-localization between mass spectrometry images. BioRxiv 758425. https://doi.org/10.1101/758425

32. Palmer A, Ovchinnikova E, Thuné M, et al (2015) Using collective expert judgements to evaluate quality measures of mass spectrometry images. In: Bioinformatics. pp i375–i384

33. Verbeeck N, Caprioli RM, Van de Plas R (2020) Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. Mass Spectrom Rev 39:245–291. https://doi.org/10.1002/mas.21602

34. Van De Plas R, De Moor B, Waelkens E (2007) Imaging mass spectrometry based exploration of biochemical tissue composition using peak intensity weighted PCA. In: 2007 IEEE/NIH Life Science Systems and Applications Workshop, LISA. IEEE Computer Society, pp 209–212

35. McDonnell LA, van Remoortere A, de Velde N, et al (2010) Imaging mass spectrometry data reduction: Automated feature identification and extraction. J Am Soc Mass Spectrom 21:1969–1978. https://doi.org/10.1016/j.jasms.2010.08.008

36. Guidotti R, Monreale A, Ruggieri S, et al (2018) A survey of methods for explaining black box models. ACM Comput Surv 51:. https://doi.org/10.1145/3236009

37. Yang E, Fournelle F, Chaurand P (2019) Silver spray deposition for AgLDI imaging MS of cholesterol and other olefins on thin tissue sections. J Mass Spectrom. https://doi.org/10.1002/jms.4428

38. Dreisewerd K (2003) The desorption process in MALDI. Chem. Rev. 103:395–425

39. Heeren RMA, Smith DF, Stauber J, et al (2009) Imaging Mass Spectrometry: Hype or Hope? J Am Soc Mass Spectrom 20:1006–1014. https://doi.org/10.1016/j.jasms.2009.01.011

40. MacAleese L, Stauber J, Heeren RMA (2009) Perspectives for imaging mass spectrometry in the proteomics landscape. Proteomics 9:819–834

41. Chiang CK, Chen WT, Chang HT (2011) Nanoparticle-based mass spectrometry for the analysis of biomolecules. Chem. Soc. Rev. 40:1269–1281

42. Davis J, Goadrich M (2006) The relationship between precision-recall and ROC curves. In: ACM International Conference Proceeding Series. pp 233–240

8. Supplementary Material

8.1. Visual Report

```
##############################################################################
– Package Version: 0.1.1
– Time: 2019–07–23 13:39:26
##############################################################################
IMAGE INFORMATION
– Peak matrix: 2016_06_01_Brain_Control–Ag.tar
– Full spectrum: NULL
– Number of peaks: 135
– Number of pixels: 10010
– Mass Range: [ 81.0431355625551 , 1002.15406104405 ]
##############################################################################
MATRIX INFORMATION
– Matrix formula: Ag1; Ag1F1; Ag1Cl1; Ag1Br1; Ag1I1; Ag1H1; Ag1H2; Ag1He1; Ag1N1O3; Ag1Th2; Ag1F2; Ag1B1F4; Ag1C27H5
Ag1C26H54O1; Ag1C28H58O1; Ag1C30H62O1; Ag1C26H52O2; Ag1C30H60O2
– Add list: F1; Cl1; Br1; I1; H1; H2; He1; N1O3; Th2; F2; B1F4; C27H56; C29H60; C31H64; C26H54O1; C28H58O1; C30H62O1;
– Substract list: NULL
– Maximum cluster multiplication: 10
– Base forms: Ag1; Ag1H1; Ag1H2; Ag1He1; Ag1F1; Ag1Cl1; Ag1F2; Ag1N1O3; Ag1Br1; Ag1B1F4; Ag2; Ag2H2; Ag2H4; Ag2He2
Ag2F4; Ag3; Ag3H3; Ag3H6; Ag3He3; Ag2N2O6; Ag2Br2; Ag3F3; Ag2B2F8; Ag3Cl3; Ag4; Ag4H4; Ag3F6; Ag4H8; Ag4He4; Ag2I2
Ag1C26H52O2; Ag4F4; Ag3N3O9; Ag1C29H60; Ag1C28H58O1; Ag5; Ag5H5; Ag1C31H64; Ag5H10; Ag1C30H62O1; Ag5He5; Ag
Ag4F8; Ag3B3F12; Ag5F5; Ag6; Ag6H6; Ag6H12; Ag6He6; Ag4N4O12; Ag3I3; Ag5Cl5; Ag5F10; Ag4Br4; Ag7; Ag7H7; Ag6F6; Ag
Ag5N5O15; Ag6Cl6; Ag8; Ag8H8; Ag6F12; Ag8H16; Ag7F7; Ag8He8; Ag5Br5; Ag4I4; Ag9; Ag5B5F20; Ag9H9; Ag2C54H112; Ag2
Ag9He9; Ag2C52H104O4; Ag8F8; Ag6N6O18; Ag7F14; Ag2C58H120; Ag2C56H116O2; Ag10; Ag10H10; Ag2C62H128; Ag10H20
Ag2C60H120O4; Ag9F9; Ag8Cl8; Ag2Th4; Ag8F16; Ag6B6F24; Ag5I5; Ag7N7O21; Ag10F10; Ag9Cl9; Ag7Br7; Ag9F18; Ag8N8O2
Ag10F20; Ag3C81H168; Ag3C78H162O3; Ag8Br8; Ag3C78H156O6; Ag9N9O27; Ag3C87H180; Ag8B8F32; Ag3C84H174O3; Ag3
Ag3C90H180O6; Ag10N10O30; Ag3Th6; Ag9B9F36; Ag10Br10; Ag8I8; Ag10B10F40; Ag4C108H224; Ag4C104H216O4; Ag4C10
Ag9I9; Ag4C124H256; Ag4C120H248O4; Ag4C120H240O8; Ag4Th8; Ag10I10; Ag5C135H280; Ag5C130H270O5; Ag5C130H260
Ag5C155H320; Ag5C150H310O5; Ag5C150H300O10; Ag5Th10; Ag6C162H336; Ag6C156H324O6; Ag6C156H312O12; Ag6C174
Ag6C180H372O6; Ag6C180H360O12; Ag7C189H392; Ag7C182H378O7; Ag6Th12; Ag7C182H364O14; Ag7C203H420; Ag7C196
Ag8C216H448; Ag8C208H432O8; Ag7C210H420O14; Ag7Th14; Ag8C208H416O16; Ag8C232H480; Ag8C224H464O8; Ag8C248
Ag9C234H486O9; Ag8C240H480O16; Ag9C234H468O18; Ag8Th16; Ag9C261H540; Ag9C252H522O9; Ag10C270H560; Ag9C27
Ag10C260H520O20; Ag9C270H540O18; Ag9Th18; Ag10C290H600; Ag10C280H580O10; Ag10C310H640; Ag10C300H620O10;
##############################################################################
PROCESSING INFORMATION
– S1 threshold: 0.8
– S2 threshold: 0.8
– Similarity method: euclidean
– Magnitude of interest: intensity
– Tolerance mode: scans
– Tolerance scans: 4
##############################################################################
```

Figure S1. Initial summary of the results including main metrics of the images, the chemical formulae, the potential cluster adduct and neutral losses.

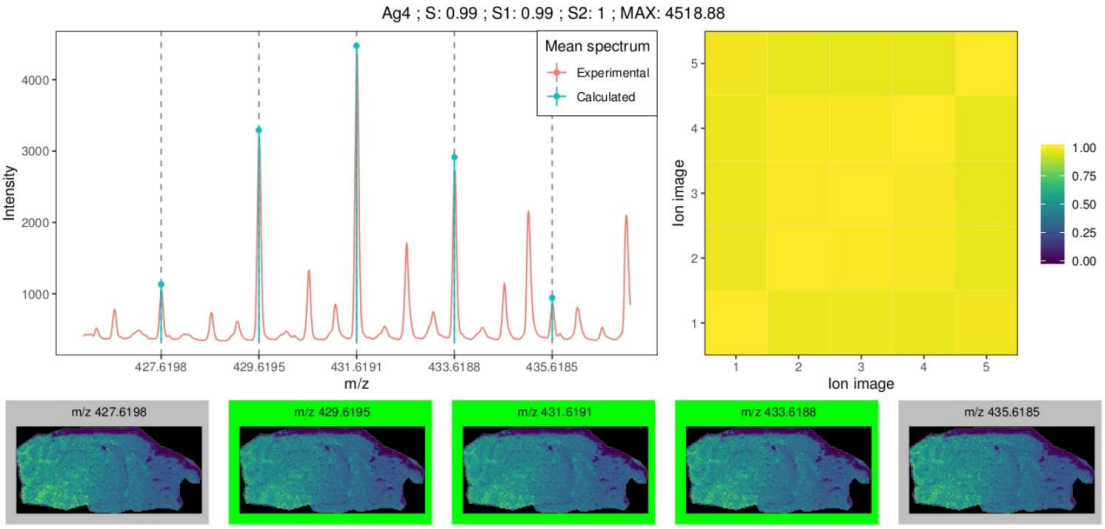


Figure S2. Visual report for cluster Ag4 in Dataset 7. The report includes the comparison between experimental and calculated peaks, the correlation map and all ionic images. The ionic images with a green border are tagged as silver-related. The ionic images with a grey border are not found in the peak list provided and are thus not classified.

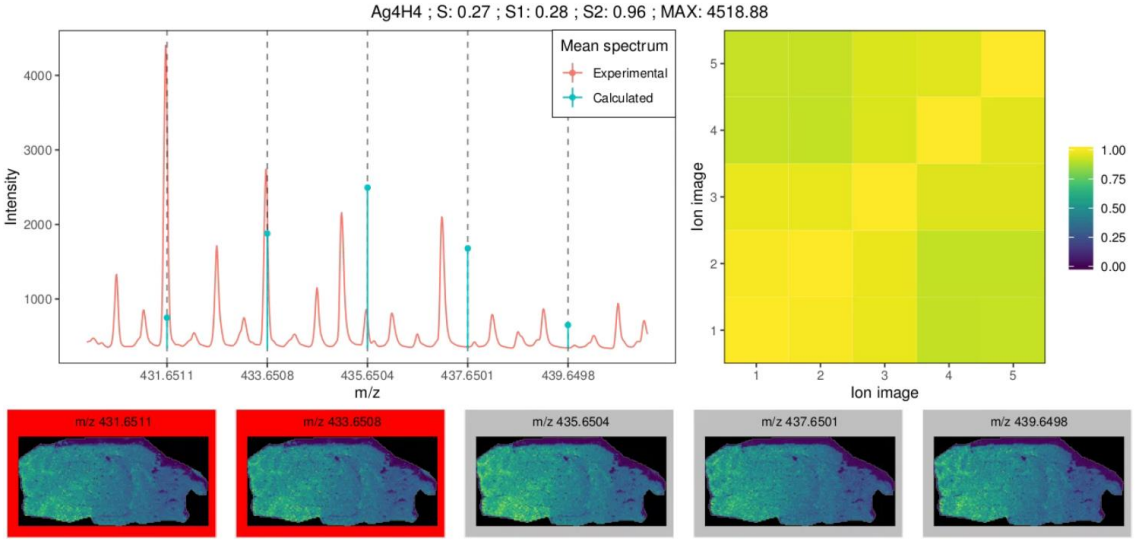


Figure S3. Visual report for cluster Ag4H4 in Dataset 7. The report includes the comparison between experimental and calculated peaks, the correlation map and all ionic images. The ionic images with a red border are tagged as not silver-related. The ionic images with a grey border are not found in the peak list provided and are thus not classified.

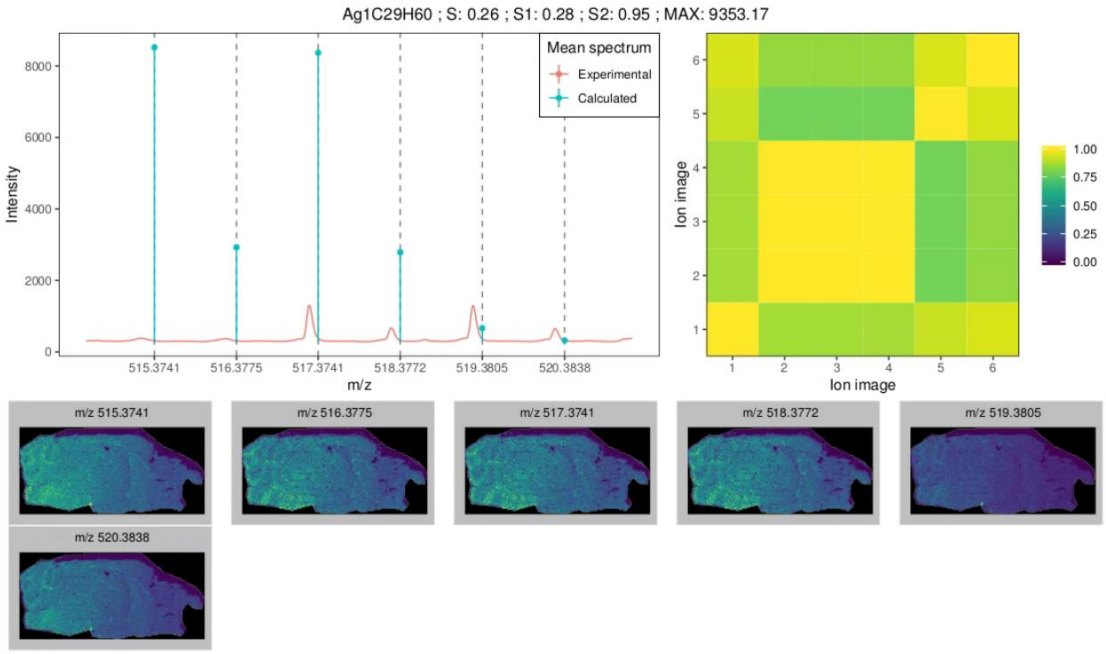


Figure S4. Visual report for cluster AgC29H60 in Dataset 7. The report includes the comparison between experimental and calculated peaks, the correlation map and all ionic images. The ionic images with a grey border are not found in the peak list provided and are thus not classified.

8.2. Table of cluster numbers

1	Ag_1	28	Ag_4Cl_4	55	Ag_7F_7
2	Ag_2	29	Ag_6He_6	56	$C_{60}H_{124}O_2 + Ag_2$
3	Ag_5	30	Ag_3Br_3	57	$Ag_5B_5F_{20}$
4	Ag_7	31	Ag_4H_4	58	Ag_8Cl_8
5	Ag_4	32	$Ag_1N_1O_3$	59	Ag_8H_8
6	Ag_9	33	$C_{26}H_{52}O_2 + Ag$	60	$C_{26}H_{54}O_1 + Ag_1$
7	Ag_3	34	$C_{29}H_{60} + Ag$	61	$C_{60}H_{120}O_4 + Ag_2$
8	$C_{28}H_{58}O_1 + Ag$	35	Ag_5F_{10}	62	Ag_9H_9
9	Ag_8	36	Ag_2H_4	63	$Ag_4B_4F_{16}$
10	Ag_6	37	Ag_5Cl_5	64	Ag_2F_4
11	Ag_3Cl_3	38	Ag_1I_1	65	Ag_5He_5
12	Ag_2TH_4	39	Ag_9He_9	66	$Ag_7N_7O_{21}$
13	Ag_2H_2	40	Ag_6F_{12}	67	Ag_9F_9
14	$C_{30}H_{60}O_2 + Ag$	41	Ag_1Cl_1	68	Ag_8F_{16}
15	Ag_6F_6	42	$C_{54}H_{112} + Ag_2$	69	$Ag_{10}He_{10}$
16	Ag_7H_7	43	Ag_3H_6	70	$Ag_4N_4O_{12}$
17	Ag_1F_2	44	Ag_7He_7	71	$Ag_6B_6F_{24}$
18	Ag_3I_3	45	$C_{52}H_{108}O_2 + Ag_2$	72	Ag_7F_{14}
19	Ag_1H_2	46	Ag_6H_{12}	73	$C_{52}H_{104}O_4 + Ag_2$
20	Ag_5F_5	47	$C_{62}H_{128} + Ag_2$	74	$C_{56}H_{116}\,O_2 + Ag_2$
21	Ag_1He_1	48	$Ag_1B_1F_4$	75	Ag_9H_{18}
22	Ag_6Cl_6	49	Ag_6H_6	76	Ag_8F_8
23	Ag_{10}	50	Ag_7H_{14}	77	$C_{58}H_{120} + Ag_2$
24	Ag_4Br_4	51	$Ag_5N_5O_{15}$	78	Ag_6Br_6
25	Ag_3F_6	52	Ag_5I_5	79	$Ag_{10}H_{10}$
26	Ag_4I_4	53	Ag_5H_{10}	80	$Ag_{10}H_{20}$
27	Ag_8He_8	54	Ag_8H_{16}	81	Ag_7Cl_7

Table S1. Cluster numbers used in Figure 1C in decreasing order of mean S1·S2 performance

8.3. Example clusters

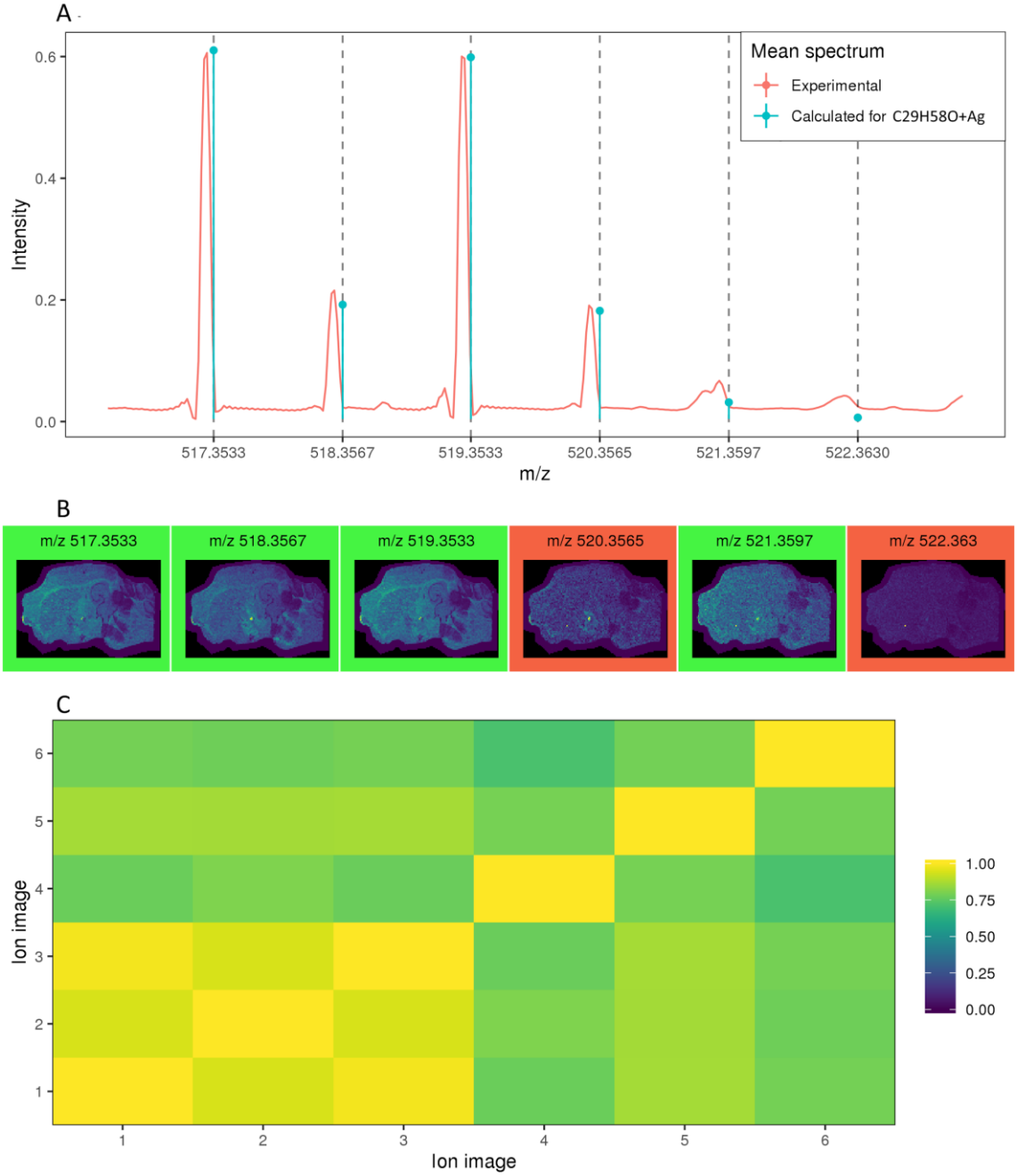


Figure S5. Classification results of cluster $C_{28}H_{58}O + Ag$ in Dataset 4. (A) Comparison between the mean experimental spectra and the theoretical pattern. (B) Spatial distributions of the experimental cluster peaks. (C) Correlation matrix between the experimental ionic images of the cluster. The cluster is misclassified as silver-related (false positive). Further study and annotation of these peaks would be needed to assess if the compound is indeed present in the sample implying that this specific compound should not be included in the "ground truth" as a negative class. Nevertheless, the constant and notable mass error between experimental and theoretical peaks allows us to infer that the experimental pattern might be due to a different compound. Adjusting the mass tolerance of the algorithm would get rid of these false positives.

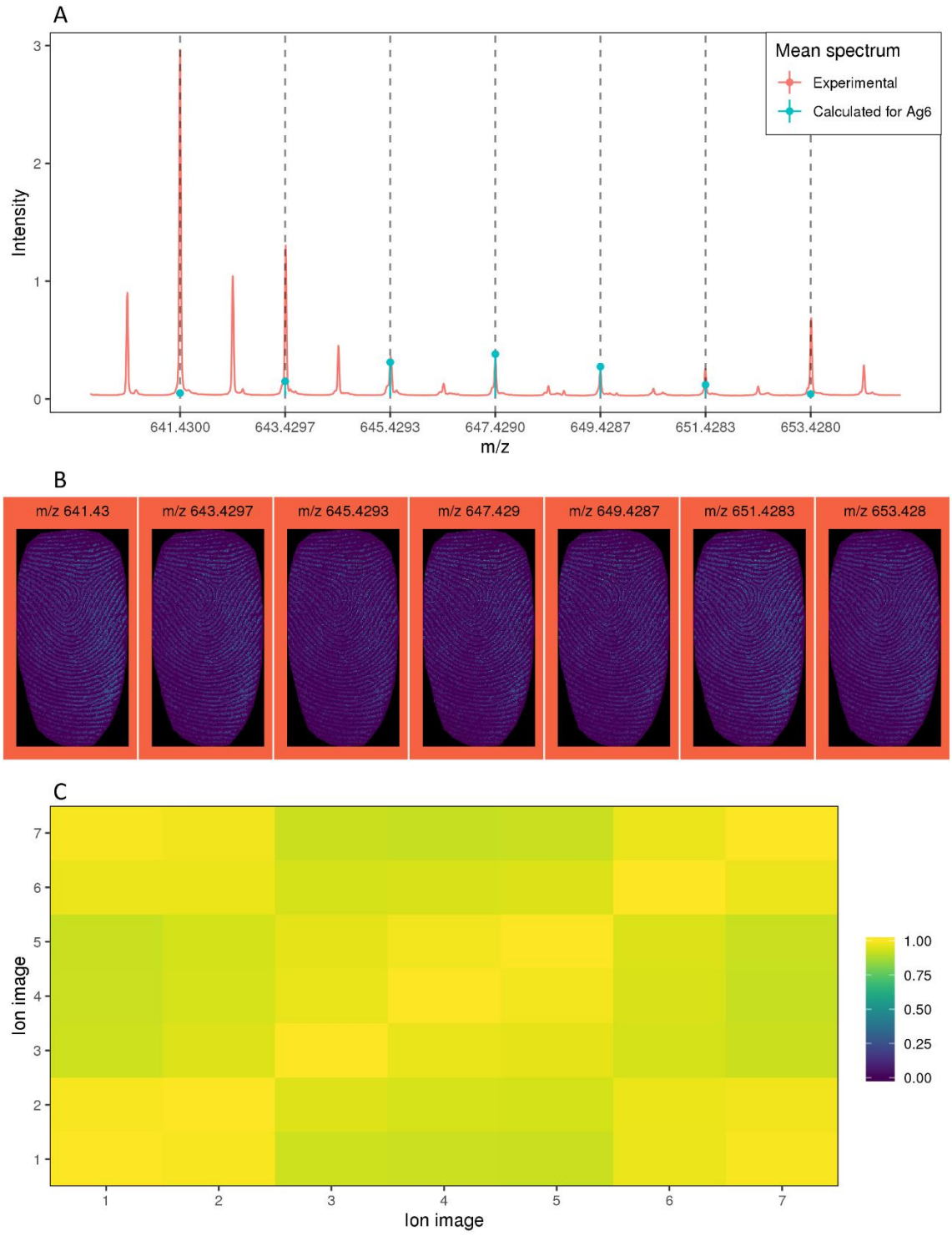


Figure S6. Classification results of cluster Ag_6 in Dataset 12. (A) Comparison between the mean experimental spectra and the theoretical pattern. (B) Spatial distributions of the experimental cluster peaks. (C) Correlation matrix between the experimental ionic images of the cluster. The cluster is misclassified as not silver-related (false negative). Like the example in Figure 2, peaks m/z 641.43, m/z 643.43 and m/z 653.43 clearly suffer from overlapping. Nevertheless, due to the high homogeneity of the fingerprint sample, the morphological correlation between the overlapped and the non-overlapped ions is relatively high. The overlapping detection algorithm fails to detect the overlapped peaks.

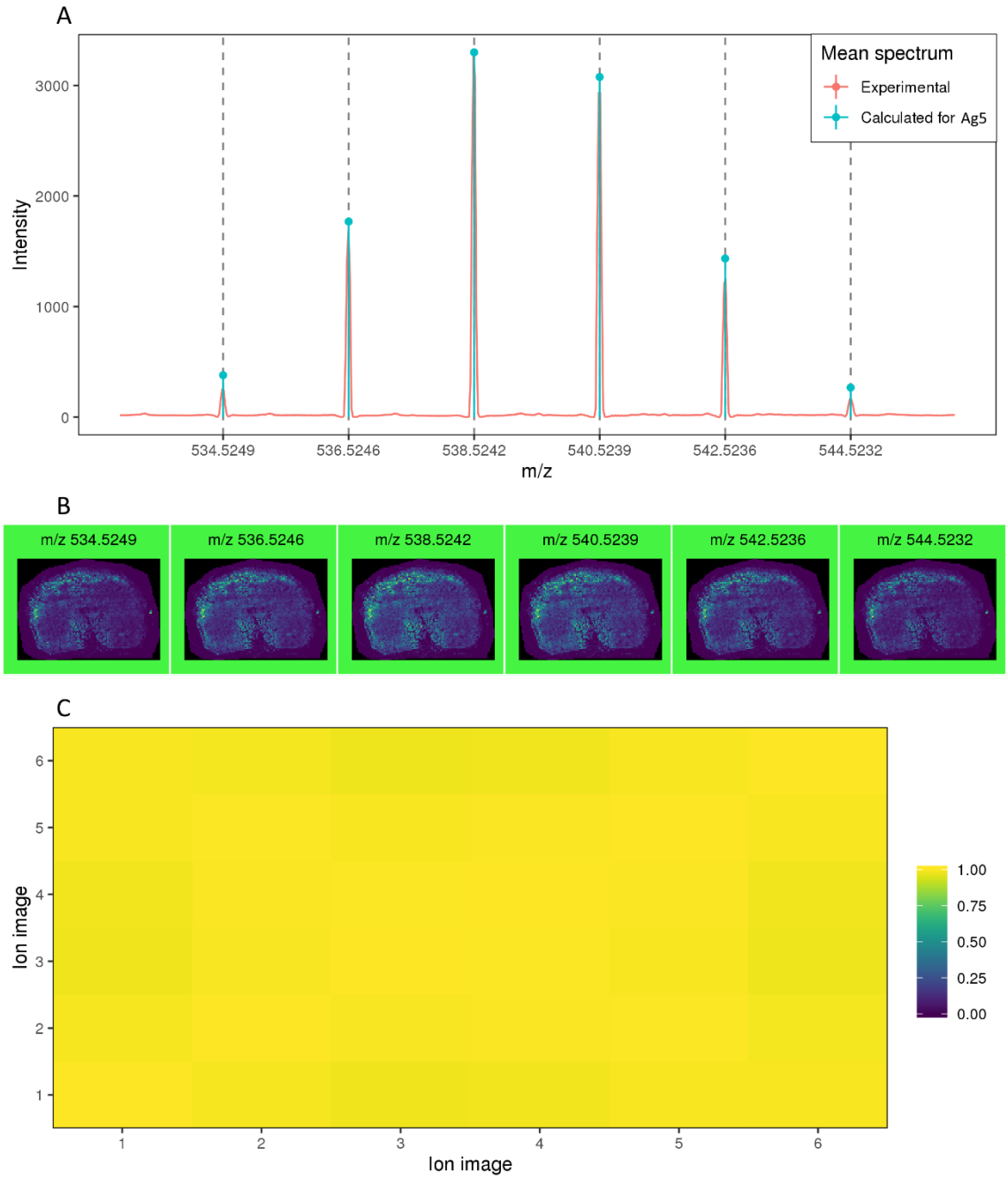


Figure S7. Classification results of cluster Ag_5 in Dataset 3. (A) Comparison between the mean experimental spectra and the theoretical pattern. (B) Spatial distributions of the experimental cluster peaks. (C) Correlation matrix between the experimental ionic images of the cluster. The cluster is correctly classified as silver-related (true positive).

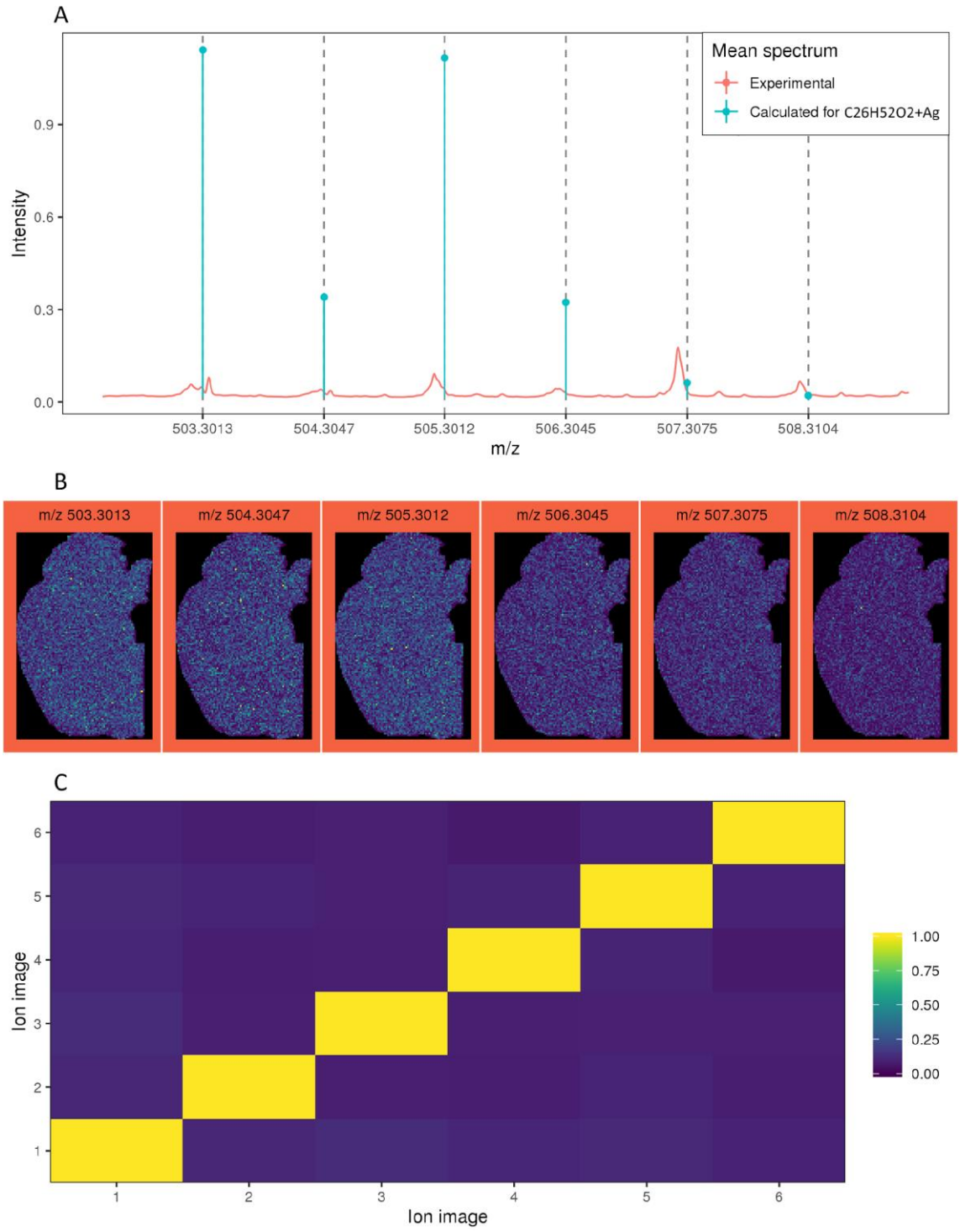


Figure S8. Classification results of cluster $C_{26}H_{52}O_2 + Ag$ in Dataset 11. (A) Comparison between the mean experimental spectra and the theoretical pattern. (B) Spatial distributions of the experimental cluster peaks. (C) Correlation matrix between the experimental ionic images of the cluster. The cluster is correctly classified as not silver-related (false positive).

8.4. Effects of overlapping peak detection

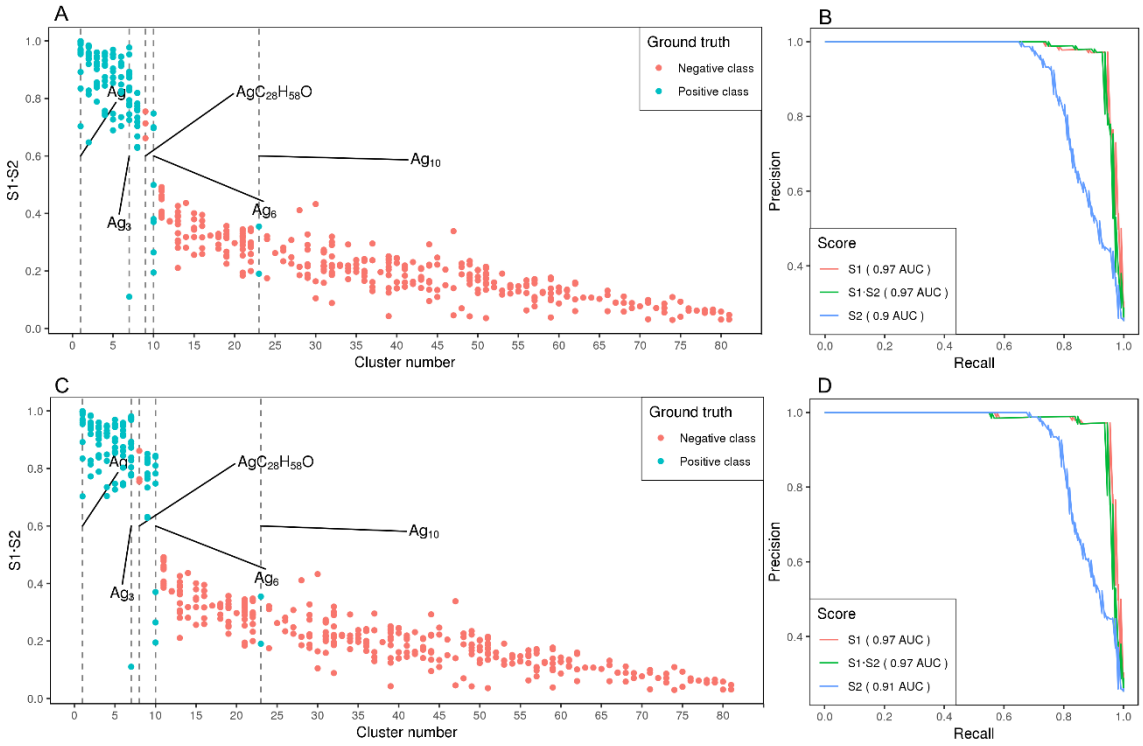


Figure S9. Similarity score S1·S2 vs. Cluster number and Precision vs. Recall curves with overlapping peak detection disabled or enabled. (A) & (B) Overlapping peak detection disabled. Multiple Ag6 clusters receive a low score and are thus misclassified as not Ag-related. (C) & (D) Overlapping peak detection enabled. The number of misclassified Ag6 clusters is considerably reduced. Additionally, the gap between the positive and negative class is now bigger leading to a more robust thresholding.

8.5. Complete exploratory analysis using PCA

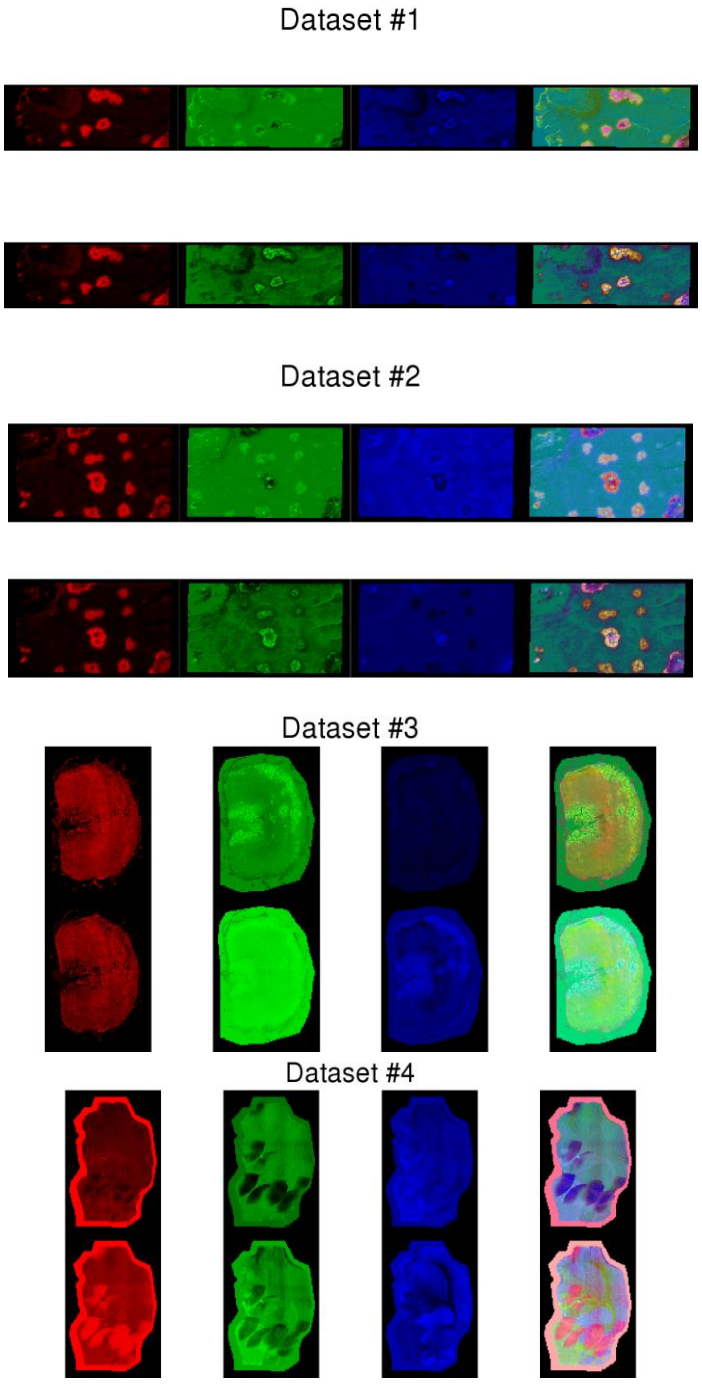


Figure S10. Exploratory analysis with PCA before (top row) and after (bottom row) removing matrix-related peaks for Datasets 1-4. Red, green and blue are used to represent the spatial distribution of PC1, PC2 and PC3, respectively. The last column uses the Red Green Blue colour model (RGB) to represent the first three principal components in a single image.

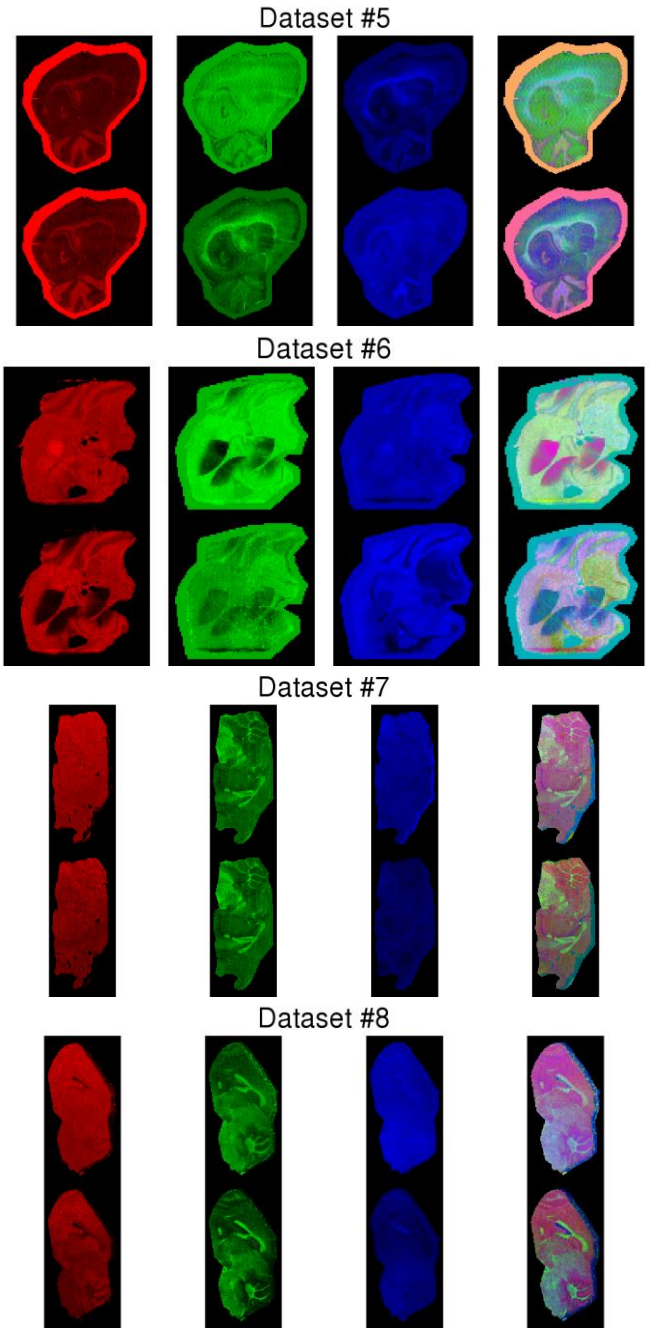


Figure S11. Exploratory analysis with PCA before (top row) and after (bottom row) removing matrix-related peaks for Datasets 5-8. Red, green and blue are used to represent the spatial distribution of PC1, PC2 and PC3, respectively. The last column uses the Red Green Blue colour model (RGB) to represent the first three principal components in a single image.

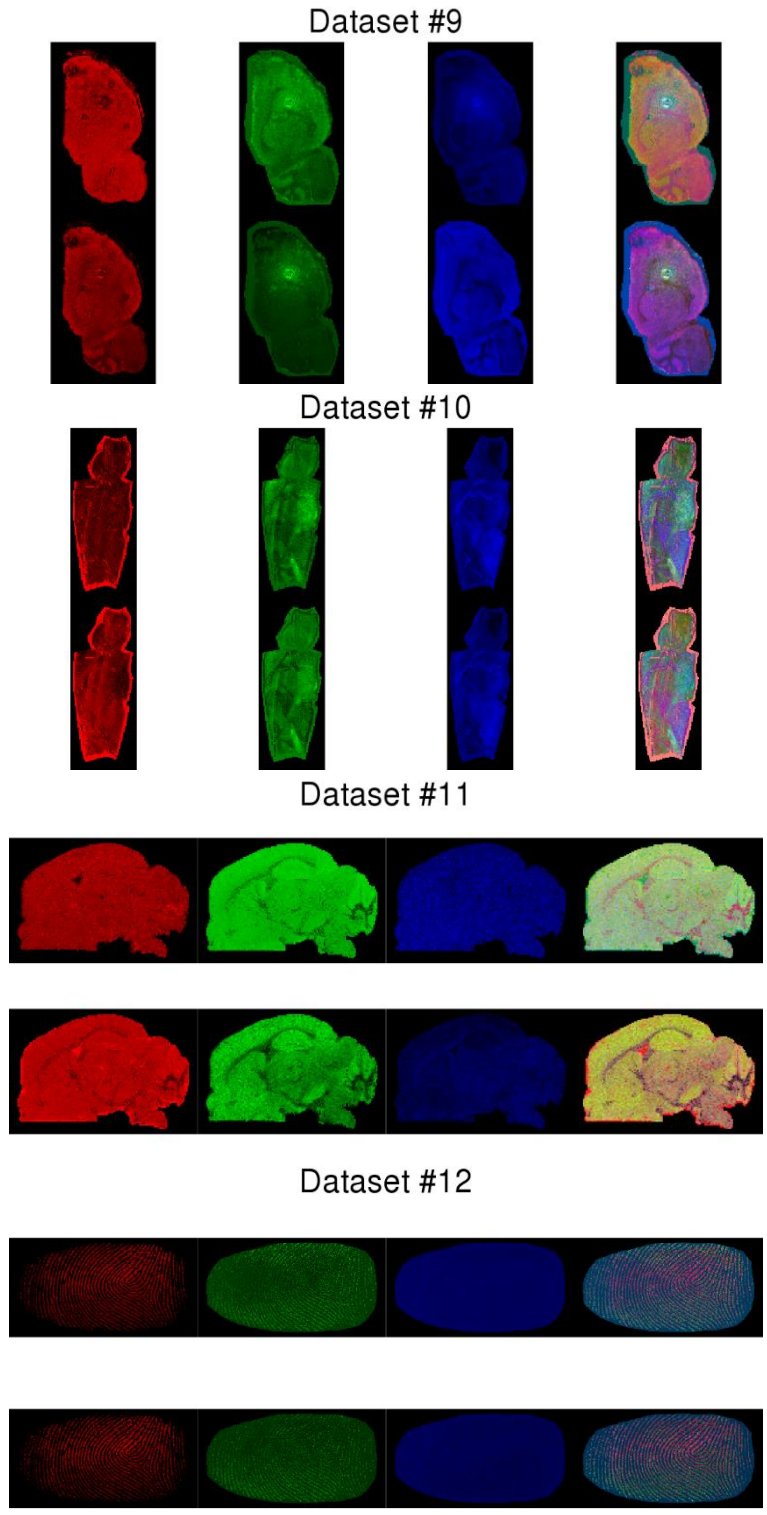


Figure S12. Exploratory analysis with PCA before (top row) and after (bottom row) removing matrix-related peaks for Datasets 9-12. Red, green and blue are used to represent the spatial distribution of PC1, PC2 and PC3, respectively. The last column uses the Red Green Blue colour model (RGB) to represent the first three principal components in a single image.

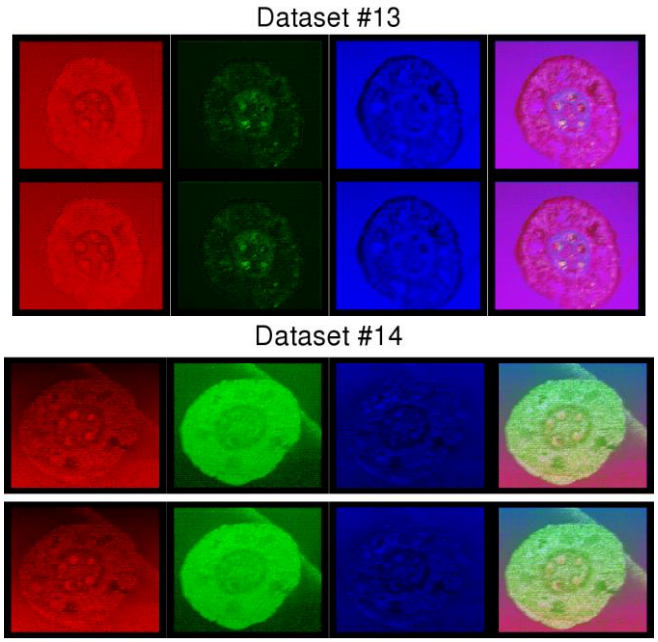


Figure S13. Exploratory analysis with PCA before (top row) and after (bottom row) removing matrix-related peaks for Datasets 13-14. Red, green and blue are used to represent the spatial distribution of PC1, PC2 and PC3, respectively. The last column uses the Red Green Blue colour model (RGB) to represent the first three principal components in a single image.

Table S2. Number of peaks, number of annotated Ag_n^+ peaks, reduction ratio and percentage of the Total Ion Count (TIC) accounted by Ag_n^+ peaks for all datasets.

Dataset	# Peaks	# Ag_n^+ peaks	Reduction ratio	TIC % (Ag_n^+ peaks)
1	1164	53	4.55%	31.78%
2	1164	51	4.38%	34.47%
3	381	46	12.07%	50.41%
4	621	55	8.86%	55.63%
5	625	45	7.2%	45.59%
6	585	41	7%	52.89%
7	135	39	28.89%	50.65%
8	135	39	28.89%	54.88%
9	174	43	24.71%	57.92%
10	174	43	24.71%	57.39%
11	1028	51	4.96%	49.99%
12	544	57	10.48%	32.75%
13	693	6	0.87%	2.71%
14	488	2	0.41%	0.07%

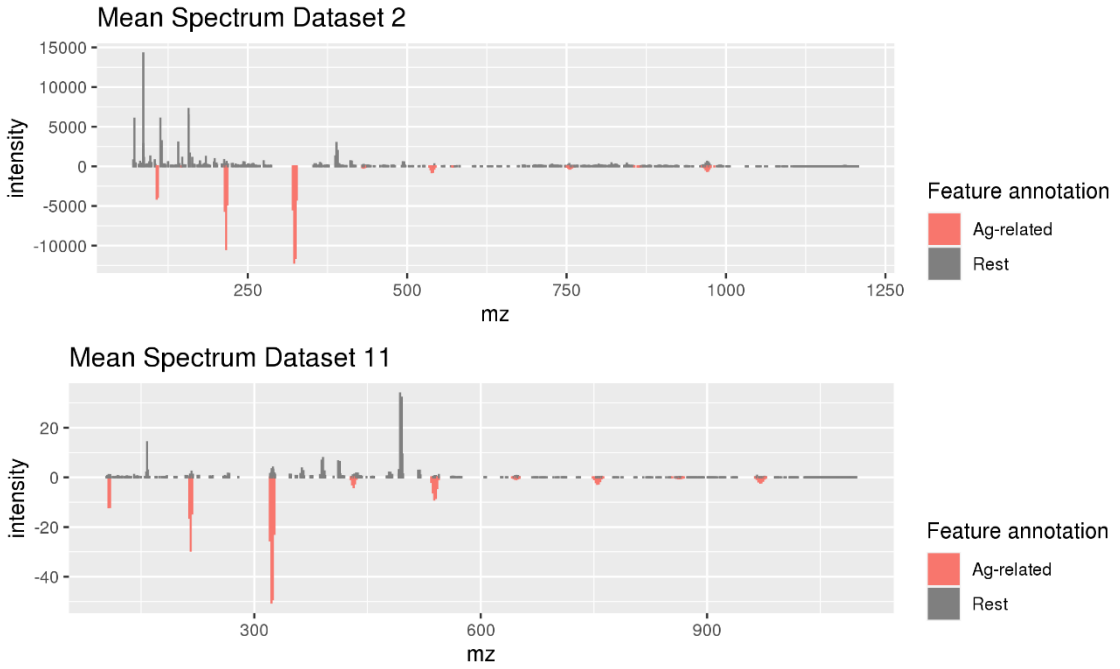


Figure S14. Mean spectra of Datasets 2 and 11. Red highlights Ag-related peaks.

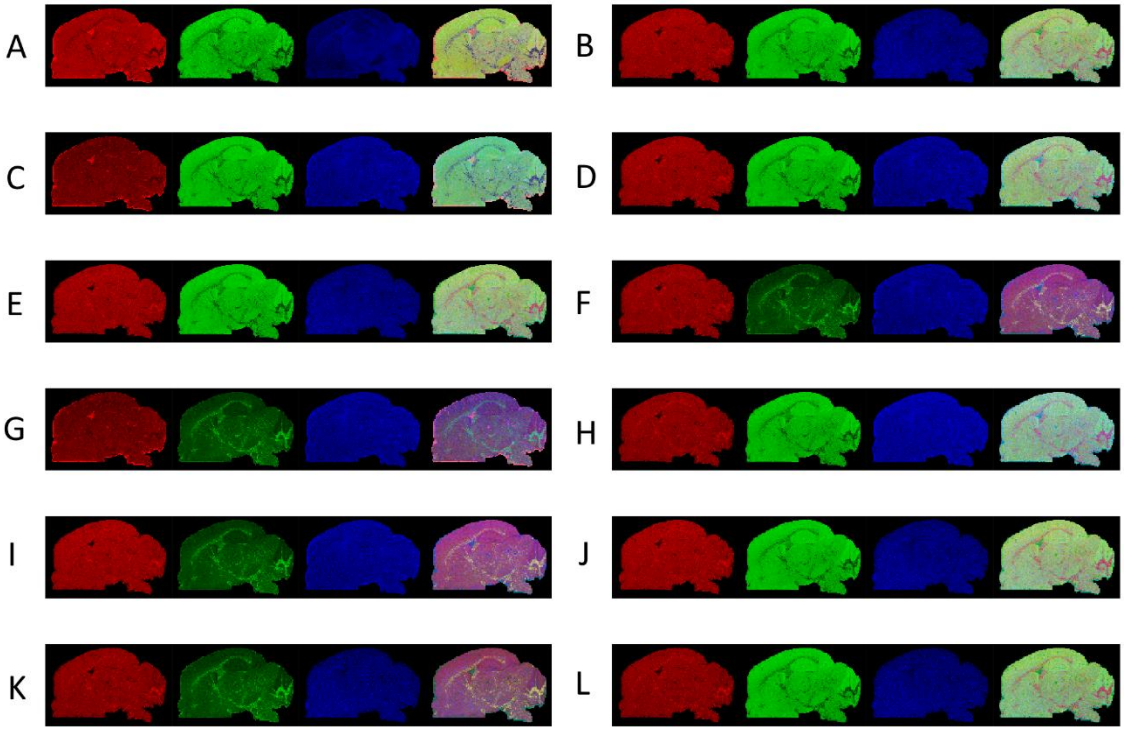


Figure S15. Exploratory analysis with PCA for Dataset 11. Matching the same number of features before and after removing matrix-related peaks. The Red, green and blue are used to represent the spatial distribution of PC1, PC2 and PC3, respectively. The last column uses the Red Green Blue colour model (RGB) to represent the first three principal components in a single image. (A) After removing matrix-related peaks. Containing 977 features. (B) Before removing matrix-related peaks. Selecting the 977 most intense features. (C-L) Before removing matrix-related peaks. Selecting 977 features randomly 10 times.

8.6. Performace comparison to blank subtraction

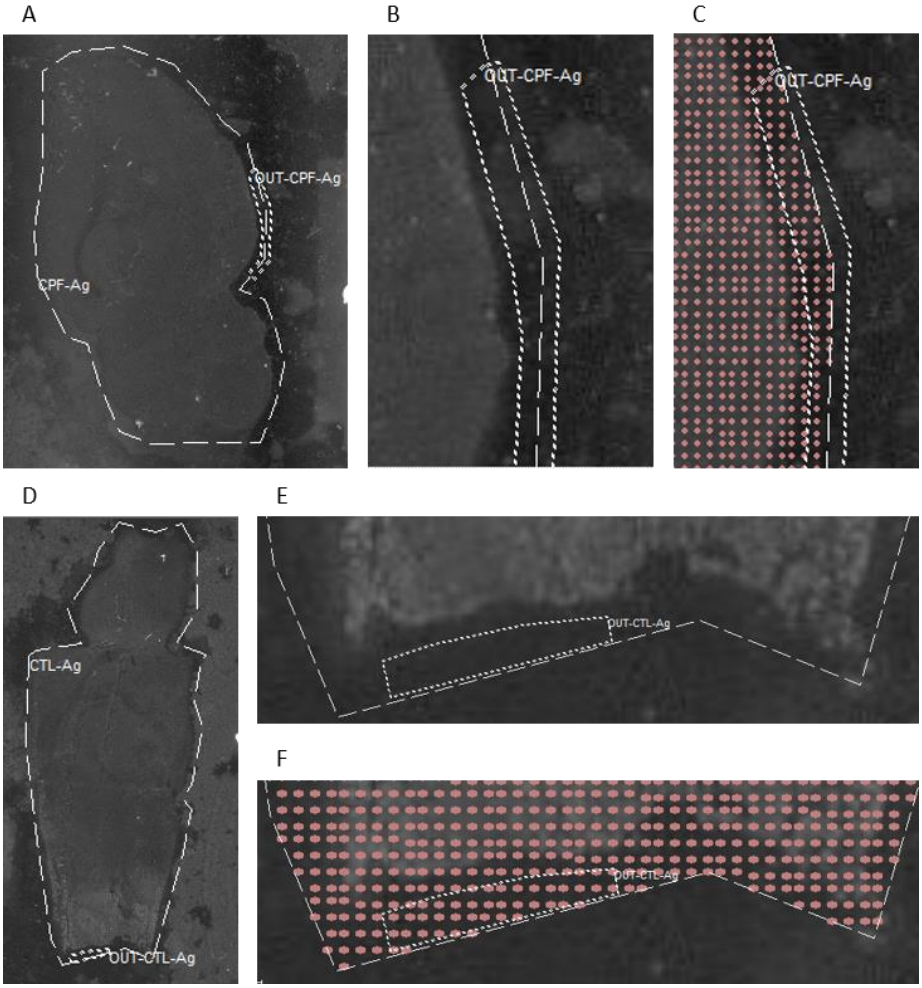


Figure S16. Regions Of Interest (ROI) outside of sample used to perform blank substraction. (A) Optical image of the brain slice used for Dataset 9. (B) Zoom-in of the off-sample ROI (C) Laser spots detail (D) Optical image of the brain slice used for Dataset 10. (E) Zoom-in of the off-sample ROI (F) Laser spots detail

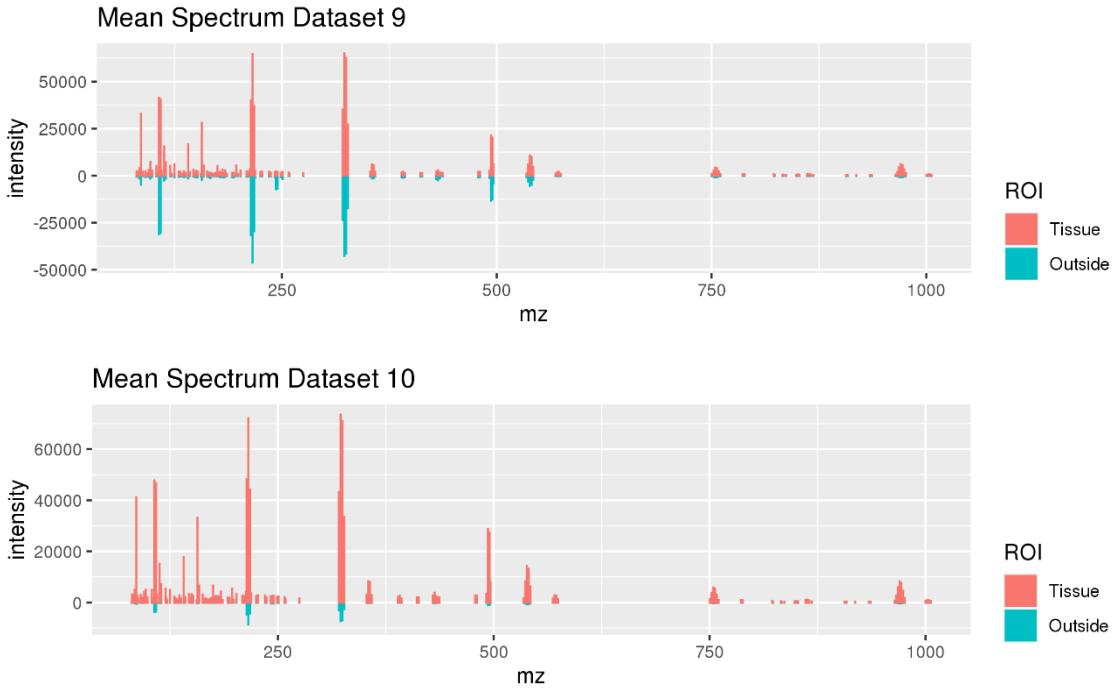


Figure S17. Mean spectra comparison between on-sample ROI (red) and off-sample ROI (blue) for Dataset 9 (top) and Dataset 10 (bottom). The out-sample spectra are considerably less intense and it is apparent that there are signals other than Ag-related signals.

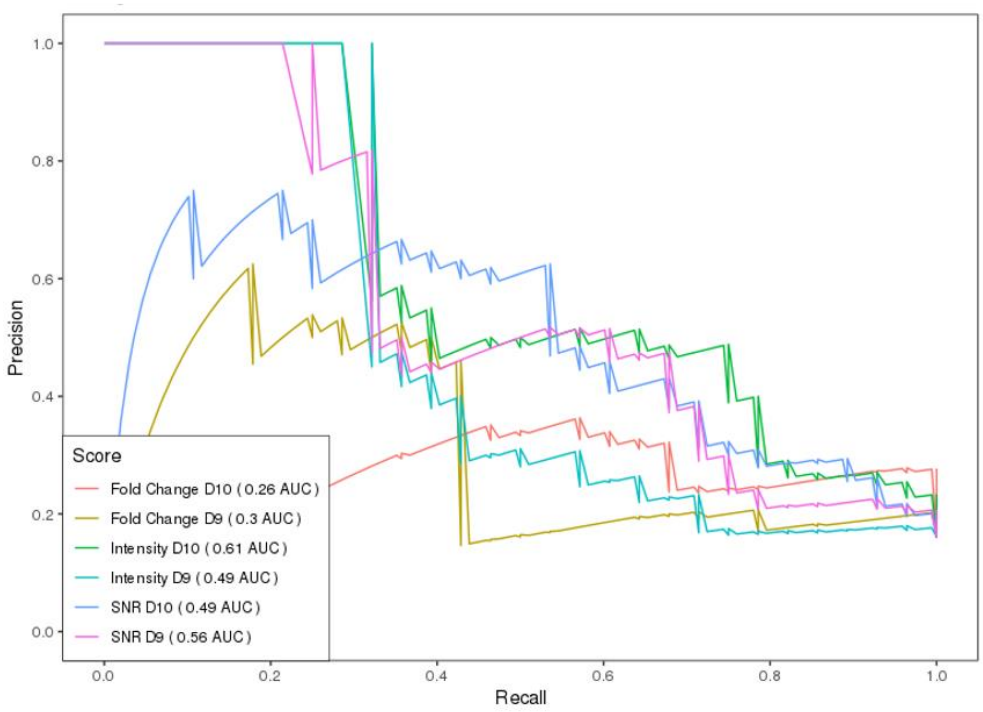


Figure S18. Precision and recall (PR) curve for background subtraction using three different metrics (Fold Change, Intensity and SNR) for Datasets 9 and 10. SNR and intensity are better classifiers than Fold Change as they report considerably higher area under the curve. The highest AUC of 0.61 reported for Dataset 10 using Intensity as the classification metric.

CHAPTER 5:

Stable Isotope Labeled MALDI matrix enables FDR-controlled discovery of endogenous and exogenous matrix-containing adducts in Mass Spectrometry Imaging

Gerard Baquer [1,#], Miguel Bernús [2,#], Lluc Sementé [1], René van Zeil [3], Maria García-Altares [1,2,*], Bram Heijs [3,4], Christoph Bookmeyer [1,5], Omar Boutureira [2], Xavier Correig [1,6,7], Pere Ràfols [1,6,7]

1 Department of Electronic Engineering, University Rovira i Virgili, Tarragona, Spain

2 Department of Analytical Chemistry and Organic Chemistry, Universitat Rovira i Virgili, Tarragona, Spain.

3 Center of Proteomics and Metabolomics, Leiden University Medical Center, Leiden, the Netherlands

4 The Novo Nordisk Foundation Center for Stem Cell Medicine (reNEW), Leiden University Medical Center, the Netherlands

5 Institute of Hygiene, University of Münster, Münster, Germany

6 Spanish Biomedical Research Center in Diabetes and Associated Metabolic Disorders (CIBERDEM), Madrid, Spain

7 Institut D'Investigacio Sanitaria Pere Virgili (IISPV), Tarragona, Spain

# These authors contributed equally to this work

*Correspondence:

Maria García-Altares, Department of Electronic Engineering, University Rovira i Virgili, Tarragona, Spain. Email: maria.garcia-altares@urv.cat

Abstract

Matrix Assisted Laser Desorption Ionization Mass Spectrometry Imaging (MALDI-MSI) has become a mature, widespread analytical technique to perform non-targeted spatial metabolomics. However, the matrix used to promote desorption and ionization causes spectral interferences in the low mass range that hinder downstream data processing. Matrix-containing adducts should be annotated and properly dealt with. Exogenous matrix-containing adducts should be removed to reduce non-biologically-relevant variables in the dataset. While matrix adducts with endogenous compounds, promise to shed light into the dark metabolome of MALDI-MSI.

Current automatic tools suffer from multiple of the following pitfalls: (1) focus exclusively on the spatial distribution, (2) do not control the False Discovery Rate (FDR), (3) do not consider adducts with endogenous metabolites, and (4) rely on a predefined list of theoretical matrix adducts.

Here we develop an experimental and computational workflow to discover matrix-containing adducts using $^{13}C_6$-labeled 2,5-Dihydroxybenzoic acid ($^{13}C_6$-DHB). By exploiting the labeling-induced m/z shift and unique spatial distribution of matrix-containing ions we can discover and annotate matrix-containing adducts formed with exogenous and endogenous compounds. This theoretical list can be used by rMSIcleanup or general purpose MSI annotation software to annotate matrix-related signals in MSI experiments under the same experimental conditions (tissue type, matrix, and ionization mode).

The computational part of the workflow is included in rMSIcleanup, an open-source R package to annotate and remove signals from the matrix. The results call for the community to incorporate matrix-related peak annotation tools such as rMSIcleanup in their MALDI-MSI workflows.

Keywords: MALDI, mass spectrometry imaging, annotation, matrix, DHB, stable isotope labeling, metabolomics, small molecule

1. Introduction

Matrix-Assisted Laser Desorption Ionization Mass Spectrometry Imaging (MALDI-MSI) is a booming analytical technique capable of spatially-resolving biomolecules in tissue samples (Rohner, Staab, and Stoeckli 2005). It has quickly become an invaluable technique in the analysis of complex and large biomolecules like proteins and peptides (Chaurand et al. 2006; Rohner, Staab, and Stoeckli 2005) and in recent years the interest has shifted towards metabolomics (Alexandrov 2020; Greer, Sturm, and Li 2011; Gao et al. 2022), the analysis of small molecules such as lipids, metabolites, and drugs. MALDI-MSI has successfully unraveled the metabolite complexity of molecular networks in complex diseases such as cancer (Notarangelo et al. 2022; Coy et al. 2022; L. Wang et al. 2022).

In the classical MALDI-MSI workflow, an organic compound (e.g. matrix) is deposited onto the sample to promote the desorption and ionization of endogenous analytes. Unfortunately, this low-weight exogenous compound adds several undesired MS signals to the MALDI-MSI spectra. Including exogenous matrix signals (adducts, multiple charges, and in-source fragments) and matrix adducts with endogenous biomolecules. These signals add an undesired layer of complexity to core MSI processing pipelines like untargeted statistical analyses (Baquer et al. 2020) or molecular annotation (Baquer et al. 2022). This is particularly worrying in metabolomics and lipidomics, as matrix-related signals are densely concentrated in the low *m/z* range.

This problem can be mitigated at the sample preparation level through derivatization, doping, or specialized deposition techniques (Calvano et al. 2018). Alternatively, rationally-designed MALDI matrices can reduce matrix-related signal interference. Some examples include organic matrices like 9-Aminoacridine (9AA) (Vermillion-Salsbury and Hercules 2002) and 1,5-Diaminonaphthalene (DAN) (Dong et al. 2013) or inorganic matrices like gold (Ràfols et al. 2018) or silver (Guan et al. 2018) sputtering, or silicon nanostructured substrates (Iakab et al. 2022).

Nevertheless, the adoption of some of these methods is still far from widespread as the choice of a matrix influences analyte/matrix co-crystalization, laser absorption, analyte ionization, and minimal analyte fragmentation. For this reason, first-generation matrices remain widely used in MALDI-MSI. In fact, 2,5-Dihydroxybenzoic acid (DHB) and alpha-Cyano-4-hydroxycinnamic acid (CHCA) represent 53% and 16% of all MALDI datasets in METASPACE (Alexandrov et al. 2019), the most widely used MSI database.

Apart from generating non-biologically relevant ions, the matrix is known to form adducts with endogenous metabolites (Janda et al. 2021). Traditionally, this phenomenon has been largely overlooked by automatic annotation tools which focus on a small list of possible adducts (M+H, M+Na, M+K, M-H, and M+Cl). Correctly annotating matrix-containing endogenous adducts would shed light on the dark metabolome (MS signals that cannot be identified) and increase the metabolome coverage and annotation confidence of MSI experiments.

In a previous study, we demonstrated the use of rMSIcleanup to automatically annotate matrix-related signals in Ag-LDI-MSI (Baquer et al. 2020). We also showed that the removal of matrix-related signals allowed Principal Component Analysis (PCA) to better focus on biologically-relevant morphology.

OffSample AI (Ovchinnikova et al. 2020) proposed several ML-based models to annotate ion images localized outside of the tissue (off-sample). Only half of the theoretical DHB

adducts detected (B. O. Keller and Li 2000) presented an off-sample spatial distribution, the rest were localized on-sample. Therefore, spatial distribution alone is insufficient to comprehensively annotate matrix-related signals. They also found that DHB-related signals comprise 5% of all detected ions. Nevertheless, their combinational model focuses on common adducts (-H,+Na,+K, and -H2O) and does not explore the prevalence of matrix adducts with endogenous biomolecules.

In this regard, mass2adduct (Janda et al. 2021) found a non-negligible number of metabolite-matrix adducts in DHB and CHCA datasets (up to 36% of all ions in certain datasets). As an example, amine-containing metabolites are prone to form matrix adducts (e.g. [M+(DHB-H2O)+H]+) in contrast to metabolites without nitrogen. Their approach exploits mass differences between all possible pairs of ions to automatically find metabolite-matrix adducts. The main caveats of their approach are: (1) it can only discover matrix adducts with endogenous metabolites already forming protonated or alkali adducts, (2) it does not control the False Discovery Rate (FDR), and (3) it only considers a limited number of matrix adducts.

Here, we present an experimental and cheminformatic workflow to identify matrix-related signals using stable isotope-labeled (SIL) organic matrices. We use a DHB analog with a 13C-labeled aromatic ring (*13C*^{6}-DHB) to introduce a known m/z shift and isotopic pattern to all matrix-containing ions. By comparing consecutive tissue slices prepared with DHB and *13C*^{6}-DHB, we are able to classify all ions into (1) endogenous, (2) exogenous matrix adduct, and (3) endogenous matrix adduct. The matrix adducts are then matched against a theoretical list of DHB adducts and in-source fragments and the Human Metabolome Database (Wishart et al. 2018) to provide a list of molecular annotations. Our annotations are FDR-controlled using a novel approach combining a decoy matrix and a decoy m/z shift distribution in a target-decoy setting (Elias and Gygi 2007). We show that matrix-related signals represent 15% of all ions and their removal is beneficial in untargeted applications, as it improves metabolite annotation and multivariate analyses. Additionally, we found that 23% of all ions correspond to matrix adducts with endogenous compounds. Considering these adducts in routinary metabolite annotation clears out a significant portion of the dark metabolome and increases the coverage of untargeted MSI metabolomics.

The experimental and computational protocol introduced can be deployed to discover matrix-containing adducts with other organic matrices, MALDI sources, and tissue types. The list can then be used as an in-house library to annotate matrix-containing adducts in experiments prepared with unlabeled matrices. All protocols are freely available at www.protocols.io/. Additionally, an R implementation of the computational protocol is released as a module of rMSIcleanup, an open-source R package for the annotation of matrix-related signals in MSI.

2. Results

2.1. $13C^6$-DHB produces high-quality MALDI-MS images

Our method for matrix-related signal discovery uses 4 consecutive slices of the same mouse brain: 2 prepared with DHB and 2 with *13C*^{6}-DHB (Figure 1). In all samples, the corpus callosum shows a higher TIC (Fig. 1A), commonly attributed to the dissimilar ion suppression compared to more densely packed brain tissue (Taylor, Dexter, and Bunch 2018). At a first glance, this structure shows finer details in the DHB samples. Overall,

the mean spectra of the two groups are similar, with comparable intensities and peaks (Fig. 1B). In fact, the distribution of pixel TIC across groups is identical (Fig. 1E).

We measure spatial autocorrelation (Moran's I) (Smets et al. 2019), as a proxy for morphology definition in ion images. Figure 1 D shows that off-sample regions (only matrix) show negligible autocorrelation (spatially noisy) while the tissue regions present positive autocorrelation (fine definition of morphological features). Sample A1 (DHB) presents higher levels of autocorrelation, indicating an overall better definition of anatomical structures, nevertheless, this difference is not statistically significant across groups (pval=0.83). Optical images of the samples (Fig. 1C) reveal more sectioning artifacts (cracks and folds) in the $^{13}C_6$-DHB samples which partially explains the perceived lower-quality images.

Finally, the UMAP (McInnes, Healy, and Melville 2018) projection of all pixels (Fig. 1G) reveals exceptionally low variability among technical replicates and clear separability between DHB and $^{13}C_6$-DHB. This reinforces the importance of matrix-related signals in unsupervised multivariate analyses. When mapped spatially, the first component (UMAP 1) (Fig. 1H) is similarly expressed across groups and highlights different anatomical features: cerebral cortex and midbrain (red), cerebral nuclei (black), thalamus and hypothalamus (pink), corpus callosum (green) and medulla (blue). Additionally, UMAP 1 also highlights the sectioning artifacts mostly affecting sample A4. On the other hand, the second component (UMAP 2) (Figure 1J) focuses on differences across groups (DHB vs $^{13}C_6$-DHB). These differences are expected, given the m/z shift introduced by the labeling, only endogenous ions are shared across groups. All matrix-related ions are expressed in one group and not present in the other. Interestingly, UMAP 2 also overlooks sectioning artifacts and brings out a fine definition of the corpus callosum in samples A3 and A4.

Collectively, these results show that $^{13}C_6$-DHB images are of marginally lower quality. Probably due to lower matrix purity (95% vs 99%) and a confounding sectioning artifact. But these differences are not statistically significant (p>0.5) and $^{13}C_6$-DHB produces ion images of high quality that can be used to discover matrix-related signals.

2.2. $^{13}C_6$-DHB enables the discovery of matrix-related signals

2.2.1. List of discovered matrix-containing adducts

Following the proposed workflow on mouse brain slices prepared with DHB and $^{13}C_6$-DHB we are able to annotate 12 matrix-containing exogenous adducts (Table 1) and 112 matrix-containing endogenous adducts (Table 2).

2.2.2. Workflow overview

The main experimental and computational workflow for matrix-related signal discovery is summarized in Figure 2. The complete protocols are available at www.protocols.io/. DHB and $^{13}C_6$-DHB are deposited onto consecutive tissue sections. All matrix-related signals will be characterized by (1) a known m/z shift (e.g. +6, +12, and +18 Da), and (2) matching spatial distribution. Additionally, their presence in off-sample regions (only matrix) will discern between endogenous and exogenous adducts. We can thus classify

all ions into (1) endogenous, (2) exogenous matrix adducts, and (3) endogenous matrix adducts (Fig. 2A).

The experimental spectra are matched against a list of theoretical matrix-related clusters, and annotation confidence is divided into 3 levels based on the evidence found: theoretical match, spatial classification, and SIL m/z shift (Fig. 2B). The top 2 levels (+++ and ++) correspond to hits to the theoretical DHB adducts and fragments (see Methods) while the bottom level (+) represent candidate adducts of DHB with endogenous metabolites. This level is later searched in the HMDB to find putative annotations.

Finally, we introduce a novel target-decoy approach to estimate the FDR using CHCA as a decoy matrix and a bimodal distribution of decoy m/z shifts (Fig. 2C).

2.2.3. Analysis of labeling induced m/z shifts

The first foundation of our computational method is the detection of m/z shift associated with the SIL. Figure 3A shows the most abundant m/z shifts between all possible pairs of peaks found in the DHB and *13C*^{6}-DHB samples. The three most abundant m/z shifts (z-score > 2$\boldsymbol{\sigma}$) are 0, +1, and +6 Da corresponding to endogenous, M+1 isotopes and [*13C*^{6}-DHB - DHB] shifts respectively. The 2*[*13C*^{6}-DHB - DHB] m/z shift (+12 Da) is also higher expressed (z-score > $\boldsymbol{\sigma}$). The +6 and +12 Da disappear when comparing the two DHB replicates (Fig. 3B). Figure 3C shows the density distribution of ppm error from +6 Da. There is a clear peak around 0 ppm when comparing DHB and *13C*^{6}-DHB (precision of 0.34). When comparing the two DHB samples, the distribution is uniform (precision of 0.12). As an example of negative control, when searching for an m/z shift that should not be dominant such as +2 Da both comparisons yield imprecise uniform distributions (Supplementary Fig. 1). To minimize false discoveries we rely on precision (AUC_{5ppm}/AUC_{50ppm}) instead of AUC_{5ppm}.

2.2.4. Ion classification based on spatial correlations

The second foundation of our computational workflow is the classification of ions based on their spatial distribution. Firstly, all samples are registered to enable spatial correlation of each ion across samples. For each ion, we compute the following metrics: (1) spatial correlation across all samples (DHB and *13C*^{6}-DHB), (2) spatial correlation within DHB samples, (3) spatial correlation within *13C*^{6}-DHB, and (4) mean intensity in each matrix control. Figure 4 maps these metrics into a UMAP projection that captures the spatial similarity between all ions (Fig. 4A). The three spatial correlations designate distinct regions in the projection corresponding to endogenous, DHB-related, and *13C*^{6}-DHB-related (Fig. 4B, C, D). The mean ion intensity in the matrix control regions allows us to further split each matrix-related group into on-sample and off-sample (Fig. 4E, F). Ions excluded by the criteria above are considered background as they convey minimal spatial information and have an abnormally low SNR.

This method relies on the use of technical replicates to compute the spatial correlation within each group. We highly encourage the use of technical replicates, but studies without replication could use spatial autocorrelation (Moran's I) as it shows a comparable ion classification (Supplementary Figure 2).

Figure 5 shows some example ions of the three main classes and their spatial correlations across samples. Endogenous ions m/z 203.2233 (Fig. 5A, B) and m/z 190.0122 (Fig. 5C, D) show a high spatial correlation across all samples. DHB-related on-sample ions like m/z 213.9642 only show a high spatial correlation within DHB

samples (Fig. 5E, F). And $13C^6$-DHB-related on-sample ions like m/z 219.9839 only show a high spatial correlation within $13C^6$-DHB samples (Fig. 5G, H). These two ions actually correspond to the same matrix-related signal as they show a +6 Da m/z shift and a high spatial correlation (r=0.34, p<0.01).

2.3. Validation in other DHB datasets

We used the list of discovered DHB-related matrix adducts to annotate several datasets using rMSIcleanup.

As a first validation, we annotated 21 datasets B1-B20 (including on-sample and off-sample regions) of consecutive mouse brain slices prepared with different matrices: DHB and Au acquired in positive ion mode and 9AA, NEDC, and Norharmane acquired in negative ion mode (Fig. 6). The high-confidence DHB-related adducts (+++ or ++) were only found in the samples prepared with DHB. With exception of most potassium adducts and the heavy adduct [4DHB+K+Na-H2O]2+ all high-confidence adducts were found in the DHB samples. Adducts [2DHB+Na-2H2O]+ and [2DHB+Na]+ were only found in off-sample regions while [M+2K-H]+ was only found on tissue (Fig 6A). Figure 6B shows the spatial distribution of some example DHB adducts.

When searching for the lower-confidence annotations (+) with no theoretical formula assigned we detect a few of them in non-DHB datasets (Fig 6B). The most notable is m/z 418.3054 which is present in all Au samples hinting at an endogenous metabolite ionized by both matrices. Overall the coverage of these m/z's in DHB samples is comprehensive. With some only present on-sample (m/z 197.4783 or 232.9446), off-sample (m/z 269.0424, 375.0089 or 471.0412), and in both (m/z 136.5201 or 188.4930).

As a second validation, we repeated the same annotation on datasets C1-C14 from METASPACE (Alexandrov et al. 2019) (Figure 7). They all are human biopsies of different tissues (brain, lung, kidney, and liver), prepared with DHB and acquired using different MS analyzers (FTICR and Orbitrap). Overall, all samples cover several DHB-related adducts. Interestingly, the coverage in lung and kidney samples (#1-#6) was considerably lower than in the brain samples (#9-#13). This could be explained by the higher sodium levels in the brain. It is important to highlight that ion images of exogenous matrix-related annotations correlate with off-sample regions (Fig 7B) while endogenous matrix-related annotations show other morphologies (Fig 7D).

2.4. The removal of matrix-related signals improves post-processing

2.4.1. Effects of matrix-related peaks removal on dimensionality reduction techniques

In this section, we explore the influence of matrix-related peaks in a typical untargeted analysis including dimensionality reduction and small molecule annotation.

Firstly we conduct a UMAP dimensionality reduction of all pixels in all consecutive mouse slices prepared with DHB and *13C*^{6}-DHB. As discussed in the quality control section, UMAP mainly highlights the difference between the two matrices (Fig 8A left) while the technical replicates within each group are closely intertwined. After the removal of all matrix-related peaks, the two groups are brought closer together and are almost

indistinguishable (Fig 8A right). These differences are better understood when contextualized in the spatial context (Fig 8B). When using all peaks in the samples both UMAP projections highlight differences between groups and convey few morphological features. After removing matrix-related signals both projections highlight different anatomical features with high contrast. In fact, both UMAP projections are almost indistinguishable irrespective of DHB labeling. The only apparent difference is the definition of the cortex in UMAP 2. These findings indicate correct and comprehensive annotation of matrix-related signals.

Similarly, when only analyzing Sample #1 (DHB), UMAP is able to better capture morphological features after the removal of matrix-related signals (Fig 8C). DBSCAN clustering (Schubert et al. 2017) of the projection with matrix-related peaks is only capable to identify 3 major clusters (Fig. 8C left) (background, cerebellum, and cerebrum+brain stem). When exclusively using endogenous ions, the same clustering identifies at least 5 major regions (background, cerebellum, cortex+interbrain, nuclei+midbrain, and hindbrain), (Fig. 8C right).

2.4.2. Effects of matrix-related peaks removal on metabolite annotation

As the second part of our untargeted analysis, we perform metabolite annotation (METASPACE + HMDB) (Alexandrov et al. 2019; Wishart et al. 2018) on Sample #1 before and after the removal of matrix-related signals (Fig 8D). When retaining all signals, METASPACE returns 97 annotations, with only 7 highly-reliable annotations (FDR<5%). When removing all matrix-related signals, the number of annotations increases to 126 with 10 highly-reliable annotations (FDR<5%). When comparing the annotated molecular formulas, the endogenous dataset retains 86% of annotations and 100% of reliable annotations (FDR<20%) found using the complete dataset. The results are clear, by removing matrix-related signals we reduce decoy hits while preserving target hits which ultimately leads to lower FDR values and more confident annotation.

Collectively, these results emphasize the importance of complete matrix-related annotation and removal, as it improves statistical analyses and small molecule annotation.

3. Discussion and Conclusion

We presented a novel experimental and computational workflow to discover matrix-related signals using SIL-MALDI-MSI based on the synthesis of a new DHB matrix, in which all the carbons of the aromatic ring have been replaced by 13C. The only previous work with a labeled matrix used a deuterated CHCAI matrix to uncover endogenous metabolites previously selected using a targeted approach (Shariatgorji et al. 2012).

We demonstrate that focusing on spatial (Ovchinnikova et al. 2020) or spectral information (Strohalm et al. 2010) alone is not enough to gain a comprehensive picture of the prevalence of matrix-related signals. The matrix forms adducts with both exogenous and endogenous compounds and it is thus present on and off-sample and lacks distinct matrix-like spatial distribution. This issue is further amplified by ion suppression effects as different molecular environments and tissue types can lead to diverse spatial patterns. Focusing on spectral information is not enough either as this can potentially add false positives with isomeric endogenous formulas. The spatial distribution helps us discern between endogenous and matrix related. Thus, the

annotation of matrix-related signals requires the integration of spatial and spectral information.

In consonance with previous studies (Janda et al. 2021) we found the number of matrix adducts with endogenous metabolites to be non-negligible. In this regard, to ensure confident annotation we introduce a novel FDR estimation paradigm based on decoy matrices and m/z shifts. This is a critical part of our workflow that enables us to work with higher levels of confidence and control the FDR.

A key finding of this work is the realization that matrix-related signals worsen the performance of typical untargeted. We found that the removal of matrix-related signals helps dimensionality reduction algorithms like UMAP (McInnes, Healy, and Melville 2018) better focus on biologically and anatomically relevant structures. Additionally, the removal of matrix-related signals also helps with the annotation of small molecules using automated tools like METASPACE (Alexandrov et al. 2019). Excluding matrix-related signals leads to an overall higher number of annotations with higher confidence (lower FDR). Echoing and expanding on previous studies (Baquer et al. 2020; Janda et al. 2021) we find that the removal of matrix-related signals improves the performance of untargeted MALDI-MSI efforts.

Finally, we provide a complete and validated database of DHB adducts with exogenous and endogenous compounds. This list can be used with rMSIcleanup to reliably and quickly annotate matrix-related signals in any dataset. It is worth noting that this new release leads to a considerable performance increase with respect to the initial release. The new release also includes an FDR estimation using a decoy library. Given a known list of adducts, rMSIcleanup can be used to confidently annotate matrix-related signals in any matrix.

This study opens a few different avenues for future work. Firstly, this methodology could be used to discover adducts in other commonly used matrices such as CHCA, DAN, or 9AA. In that regard, the main bottleneck is the limited availability of labeled MALDI matrix analog. Some aspects to be considered to be able to mass produce SIL-MALDI matrices include.

On the computational side, a really interesting avenue to explore would be modeling the MALDI matrix adduct from a molecular structure point of view. In a first exploration we could build a probabilistic model of the prevalence of different matrix-related adducts under different matrices, tissue types, and mass analyzers. What sort of adducts does a specific matrix generate? In this regard, the more than 7000 openly-available datasets in METASPACE would be really valuable. In a second and deeper iteration, we could aim to link that information to specific aspects of the molecular structure of the matrix and the adducts. Why does a specific matrix promote a certain type of adduct?

Abbreviations

2,5-Dihydroxybenzoic acid (DHB), 2,5-Dihydroxybenzoic acid with 13C-labeled aromatic ring (*13C*^{6}-DHB), alpha-Cyano-4-hydroxycinnamic acid (CHCA), 9-Aminoacridine (9AA), 1,5-Diaminonaphthalene (DAN), N-(1-naphthyl) ethylenediamine dihydrochloride (NEDC), Machine Learning (ML), Artificial Intelligence (AI), Human Metabolome Database (HMDB), Standard Isotope Labeled (SIL), False Discovery Rate (FDR), Nuclear Magnetic Resonance (NMR), Matrix Assister Laser Desorption Ionization

(MALDI), Mass Spectrometry Imaging (MSI), Total Ion Current (TIC), Uniform Manifold Approximation and Projection (UMAP), Competitive Fragmentation Modeling for Metabolite Identification (CFM-ID), Fourier-transform ion cyclotron resonance (FTICR), Time Of Flight (TOF), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Indium tin oxide (ITO), Area Under the Curve (AUC)

Acknowledgments

We acknowledge Denis Abu Sammour (HS Mannheim), Elisa Ruhland (IBMP), Brittney Gorman (PNNL), and Jessica Lukowski (PNNL) and respective colleagues as the original contributors of the METASPACE datasets used for validation.

Authors contributions

GB, MB, MG, OB, XC, and PR developed the concept for the study. MB and OB designed, performed, and validated the synthesis of $^{13}C_6$-DHB. RZ, BH, and CB designed and performed mass spectrometry imaging experiments. GB developed the computational workflow in collaboration with LS, MG, XC, and PR. GB and MB wrote the original manuscript with substantial edits and contributions from all authors. BH, OB, and XC provided supervision, project administration, and funding.

Funding

The authors acknowledge the financial support of the Spanish Ministry of Economy and Competitivity through project RTI2018096061-B-100. GB acknowledges the financial support of the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 713679 and the Universitat Rovira i Virgili (URV). LS acknowledges the financial support of Universitat Rovira i Virgili through the pre-doctoral grant 2017PMF-PIPF-60. MGA acknowledges the financial support from the Agency for Management of University and Research Grants of the Generalitat de Catalunya (AGAUR) through the postdoctoral grant 2018 BP 00188.

Availability of data and materials

The platform-independent R package rMSIcleanup presented in this publication is freely available under the terms of the GNU General Public License v3.0 at https://github.com/gbaquer/rMSIcleanup. The datasets supporting the conclusions of this article are available in the Mendeley Data repository. Datasets C1-C15 are available at https://metaspace2020.eu/ (References provided in Table S1).

4. Methods

4.1. MSI data processing

All samples were visualized and exported to .imzML using SCiLS (Bruker). The .imzML files were processed and exported to a centroid-mode peak matrix using rMSI2 (Ràfols et al. 2020). The default processing parameters were used. The Signal-to-Noise Ratio (SNR) threshold was set to 5 and the Savitzky– Golay smoothing had a kernel size of 7. Peaks appearing in less than 5% of the pixels were filtered out. Peaks within a window of 6 data points or scans were binned together as the same mass peak. No intensity normalization was performed.

Image registration of Datasets X-Y was performed using rMSIworkflows. Manually specified fiducial markers were used to calculate and apply a rigid transformation (rotation, translation, and scaling).

4.2. Statistical methods

All statistical group comparisons were performed at a pixel level using the linear mixed effects model in the nlme R package (Pinheiro 2009). We considered sample ID as a random effect and adjusted the p-values using FDR correction (Benjamini and Hochberg 1995).

Autocorrelation of ion images was computed using Moran's I test available in the moranfast R package (https://github.com/mcooper/moranfast). The spatial correlation of ion images was computed using Pearson's method. All correlation and autocorrelation p-values were FDR-corrected and considered significant if the p-val<0.05. After plotting all ranked significant correlation values we manually defined a minimum correlation threshold of 0.065.

All UMAP (McInnes, Healy, and Melville 2018) projections were computed using the instantiation in uwot R package (Melville, Lun, and Djekidel 2020). Segmentation of UMAP projections was conducted in the dbscan R package (Hahsler, Piekenbrock, and Doran 2019) ($\boldsymbol{\varepsilon}$=0.3). METASPACE (Alexandrov et al. 2019) was used for metabolite annotation against the Human Metabolome Database (Wishart et al. 2018) (10 ppm) considering all adducts available (M+, M+H, M+Na, M+K, M+NH4).

4.3. Discovery of matrix-related signals

Using the DHB and $^{13}C_6$-DHB consecutive mouse brain slices (Datasets X-Y) we compile a list of matrix-related adducts and m/z in three main steps: matrix-related signal discovery, exogenous-matrix adduct annotation, and endogenous-matrix adduct annotation.

In the matrix-related signal discovery, all ion signals are classified into endogenous, on-sample matrix-related, and off-sample matrix-related. This classification combines spatial and spectral information. The presence of all ions in each sample group (DHB on and off-sample, and $^{13}C_6$-DHB on and off sample) is determined using spatial correlation and absolute mean intensity. This classification was visually validated by mapping it in a 2D UMAP projection of the spatial similarity between ions. We also compute the m/z shift between all possible pairs of DHB and $^{13}C_6$-DHB m/z values (only when the $^{13}C_6$ m/z

is higher). All pairs of ions spatially classified as not endogenous with an m/z shift matching the isotopic labeling (+6Da, +12Da, +18Da) (5 ppm) are considered matrix-related. This discovery is FDR-controlled using a bimodal decoy distribution ($\boldsymbol{\mu}$=N±N/2,σ=0.1). Where N corresponds to the shift used in the target search (+6Da, +12Da, +18Da).

To annotate exogenous-matrix adducts all discovered matrix-related peaks were matched against a database of theoretical DHB adducts (B. O. Keller and Li 2000; Bourcier, Bouchonnet, and Hoppilliard 2001; Wallace, Arnould, and Knochenmuss 2005; Petković et al. 2009), generic positive ion mode adducts (Strohalm et al. 2010; Loos et al. 2015; Huang et al. 1999; Bernd O. Keller et al. 2008) and DHB in-silico fragments predicted with CFM-ID (F. Wang et al. 2021). This search was FDR-controlled using the decoy matrix CHCA.

Finally, to annotate endogenous-matrix adducts we matched the unannotated matrix-related peaks against HMDB (Wishart et al. 2018) considering all exogenous-matrix adducts found. Exact mass searches were conducted using the R package MS2ID (https://github.com/jmbadia/MS2ID) and isotopic pattern matching was conducted using the annotation engine in rMSIcleanup (Baquer et al. 2020). This search was FDR-controlled using the decoy matrix CHCA.

4.4. Annotation of matrix-related signals

rMSIcleanup (Baquer et al. 2020) was used to annotate matrix-related signals in Datasets C1-C14. rMSIcleanup uses the R package enviPat (Loos et al. 2015) to generate the theoretical isotopic patterns of all candidate matrix-related adducts and fragments. Each candidate is matched against the experimental data and given a similarity score (S). S is the product of 3 scores: isotopic pattern similarity (S1, cosine similarity), isotopic spatial correlation (S2, weighted Pearson's correlation), and monoisotopic ion autocorrelation (S3, Moran's I).

The False Discovery Rate (FDR) of all annotations is estimated following a target-decoy approach (Palmer et al. 2017) using a decoy MALDI matrix. In this study, we used CHCA as a decoy matrix given its similar monoisotopic weight and structure to DHB.

rMSIcleanup uses a binary search algorithm instantiated in Rfast (Papadakis et al., n.d.) to perform efficient searches in large datasets.

5. Tables and Figures

Table 1. List of discovered exogenous matrix adducts.

Experimental m/z	Adduct	FDR	Confidence level
159.0053	DHB+Na-H2O	8.33%	+++
177.0159	DHB+Na	8.33%	+++
192.9901	DHB+K	8.33%	+++
198.9980	DHB+2Na-H	8.33%	+++
230.9462	DHB+2K-H	8.33%	+++
273.0406	2DHB-2H2O+H	8.33%	+++
809.0944	5DHB+K	8.33%	+++
174.9791	DHB+K-H2O	11.11%	++
295.0232	2DHB+Na-2H2O	11.11%	++
313.0342	2DHB+Na-H2O	11.11%	++
331.0452	2DHB+Na	11.11%	++
351.9936	[4DHB+K+Na-H2O]2+	11.11%	++

Table 2. List of discovered endogenous matrix adducts.

Experimental m/z	Adduct	FDR	Confidence level
136.5201	DHB+X	13.85%	+
147.5110	DHB+X	13.85%	+
155.0087	DHB+X	13.85%	+
[…] 53 more DHB+X ions			
199.0923	2DHB+X	12.20%	+
245.0451	2DHB+X	12.20%	+
259.9850	2DHB+X	12.20%	+
[…] 33 more 2DHB+X ions			
272.1270	3DHB+X	28.13%	+
313.0148	3DHB+X	28.13%	+
330.0120	3DHB+X	28.13%	+
[…] 20 more 3DHB+X ions			

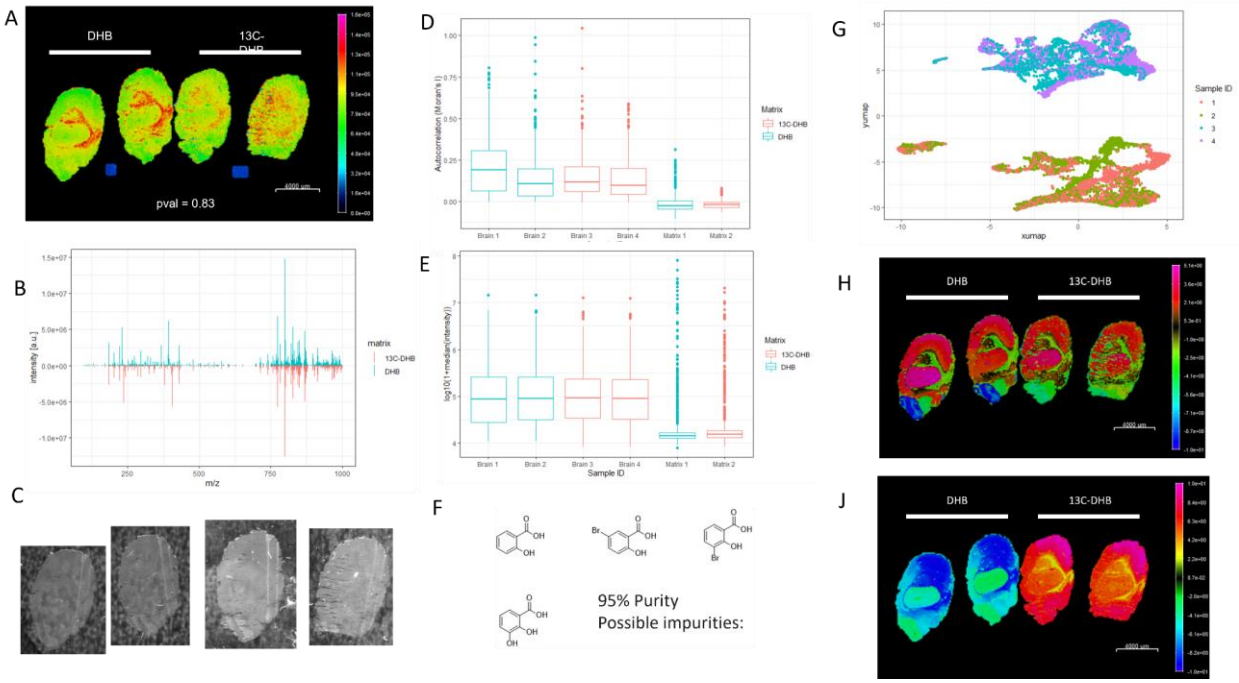


Figure 1 DHB vs *13C*^{6}-DHB MALDI-MSI quality comparison **(A)** TIC spatial distribution **(B)** Mean spectra **(C)** Optical images of the DHB (left) and *13C*^{6}-DHB (right) **(D)** Moran´s I autocorrelation test **(E)** Median intensity per MS feature **(F)** List of possible impurities **(G)** UMAP projection of all pixels in the MALDI-MSI run **(H)** UMAP 1 spatial distribution **(J)** UMAP 2 spatial distribution

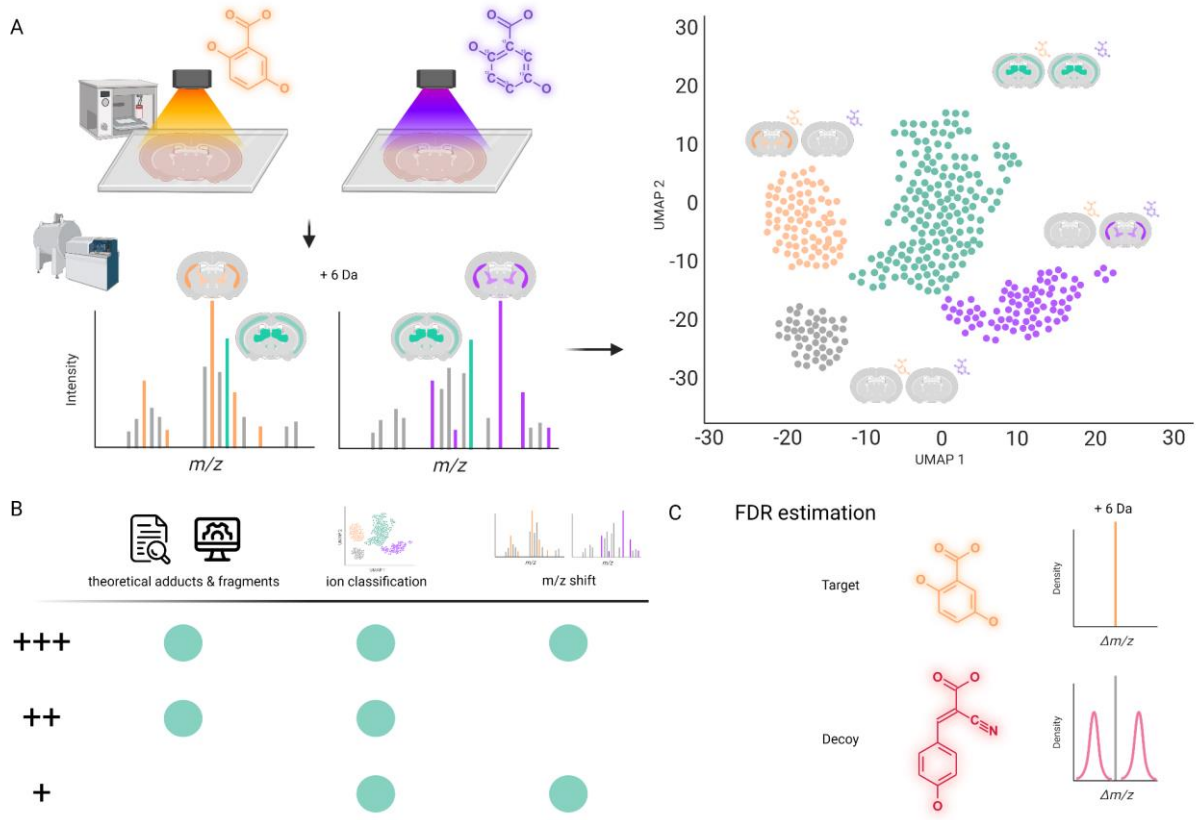


Figure 2 General workflow of matrix-related adducts discovery using SIL-matrix. **(A)** Experimental and computational foundations **(B)** Definition of confidence levels **(C)** FDR estimation based on a decoy matrix and a decoy distribution of m/z shifts

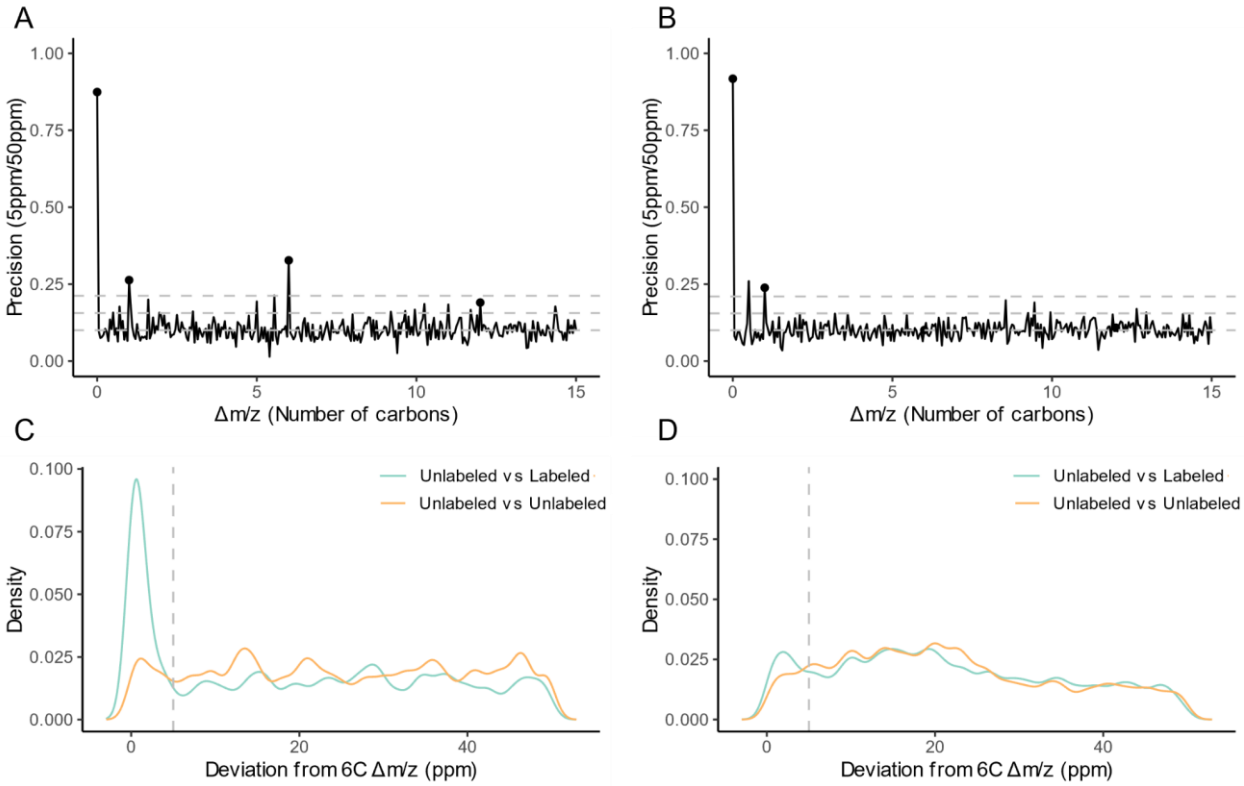


Figure 3 Exploration of m/z shifts **(A)** Precision of m/z shifts in DHB vs *13C*^{6}-DHB **(B)** Precision of m/z shifts in DHB vs DHB **(C)** Density distribution of ppm error of +6 DA m/z shift. **(D)** Density distribution of ppm error of +2 DA m/z shift.

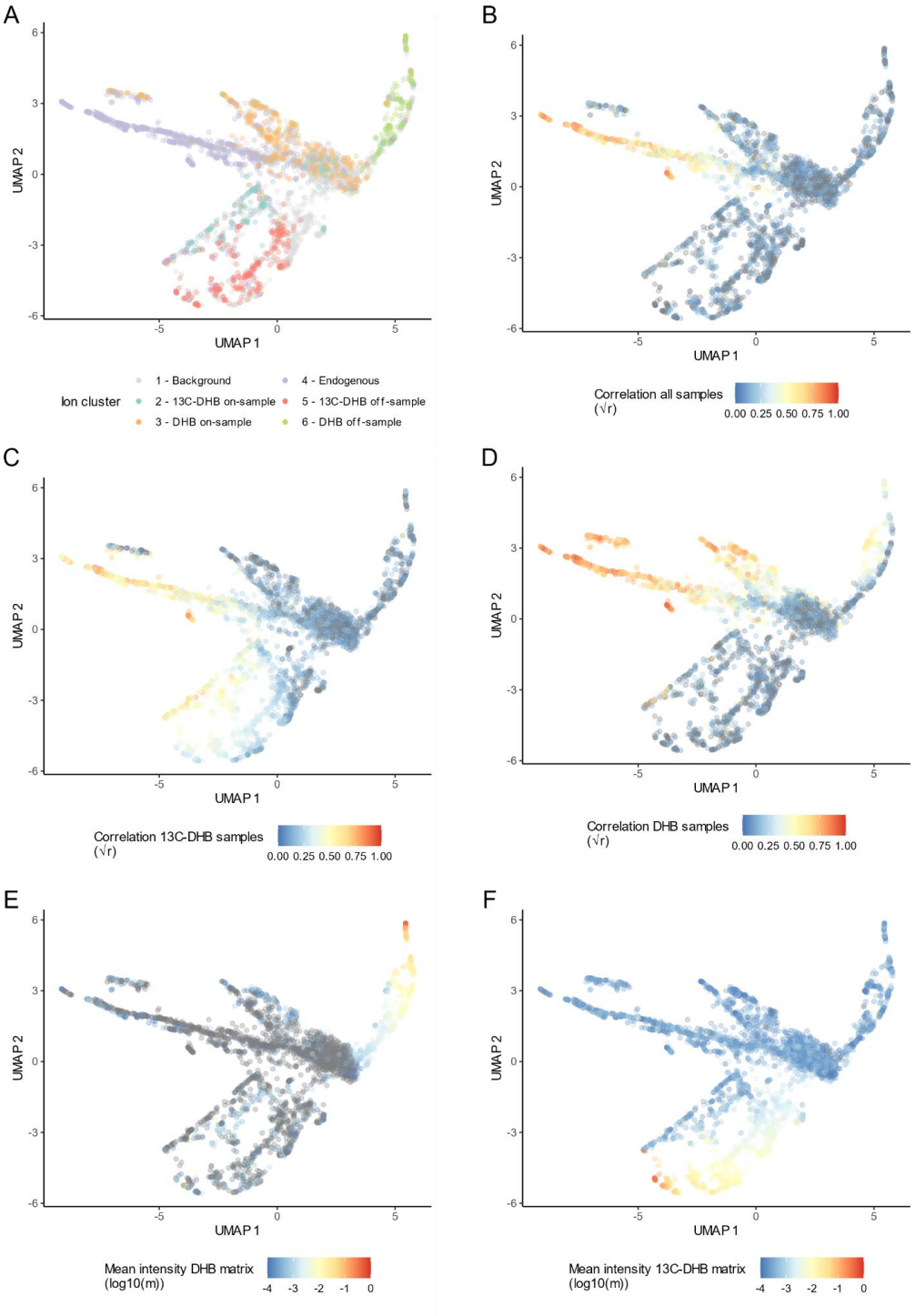


Figure 4. Classification of ion images based on their spatial distribution. **(A)** Classification of each ion image. **(B)** Spatial correlation across all samples, **(C)** *13C*^{6}-DHB samples, and **(D)** DHB samples. **(E)** Mean ion intensity in DHB off-sample region and **(F)** *13C*^{6}-DHB off-sample region.

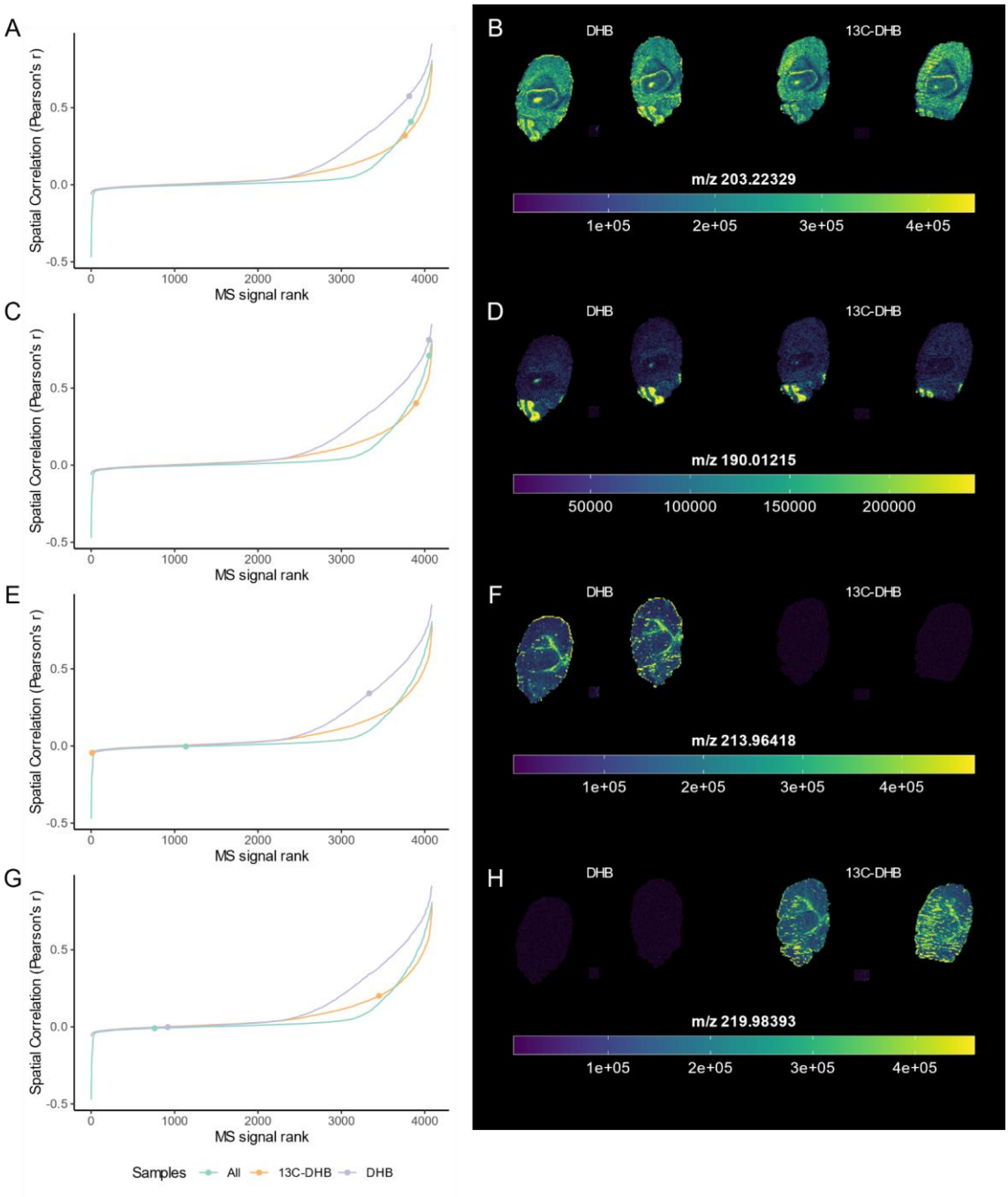


Figure 5. Examples of ion classification. Rank order plot of spatial correlation and ion images for **(A-B)** m/z 203.2233, **(C-D)** m/z 190.0122, **(E-F)** m/z 213.9642, and **(G-H)** m/z 219.9839

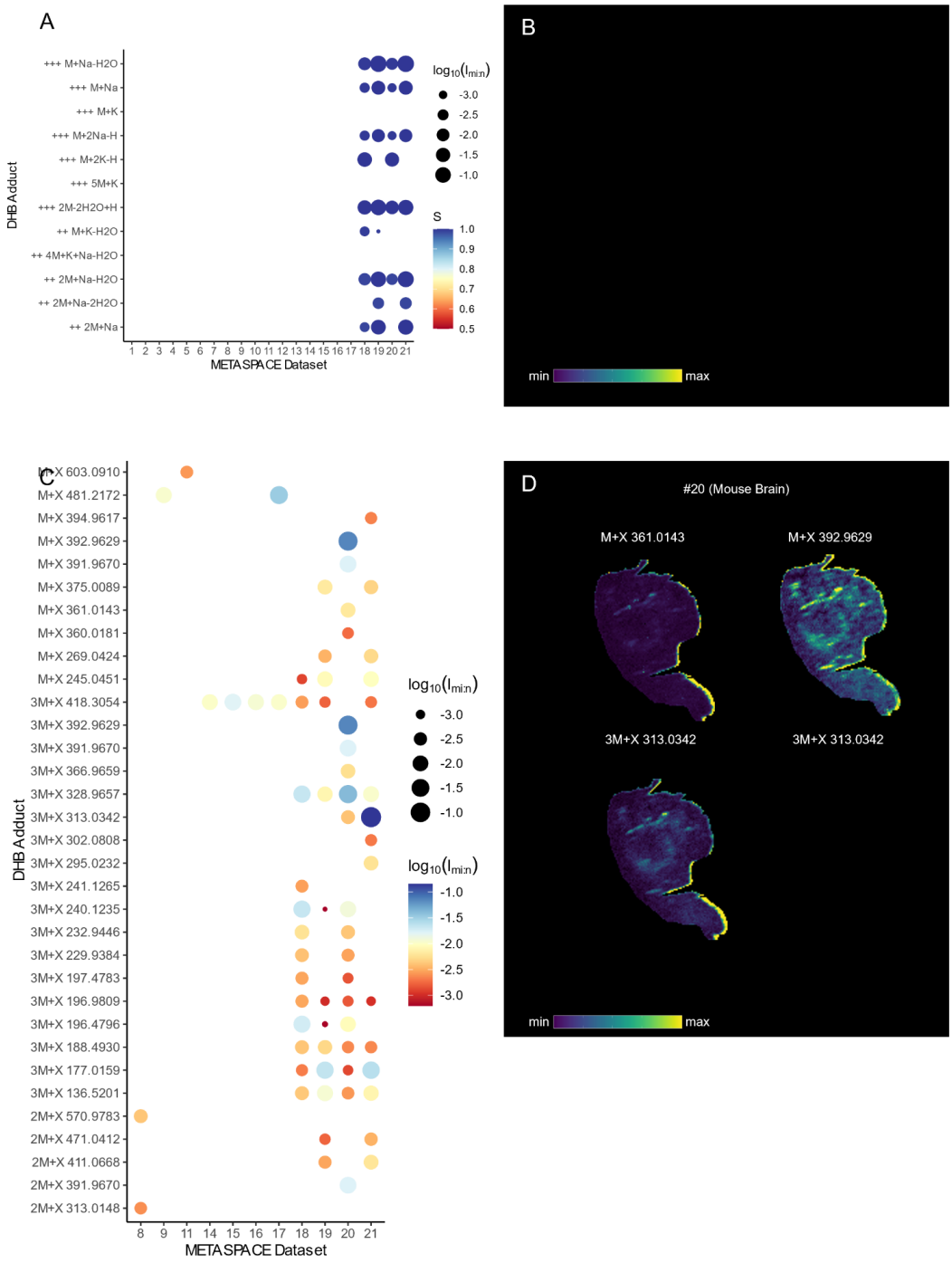


Figure 6. Annotation of mouse brain samples prepared with different matrices. (A) Exogenous matrix-related annotations (+++ and ++). The color indicates the S score (see Methods) and the size indicates the intensity normalized to the maximum peak in the mean spectrum. **(B)** Example ion images. **(C)** Endogenous matrix-related annotations (+), and **(D)** example images.

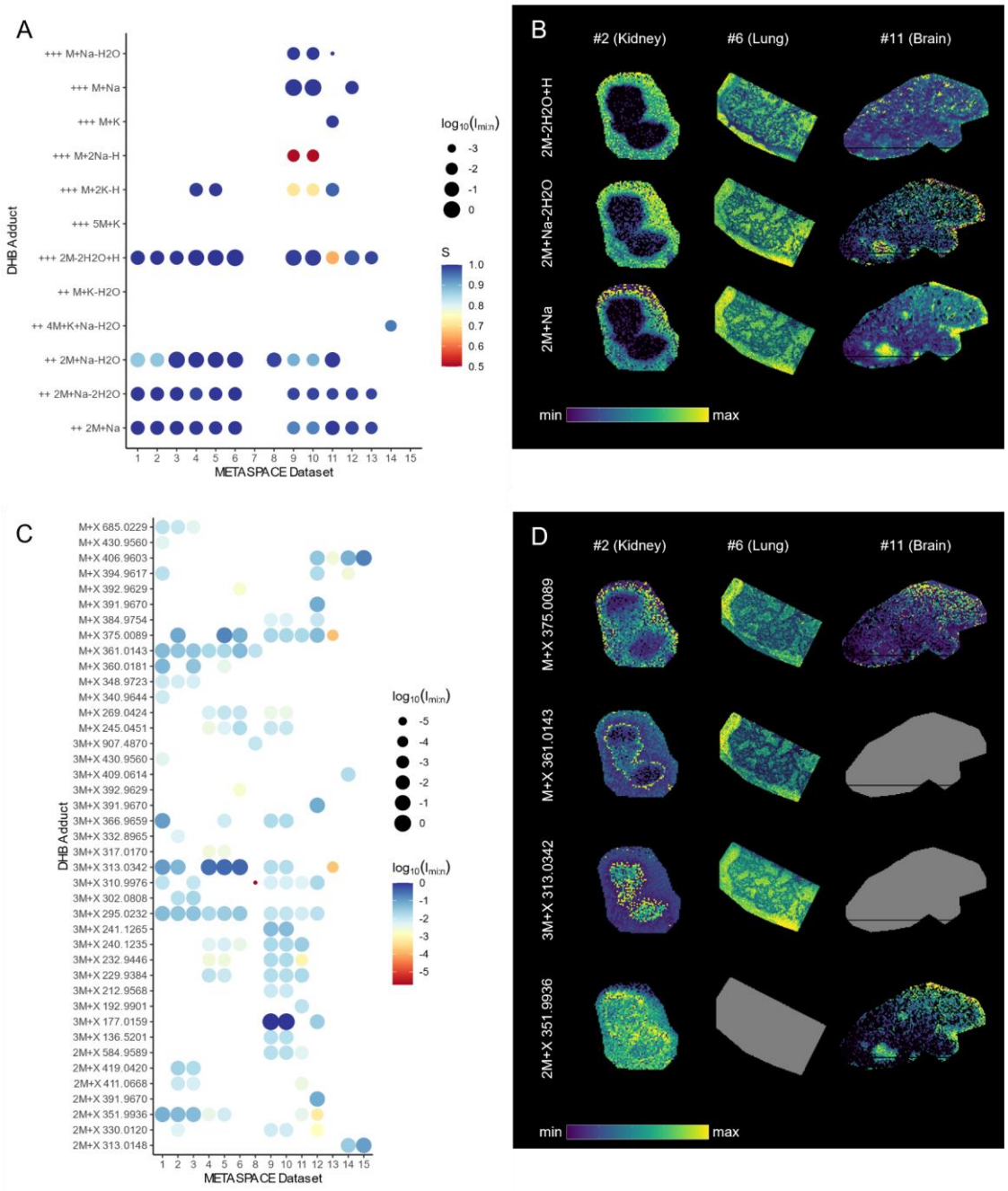


Figure 7. Annotation of human biopsies of different organs prepared with DHB (METASPACE). **(A)** Exogenous matrix-related annotations (+++ and ++). The color indicates the S score and the size indicates the intensity normalized to the maximum peak in the mean spectrum. **(B)** Example ion images. Gray images indicate the given ion is not present in the sample. **(C)** Endogenous matrix-related annotations (+), and **(D)** example ion images.

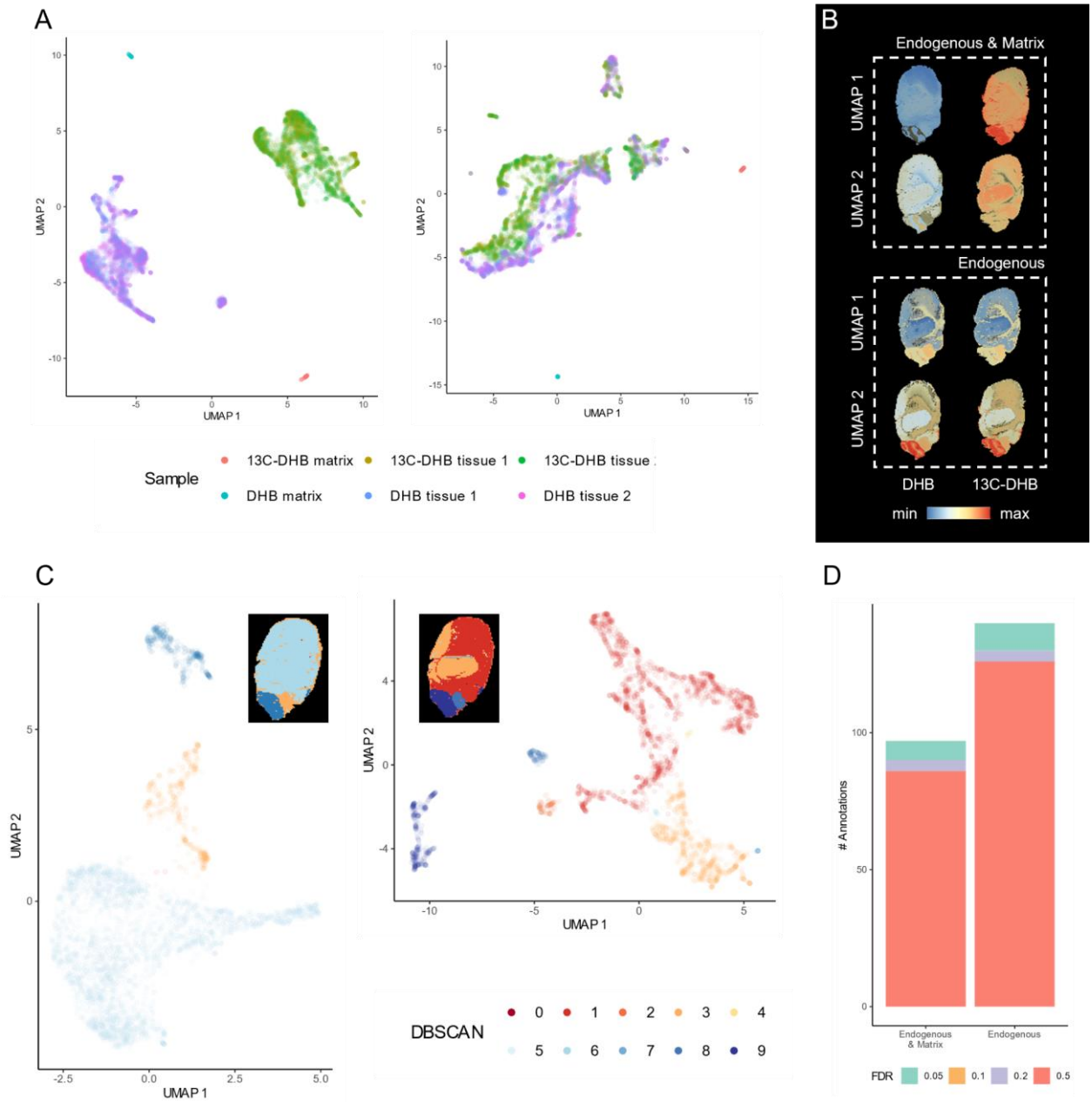


Figure 9. Influence of matrix-related signals in untargeted workflows DHB samples. **(A)** UMAP pixel projection before (left) and after (right) matrix removal **(B)** Spatial UMAP mapping before (top) and after (bottom) matrix removal **(C)** UMAP pixel projection and DBSCAN clustering before (left) and (after) matrix removal **(D)** Number of METASPACE annotations before and after matrix removal.

6. References

Alexandrov, Theodore. 2020. “Spatial Metabolomics and Imaging Mass Spectrometry in the Age of Artificial Intelligence.” *Annual Review of Biomedical Data Science* 3 (July): 61–87.

Alexandrov, Theodore, Katja Ovchinnikova, Andrew Palmer, Vitaly Kovalev, Artem Tarasov, Lachlan Stuart, Renat Nigmetzianov, Dominik Fay, and Key Metaspace Contributors. 2019. “METASPACE: A Community-Populated Knowledge Base of Spatial Metabolomes in Health and Disease.” *bioRxiv*. https://doi.org/10.1101/539478.

Baquer, Gerard, Lluc Sementé, María García-Altares, Young Jin Lee, Pierre Chaurand, Xavier Correig, and Pere Ràfols. 2020. “rMSIcleanup: An Open-Source Tool for Matrix-Related Peak Annotation in Mass Spectrometry Imaging and Its Application to Silver-Assisted Laser Desorption/ionization.” *Journal of Cheminformatics* 12 (1): 45.

Baquer, Gerard, Lluc Sementé, Toufik Mahamdi, Xavier Correig, Pere Ràfols, and María García-Altares. 2022. “What Are We Imaging? Software Tools and Experimental Strategies for Annotation and Identification of Small Molecules in Mass Spectrometry Imaging.” *Mass Spectrometry Reviews*, July, e21794.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society* 57 (1): 289–300.

Bourcier, Sophie, Stéphane Bouchonnet, and Yannik Hoppilliard. 2001. “Ionization of 2,5-Dihydroxybenzoic Acid (DHB) Matrix-Assisted Laser Desorption Ionization Experiments and Theoretical Study.” *International Journal of Mass Spectrometry*. https://doi.org/10.1016/s1387-3806(01)00446-8.

Calvano, Cosima Damiana, Antonio Monopoli, Tommaso R. I. Cataldi, and Francesco Palmisano. 2018. “MALDI Matrices for Low Molecular Weight Compounds: An Endless Story?” *Analytical and Bioanalytical Chemistry* 410 (17): 4015–38.

Chaurand, Pierre, Jeremy L. Norris, D. Shannon Cornett, James A. Mobley, and Richard M. Caprioli. 2006. “New Developments in Profiling and Imaging of Proteins from Tissue Sections by MALDI Mass Spectrometry.” *Journal of Proteome Research* 5 (11): 2889–2900.

Coy, Shannon, Shu Wang, Sylwia A. Stopka, Jia-Ren Lin, Clarence Yapp, Cecily C. Ritch, Lisa Salhi, et al. 2022. “Single Cell Spatial Analysis Reveals the Topology of Immunomodulatory Purinergic Signaling in Glioblastoma.” *Nature Communications* 13 (1): 4814.

Dong, Wei, Qing Shen, Joewel T. Baibado, Yimin Liang, Ping Wang, Yeqing Huang, Zhifeng Zhang, Yixuan Wang, and Hon-Yeung Cheung. 2013. “Phospholipid Analyses by MALDI-TOF/TOF Mass Spectrometry Using 1,5-Diaminonaphthalene as Matrix.” *International Journal of Mass Spectrometry* 343-344 (June): 15–22.

Elias, Joshua E., and Steven P. Gygi. 2007. “Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry.” *Nature Methods* 4 (3): 207–14.

Gao, Si-Qi, Jin-Hui Zhao, Yue Guan, Ying-Shu Tang, Ying Li, and Li-Yan Liu. 2022. “Mass Spectrometry Imaging Technology in Metabolomics: A Systematic Review.” *Biomedical Chromatography: BMC*, August, e5494.

Greer, Tyler, Robert Sturm, and Lingjun Li. 2011. “Mass Spectrometry Imaging for Drugs and Metabolites.” *Journal of Proteomics* 74 (12): 2617–31.

Guan, Ming, Zhen Zhang, Shilei Li, Jian ’an Liu, Lu Liu, Hui Yang, Yangyang Zhang, Tie Wang, and Zhenwen Zhao. 2018. “Silver Nanoparticles as Matrix for MALDI FTICR MS Profiling and Imaging of Diverse Lipids in Brain.” *Talanta* 179: 624–31.

Hahsler, Michael, Matthew Piekenbrock, and Derek Doran. 2019. “Dbscan: Fast Density-Based Clustering with R.” *Journal of Statistical Software* 91: 1–30.

Huang, Nelson, Marshall M. Siegel, Gary H. Kruppa, and Frank H. Laukien. 1999. “Automation of a Fourier Transform Ion Cyclotron Resonance Mass Spectrometer for Acquisition, Analysis, and E-Mailing of High-Resolution Exact-Mass Electrospray Ionization Mass Spectral Data.” *Journal of the American Society for Mass Spectrometry*. https://doi.org/10.1016/s1044-0305(99)00089-6.

Iakab, Stefania-Alexandra, Gerard Baquer, Marta Lafuente, Maria Pilar Pina, José Luis Ramírez, Pere Ràfols, Xavier Correig-Blanchar, and María García-Altares. 2022. “SALDI-MS and SERS Multimodal Imaging: One Nanostructured Substrate to Rule Them Both.” *Analytical Chemistry* 94 (6): 2785–93.

Janda, Moritz, Brandon K. B. Seah, Dennis Jakob, Janine Beckmann, Benedikt Geier, and Manuel Liebeke. 2021. “Determination of Abundant Metabolite Matrix Adducts Illuminates the Dark Metabolome of MALDI-Mass Spectrometry Imaging Datasets.” *Analytical Chemistry* 93 (24): 8399–8407.

Keller, Bernd O., Jie Sui, Alex B. Young, and Randy M. Whittal. 2008. “Interferences and Contaminants Encountered in Modern Mass Spectrometry.” *Analytica Chimica Acta* 627 (1): 71–81.

Keller, B. O., and L. Li. 2000. “Discerning Matrix-Cluster Peaks in Matrix-Assisted Laser Desorption/ionization Time-of-Flight Mass Spectra of Dilute Peptide Mixtures.” *Journal of the American Society for Mass Spectrometry* 11 (1): 88–93.

Loos, Martin, Christian Gerber, Francesco Corona, Juliane Hollender, and Heinz Singer. 2015. “Accelerated Isotope Fine Structure Calculation Using Pruned Transition Trees.” *Analytical Chemistry* 87 (11): 5738–44.

McInnes, Leland, John Healy, and James Melville. 2018. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1802.03426.

Melville, J., A. Lun, and M. Djekidel. 2020. “Uwot: The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction. R Package Version 0.1. 8.”

Notarangelo, Giulia, Jessica B. Spinelli, Elizabeth M. Perez, Gregory J. Baker, Kiran Kurmi, Ilaria Elia, Sylwia A. Stopka, et al. 2022. “Oncometabolite D-2HG Alters T Cell Metabolism to Impair CD8+ T Cell Function.” *Science* 377 (6614): 1519–29.

Ovchinnikova, Katja, Vitaly Kovalev, Lachlan Stuart, and Theodore Alexandrov. 2020. “OffsampleAI: Artificial Intelligence Approach to Recognize off-Sample Mass Spectrometry Images.” *BMC Bioinformatics* 21 (1): 129.

Palmer, Andrew, Prasad Phapale, Ilya Chernyavsky, Regis Lavigne, Dominik Fay, Artem Tarasov, Vitaly Kovalev, et al. 2017. “FDR-Controlled Metabolite Annotation for High-Resolution Imaging Mass Spectrometry.” *Nature Methods* 14 (1): 57–60.

Papadakis, M., M. Tsagris, M. Dimitriadis, and S. Fafalios. n.d. “Rfast: A Collection of Efficient and Extremely Fast R Functions.” *R Package Version*.

Petković, Marijana, Jürgen Schiller, Matthias Müller, Rosmarie Süß, Klaus Arnold, and Jürgen Arnhold. 2009. “Detection of Adducts with Matrix Clusters in the Positive and Negative Ion Mode MALDI-TOF Mass Spectra of Phospholipids.” *Zeitschrift Für Naturforschung B* 64 (3): 331–34.

Pinheiro, J. 2009. “Nlme : Linear and Nonlinear Mixed Effects Models. R Package Version 3.1-96.” *Http://cran.r-Project.org/web/packages/nlme/*. https://ci.nii.ac.jp/naid/10029283579/.

Ràfols, Pere, Bram Heijs, Esteban Del Castillo, Oscar Yanes, Liam A. McDonnell, Jesús Brezmes, Iara Pérez-Taboada, Mario Vallejo, María García-Altares, and Xavier Correig. 2020. “RMSIproc: An R Package for Mass Spectrometry Imaging Data Processing.” *Bioinformatics* 36 (11): 3618–19.

Ràfols, Pere, Dídac Vilalta, Sònia Torres, Raul Calavia, Bram Heijs, Liam A. McDonnell, Jesús Brezmes, et al. 2018. “Assessing the Potential of Sputtered Gold Nanolayers in Mass Spectrometry Imaging for Metabolomics Applications.” *PloS One* 13 (12): e0208908.

Rohner, Tatiana C., Dieter Staab, and Markus Stoeckli. 2005. “MALDI Mass Spectrometric Imaging of Biological Tissue Sections.” *Mechanisms of Ageing and Development* 126 (1): 177–85.

Schubert, Erich, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. 2017. “DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN.” *ACM Trans. Database Syst.*, 19, 42 (3): 1–21.

Shariatgorji, Mohammadreza, Anna Nilsson, Richard J. A. Goodwin, Per Svenningsson, Nicoletta Schintu, Zoltan Banka, Laszlo Kladni, Tibor Hasko, Andras Szabo, and Per E. Andren. 2012. “Deuterated Matrix-Assisted Laser Desorption Ionization Matrix Uncovers Masked Mass Spectrometry Imaging Signals of Small Molecules.” *Analytical Chemistry* 84 (16): 7152–57.

Smets, Tina, Nico Verbeeck, Marc Claesen, Arndt Asperger, Gerard Griffioen, Thomas Tousseyn, Wim Waelput, Etienne Waelkens, and Bart De Moor. 2019. “Evaluation of Distance Metrics and Spatial Autocorrelation in Uniform Manifold Approximation and Projection Applied to Mass Spectrometry Imaging Data.” *Analytical Chemistry* 91 (9): 5706–14.

Strohalm, Martin, Daniel Kavan, Petr Novák, Michael Volný, and Vladimír Havlícek. 2010. “mMass 3: A Cross-Platform Software Environment for Precise Analysis of Mass Spectrometric Data.” *Analytical Chemistry* 82 (11): 4648–51.

Taylor, Adam J., Alex Dexter, and Josephine Bunch. 2018. “Exploring Ion Suppression in Mass Spectrometry Imaging of a Heterogeneous Tissue.” *Analytical Chemistry* 90 (9): 5637–45.

Vermillion-Salsbury, Rachal L., and David M. Hercules. 2002. “9-Aminoacridine as a Matrix for Negative Mode Matrix-Assisted Laser Desorption/ionization.” *Rapid Communications in Mass Spectrometry: RCM* 16 (16): 1575–81.

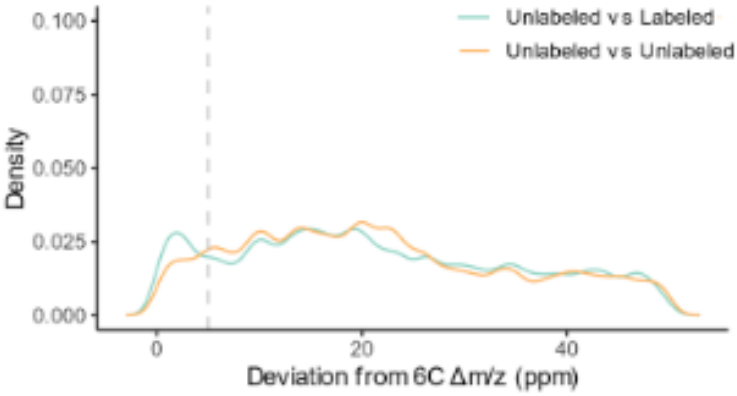
Wallace, W. E., M. A. Arnould, and R. Knochenmuss. 2005. “2,5-Dihydroxybenzoic Acid: Laser Desorption/ionisation as a Function of Elevated Temperature.” *International Journal of Mass Spectrometry*. https://doi.org/10.1016/j.ijms.2004.11.011.

Wang, Fei, Jaanus Liigand, Siyang Tian, David Arndt, Russell Greiner, and David S. Wishart. 2021. “CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification.” *Analytical Chemistry* 93 (34): 11692–700.

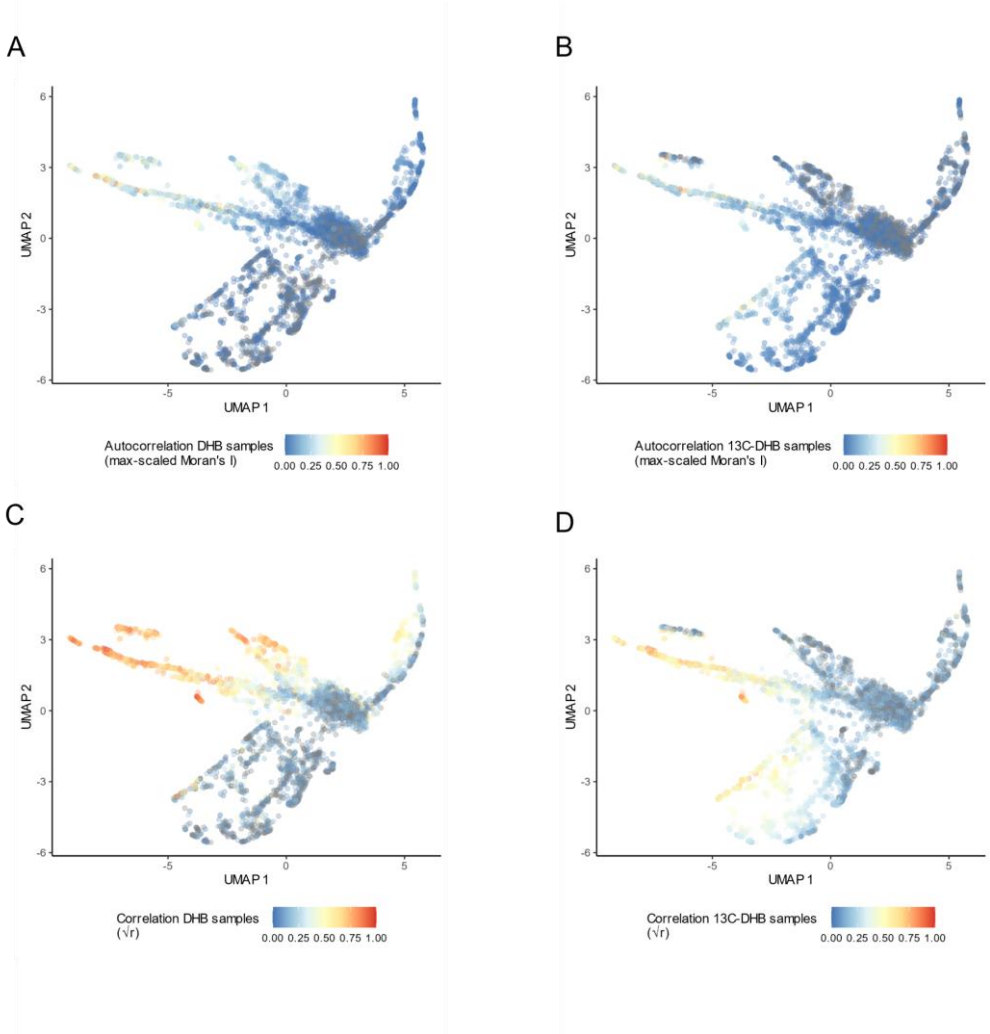
Wang, Lin, Dan Wang, Olmo Sonzogni, Shizhong Ke, Qi Wang, Abhishek Thavamani, Felipe Batalini, et al. 2022. “PARP-Inhibition Reprograms Macrophages toward an Anti-Tumor Phenotype.” *Cell Reports* 41 (2): 111462.

Wishart, David S., Yannick Djoumbou Feunang, Ana Marcu, An Chi Guo, Kevin Liang, Rosa Vázquez-Fresno, Tanvir Sajed, et al. 2018. “HMDB 4.0: The Human Metabolome Database for 2018.” *Nucleic Acids Research* 46 (D1): D608–17.

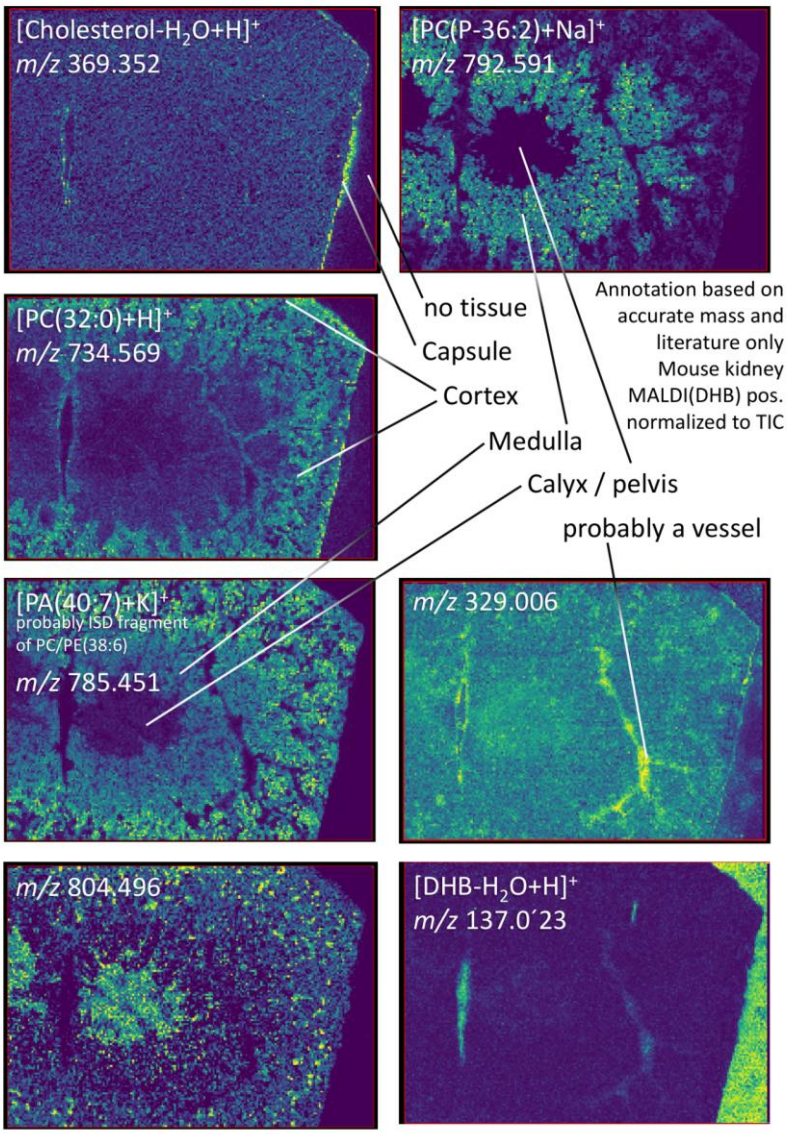
7. Supplementary Materials



Supplementary Figure 1. Density distribution of ppm error of +2 DA m/z shift.



Supplementary Figure 2. Comparison between autocorrelation of a single sample and spatial correlation between all samples within one group.



Supplementary Figure 3. Example molecular and anatomical annotations for Dataset D1

Supplementary Table 1. List of the 25 MALDI MSI datasets used for method development and validation. Sample type, sample preparation, and MALDI-MSI acquisition parameters.

No.	**Species**	**Tissue type**	**Matrix deposition**	**Lateral Res. (um)**	***m/z* range**	**Mass spectrometer**	**Acq. Mode**
A1-A2	*Mus musculus*	Brain	DHB, TM sprayer	25 µm	100-1000	9.4T Solarix FTICR (Bruker)	Positive/Profile
A3-A4	*Mus musculus*	Brain	$^{13}C_6$-DHB, TM sprayer	25 µm	100-1000	9.4T Solarix FTICR (Bruker)	Positive/Profile
B1-B2	*Mus musculus*	Brain	9AA, TM sprayer	25 µm	100-1000	9.4T Solarix FTICR (Bruker)	Negative/Profile
B3-B4	-	Matrix control	9AA, TM sprayer	25 µm	100-1000	9.4T Solarix FTICR (Bruker)	Negative/Profile
B5-B6	*Mus musculus*	Brain	NEDC, TM sprayer	25 µm	100-1000	9.4T Solarix FTICR (Bruker)	Negative/Profile
B7-B8	-	Matrix control	NEDC, TM sprayer	25 µm	100-1000	9.4T Solarix FTICR (Bruker)	Negative/Profile

B9-B10	*Mus musculus*	Brain	NOR, TM sprayer	25 µm	100-1000	9.4T Solarix FTICR (Bruker)	Negative/Profile
B11-B12	-	Matrix control	NOR, TM sprayer	25 µm	100-1000	9.4T Solarix FTICR (Bruker)	Negative/Profile
B13-B14	*Mus musculus*	Brain	Au, TM sprayer	25 µm	100-1000	9.4T Solarix FTICR (Bruker)	Positive/Profile
B15-B16	-	Matrix control	Au, TM sprayer	25 µm	100-1000	9.4T Solarix FTICR (Bruker)	Positive/Profile
B17-B18	*Mus musculus*	Brain	DHB, TM sprayer	25 µm	100-1000	9.4T Solarix FTICR (Bruker)	Positive/Profile
B19-B20	-	Matrix control	DHB, TM sprayer	25 µm	100-1000	9.4T Solarix FTICR (Bruker)	Positive/Profile
D1	*Mus musculus*	Kidney (20 µm)	DHB, sublimated and re-cristallized	25 µm	100-1000	Spectroglyph, modified (7.5 torr)/ Orbitrap Q Exactive Plus	Positive/Profile

Supplementary Table 2. List of the 14 METASPACE (Alexandrov et al. 2019) MALDI MSI datasets used for validation. Sample type, sample preparation, and MALDI-MSI acquisition parameters.

No.	**Species**	**Tissue type**	**Matrix deposition**	**Lateral Res. (um)**	***m/z* range**	**Mass spectrometer**	**Acq. Mode**	**Contributor**
C1-C4	*Homo Sapiens*	Liver	DHB, TM sprayer	N/A	300-2000	TOF	Positive/Centroid	Denis Abu Sammour (HS Mannheim)
C5-C6	*Homo Sapiens*	Liver	DHB, TM sprayer	N/A	150-1850	FTICR	Positive/Centroid	Denis Abu Sammour (HS Mannheim)
C7-C8	*Homo Sapiens*	Brain	DHB, TM sprayer	N/A	100-1150	FTICR	Positive/Centroid	Elisa Ruhland (IBMP)
C9-C11	*Homo Sapiens*	Lung	DHB, HTX M5 Sprayer	35 µm	200-1200	FTICR	Positive/Centroid	Brittney Gorman (PNNL)
C12-C14	*Homo Sapiens*	Kidney	DHB, TM sprayer	35 µm	200-1300	FTICR	Positive/Centroid	Jessica Lukowski(PNNL)

CHAPTER 6:

Final discussion and conclusions

This thesis has revolved around the study and annotation of two types of MS signals traditionally underestimated and overlooked in MALDI-MSI: in-source ion fragments and MALDI matrix-adducts ions. To address these issues we have presented two automatic annotation tools: rMSIfragment and rMSIcleanup, that are intended to identify and annotate these ions, respectively. We have also developed a full experimental and computational workflow based on SIL-MALDI matrices to discover de-novo matrix-related adducts. This method allows the complete characterization of MALDI matrix adducts under different matrices, tissue types, ion sources, and MS analyzers.

In general, we have found that the correct annotation and handling of in-source fragments and matrix-related signals should be prioritized. These signals negatively affect common computational approaches like dimensionality reduction (PCA, UMAP), and the annotation of endogenous metabolites and lipids and produces erroneous compound identifications.

As a derived result of the annotation methods developed, we also made significant contributions to the broader field of software tools for MALDI-MSI in several ways. First, we proposed a novel iterative biclustering algorithm capable of detecting overlapped MS features (Baquer et al. 2020). This is of particular interest in TOF and other MS analyzers with moderate resolving power and a higher prevalence of overlapped MS features. We next introduced several new target-decoy strategies to estimate annotation confidence and performance using highly unlikely compounds, in-source fragmentation pathways, MALDI matrices, and m/z shifts. We also developed computational methods to register and fuse MSI with complementary imaging modalities such as Raman Imaging (Iakab et al. 2022), multiplexed tissue immunofluorescence (Notarangelo et al. 2022; Coy et al. 2022), and immunohistochemistry (L. Wang et al. 2022). Finally, we made contributions at an MSI community level by proposing two new standards (Baquer et al. 2022). We proposed an adaptation of the metabolomics levels of confidence (Schymanski et al. 2014) to MALDI-MSI and we also drafted an integration of the mzTab-M (Hoffmann et al. 2019) format with .imZML (Schramm et al. 2012) in an attempt to standardize the output of different annotation software.

1. Annotation of lipid in-source fragments with rMSIfragment

We have demonstrated the performance of rMSIfragment on 15 human nevi datasets with two orthogonal approaches: (1) matching its annotations to HPLC and (2) using a target-decoy approach. Both approaches yield similar performance estimations (0.7 AUC and 0.6 AUC for the samples acquired in negative and positive mode respectively).

As a next step, we deployed rMSIfragment to annotate lipids and their fragments in 12 publicly available samples covering a wide combination of samples and experimental setups. The performances obtained are comparable and often better than the ones obtained on the human nevi datasets. Additionally, rMSIcleanup shows a high lipidome coverage overlap when compared to available annotation tools like METASPACE (Alexandrov et al. 2019). Collectively, these results indicate that rMSIfragment is capable of reliably annotating lipids and their in-source fragments across multiple experimental conditions

One key highlight of our study is the importance of considering in-source fragmentation pathways when performing molecular annotation. We have found that overlooking ISD pathways, can lead to up to 75% of the reported lipid annotations to be overlapped with at least one in-source fragment. rMSIfragment mitigates this issue through two mechanisms: (1) unlikely lipids with low occurrences (number of adducts and in-source fragments) and poor spatial correlation are filtered out, and (2) the user is aware of the overlap, allowing them to be cautious with their interpretation of the automated annotations.

We envision three new avenues to further improve the automatic annotation of lipids and their in-source fragments: (1) exploiting known ion suppression effects between different lipid classes, (2) exploiting MS/MS libraries, and (3) compiling MALDI-ISD or MALDI-MS/MS libraries.

Ion suppression effects strongly favor certain classes of lipids, difficulting the analysis of suppressed species (Boskamp and Soltwisch 2020). In positive mode, for instance, phosphatidylcholines (PC) display a strong signal in detriment to lower signals of phosphatidylethanolamines (PE) or phosphatidylserines (PS), phosphatidylglycerol (PG) or phosphatidylinositol (PI). In negative mode the effect is reversed, PC species show lower signals in samples containing other lipid species in favor of other lipid species. These interactions have been characterized in the past (Boskamp and Soltwisch 2020) and could be considered to define a new ranking score to filter out unlikely lipid annotations with intensity values that contradict them.

Initially, we approached the annotation of MALDI-MSI in-source fragmentation by modeling and exploiting similarities of MALDI-MSI spectra to publicly available MS/MS libraries. These approaches were inspired by recent advancements in the LC-MS community (Xue et al. 2020), where the in-source fragmentation in an ESI source is enhanced to yield fragmentation patterns similar to those present in METLIN to aid the quick identification of metabolites. We compared individual MALDI-MSI spectra to MS/MS fragmentation spectra available in MS2ID (https://github.com/jmbadia/MS2ID) and in-silico fragmentation algorithms such as MetFrag (Ruttkies et al. 2016), CFM-ID (F. Wang et al. 2021) and Sirius (Dührkop et al. 2019). The preliminary results suggested that the use of MS/MS fragmentation spectra available in databases and in-silico tools was not sufficient to annotate in-source fragmentation in MALDI-MSI confidently. The two main limitations are (1) the different ion sources used (MALDI vs ESI) and (2) the underrepresentation of common MALDI adducts such as M+Na and M+K (< 10%) in MS/MS libraries. This is of particular interest given that different adducts can yield different fragmentation patterns (Fuchs et al. 2007).

To overcome these limitations, one interesting avenue would be the compilation of MALDI-ISD libraries. This community-wide effort would help better characterize MALDI-ISD in a wide range of biomolecules. A MALDI-MS/MS library, perhaps a more pressing interest of the MALDI community, could already provide a lot of information due to the common ionization mechanisms. These two libraries would be invaluable tools to foster the development of the next generation of ISD annotation algorithms in MALDI-MSI.

In conclusion, neglecting in-source fragmentation leads to an increased number of false lipid annotations. rMSIfragment mitigates this effect by prioritizing annotations of lipids found forming multiple adducts and in-source fragments.

2. Annotation of Ag-related signals with rMSIcleanup

The goal of this study was to develop, optimize and validate a new algorithm to annotate signals attributed to the LDI promoting material in MSI. The developed algorithm is packaged and released as rMSIcleanup, an open-source R package freely available for the scientific community and fully integrated with rMSIproc [20], a stand-alone package for the visualization, pre-processing and analysis of MSI datasets.

As demonstrated, the widely used “blank subtraction” approach is outperformed by rMSIcleanup in the annotation Ag-related signals. In comparison to the top-performing alternatives for matrix-related peak annotation which are based on machine and deep learning [11], rMSIcleanup has the main advantage of using two intuitive scores (accounting for the isotopic ratios of clusters and the spatial distribution of their ions) and providing a visual justification of each annotation. This is a key contribution as it helps overcome the black-box problem, increases the user’s confidence in the annotation and can help researchers optimize experimental workflows (for instance, choosing LDI promoters that minimize interferences in the *m/z* range of interest). Another merit of our work is that, to our knowledge, it is the first matrix signal annotation algorithm to explicitly detect and deal with overlapping MS signals, which successfully prevents overlapped peaks from being misclassified. Given that we follow a targeted analytical approach, our classification is focused only on matrix-related signals while the algorithms presented by Ovchinnikova et al. [11] have a broader scope and also classify as off-sample other exogenous compounds. In the era of big data, these two apparently opposite approaches (namely our analytical approach based on chemical similarity scores and their untargeted approach based on machine learning) must not only coexist but also complement each other following the trend already initiated in other fields [36]. This reality urges the MSI community to develop annotation algorithms capable of, not only exploiting the knowledge in the increasingly large amounts of MSI datasets available, but also incorporating metrics that take into account the chemical context of the sample to aid transparent justification.

AgLDI MSI was chosen to validate the algorithm, due to the well-understood ionization of silver. A “validation list” was compiled from the literature, which included silver clusters that should be present in all samples and silver adducts or compounds that should not be present in any of them. Given the heterogeneity of the samples used in this study, the described validation list was adapted to each dataset. For each dataset, those clusters in the validation list for which the experimental data contained none of their theoretical masses were excluded. These adjustments in the validation list prevented an overestimation of the performance of the algorithm attributed to a high number of correctly classified “negative class” clusters (i.e. true negatives) located in mass ranges with no signal. We propose this validation strategy as a novel alternative to more common validation approaches such as chemical standards [6] or expert annotation [11, 32]. This study adds to previous work [6, 14, 17, 29, 37] and further demonstrates the potentiality of AgLDI MS imaging, a thriving technology known for its reduced background signals in spatial metabolomics that is strongly complemented by our annotation algorithm as it further removes the influence of the matrix.

In agreement with previous work on the effects of MSI data reduction [35], we have demonstrated that the annotation and removal of signals related to the LDI promoting material used can further enhance post-processing, due to the elimination of variables

attributed to exogenous compounds that do not reflect the morphology nor chemical composition of the sample. These results highlight the need to include software annotation tools such as rMSIcleanup in MSI workflows before exploring the datasets with classical data analysis techniques used in metabolomics. Here we would like to emphasize the need for a standardized quantitative metric to assess the quality of MSI images and we acknowledge the relevance of standardization initiatives such as the MALDISTAR project (www.maldistar.org).

We envision two main applications for rMSIcleanup. On the one hand, it can be used in a purely exploratory fashion to better understand ionization and adduct cluster formation in new matrices, tissues and applications. In this case, the user is advised to add a long list of potential adducts or neutral losses to assess their formation. The validation approach followed in this paper is a clear example of this exploratory application of rMSIcleanup. A second application is the automated peak annotation of well-known matrices and tissues. In this case, only the clusters that are known to be formed need to be given to the software. This curated selection increases the data-processing speed. The set of matrix-related annotated peaks can then be eliminated from the dataset prior to performing post-processing workflows such as multivariate statistical analysis. In any case, the choice of adducts and neutral losses to consider (or matrix adducts with endogenous compounds, e.g. fatty acids + Ag) is application dependent and is therefore left to the user. This list must be manually specified as an input parameter to rMSIcleanup.

Finally, the promising results obtained in the annotation of Ag_n^+-related peaks in AgLDI MSI open the door to the extension of this methodology to more widely used matrices such as 2,5-Dihydroxybenzoic acid (DHB), 1,5-Diaminonaphthalene (DAN), and 9-Aminoacridine (9AA) among others. These organic matrices pose greater challenges. Firstly, they lead to increased matrix background due to their greater fragmentation and adduct formation [38–40] and the higher quantities in which they are added [39]. Moreover, they present the problem of “hot spot” formation given their less homogeneous application process [41]. These issues highlight not only the benefits of AgLDI MSI but also that matrix-related peak annotation can benefit data post-processing even further in applications using organic matrices.

3. Discovery and annotation of matrix-related signals with SIL-MALDI matrix

We presented a novel experimental and computational workflow to discover matrix-related signals using SIL-MALDI-MSI based on the synthesis of a new DHB matrix, in which all the carbons of the aromatic ring have been replaced by 13C. The only previous work with a labeled matrix used a deuterated CHCAI matrix to uncover endogenous metabolites previously selected using a targeted approach (Shariatgorji et al. 2012).

We demonstrate that focusing on spatial (Ovchinnikova et al. 2020) or spectral information (Strohalm et al. 2010) alone is not enough to gain a comprehensive picture of the prevalence of matrix-related signals. The matrix forms adducts with both exogenous and endogenous compounds and it is thus present on and off-sample and lacks distinct matrix-like spatial distribution. This issue is further amplified by ion suppression effects as different molecular environments and tissue types can lead to diverse spatial patterns. Focusing on spectral information is not enough either as this can potentially add false positives with isomeric endogenous formulas. The spatial

distribution helps us discern between endogenous and matrix related. Thus, the annotation of matrix-related signals requires the integration of spatial and spectral information.

In consonance with previous studies (Janda et al. 2021) we found the number of matrix adducts with endogenous metabolites to be non-negligible. In this regard, to ensure confident annotation we introduce a novel FDR estimation paradigm based on decoy matrices and m/z shifts. This is a critical part of our workflow that enables us to work with higher levels of confidence and control the FDR.

A key finding of this work is the realization that matrix-related signals worsen the performance of typical untargeted. We found that the removal of matrix-related signals helps dimensionality reduction algorithms like UMAP (McInnes, Healy, and Melville 2018) better focus on biologically and anatomically relevant structures. Additionally, the removal of matrix-related signals also helps with the annotation of small molecules using automated tools like METASPACE (Alexandrov et al. 2019). Excluding matrix-related signals leads to an overall higher number of annotations with higher confidence (lower FDR). Echoing and expanding on previous studies (Baquer et al. 2020; Janda et al. 2021) we find that the removal of matrix-related signals improves the performance of untargeted MALDI-MSI efforts.

Finally, we provide a complete and validated database of DHB adducts with exogenous and endogenous compounds. This list can be used with rMSIcleanup to reliably and quickly annotate matrix-related signals in any dataset. It is worth noting that this new release leads to a considerable performance increase with respect to the initial release. The new release also includes an FDR estimation using a decoy library. Given a known list of adducts, rMSIcleanup can be used to confidently annotate matrix-related signals in any matrix.

This study opens a few different avenues for future work. Firstly, this methodology could be used to discover adducts in other commonly used matrices such as CHCA, DAN, or 9AA. In that regard, the main bottleneck is the limited availability of labeled MALDI matrix analog.

On the computational side, an interesting avenue to explore would be modeling the MALDI matrix adduct from a molecular structure point of view. In a first exploration we could build a probabilistic model of the prevalence of different matrix-related adducts under different matrices, tissue types, and mass analyzers. What sort of adducts does a specific matrix generate? In this regard, the more than 7000 openly-available datasets in METASPACE would be really valuable. In a second and deeper iteration, we could aim to link that information to specific aspects of the molecular structure of the matrix and the adducts. Why does a specific matrix promote a certain type of adduct?

4. Conclusions

Conclusion 1: rMSIfragment can annotate in-source fragments in lipids

In completion of Objective 1 of this thesis, we developed rMSIfragment, a computational tool to annotate lipid in-source fragments. We found that, if not properly annotated, in-source fragments of lipids can be overlapped with up to 50% of lipid parental annotations. rMSIfragment mitigates this issue by two mechanisms (1) the user is made aware of the overlap so they can be cautious with the interpretation of the results and (2) highly unlikely parental or fragment ions are deprioritized with a low-ranking score thus effectively reducing the number of overlaps.

Conclusion 2: rMSIcleanup can confidently annotate matrix-related signals

Objective 2 of this thesis was the development of an automated annotation tool for matrix-related signals. rMSIcleanup was developed to fulfil this objective. Apart from the main functionality of the package, rMSIcleanup is capable detect overlapped peaks. This is of special relevance when processing TOF data with lower mass resolving power. We initially demonstrated its use on Ag-LDI-MSI given the well characterized ionization and adduct formation of silver, as well as its distinct isotopic pattern. Given the successful results we moved on to the annotation of DHB, the most widely used MALDI matrix.

Conclusion 3: The combination of labeled and unlabelled MALDI matrix analogs in the same experimental setup allows for de-novo discovery of matrix-related signals.

The main limitation of rMSIcleanup is that it requires a known list of matrix-related adducts to search for. To further complete Objective 3 we proposed a novel experimental and computational pipeline to discover matrix-containing signals. By using a Stable Isotopically Labeled MALDI matrix we are able to shift all m/z related to the matrix. We compare technical replicates prepared with the labelled and unlabeled matrix to distinguish between (1) endogenous ions, (2) exogenous matrix-containing adducts, and (3) endogenous matrix containing adducts. This is the first time this methodology is introduced and it could be deployed to explore different matrices and experimental conditions.

Conclusion 4: The correct annotation of matrix-related signals and in-source fragments improves the performance of MSI untargeted metabolomics workflows

Finally, to complete Objective 3 we quantified the prevalence of in-source fragments and matrix-related signals, and we also studied their effect on MSI post processing. We found that in the study of lipids, more than 50% of the annotated lipids were isobaric to possible insource fragments. We also found that the prevalence of matrix-related exogenous ions is around 15% while the presence of matrix adducts formed with endogenous compounds reaches the 23%. In both cases, we demonstrate that the removal and proper handling of these overlooked signals improves both dimensionality reduction algorithms and metabolite annotation pipelines. Overall, we conclude that automatic annotation tools in MSI need to start considering and properly annotating in-source fragments and matrix-related adducts.

5. References

Baquer, Gerard, Lluc Sementé, María García-Altares, Young Jin Lee, Pierre Chaurand, Xavier Correig, and Pere Ràfols. 2020. “rMSIcleanup: An Open-Source Tool for Matrix-Related Peak Annotation in Mass Spectrometry Imaging and Its Application to Silver-Assisted Laser Desorption/ionization.” *Journal of Cheminformatics* 12 (1): 45.

Baquer, Gerard, Lluc Sementé, Toufik Mahamdi, Xavier Correig, Pere Ràfols, and María García-Altares. 2022. “What Are We Imaging? Software Tools and Experimental Strategies for Annotation and Identification of Small Molecules in Mass Spectrometry Imaging.” *Mass Spectrometry Reviews*, July, e21794.

Bond, Nicholas J., Albert Koulman, Julian L. Griffin, and Zoe Hall. 2017. “massPix: An R Package for Annotation and Interpretation of Mass Spectrometry Imaging Data for Lipidomics.” *Metabolomics: Official Journal of the Metabolomic Society* 13 (11): 128.

Coy, Shannon, Shu Wang, Sylwia A. Stopka, Jia-Ren Lin, Clarence Yapp, Cecily C. Ritch, Lisa Salhi, et al. 2022. “Single Cell Spatial Analysis Reveals the Topology of Immunomodulatory Purinergic Signaling in Glioblastoma.” *Nature Communications* 13 (1): 4814.

Dreisewerd, Klaus. 2003. “The Desorption Process in MALDI.” *Chemical Reviews*. https://doi.org/10.1021/cr010375i.

Dührkop, Kai, Markus Fleischauer, Marcus Ludwig, Alexander A. Aksenov, Alexey V. Melnik, Marvin Meusel, Pieter C. Dorrestein, Juho Rousu, and Sebastian Böcker. 2019. “SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information.” *Nature Methods* 16 (4): 299–302.

Garate, Jone, Sergio Lage, Lucía Martín-Saiz, Arantza Perez-Valle, Begoña Ochoa, M. Dolores Boyano, Roberto Fernández, and José A. Fernández. 2020. “Influence of Lipid Fragmentation in the Data Analysis of Imaging Mass Spectrometry Experiments.” *Journal of the American Society for Mass Spectrometry* 31 (3): 517–26.

Hillenkamp, Franz, Thorsten W. Jaskolla, and Michael Karas. 2014. “The MALDI Process and Method.” *MALDI MS. A Practical Guide to Instrumentation, Methods, and Applications, 2nd Ed. (Ed. : F. Hillenkamp, J. Peter-Katalinic), Wiley Blackwell, Weinheim, Germany*. https://onlinelibrary.wiley.com/doi/pdf/10.1002/9783527335961#page=16.

Hoffmann, Nils, Joel Rein, Timo Sachsenberg, Jürgen Hartler, Kenneth Haug, Gerhard Mayer, Oliver Alka, et al. 2019. “mzTab-M: A Data Standard for Sharing Quantitative Results in Mass Spectrometry Metabolomics.” *Analytical Chemistry* 91 (5): 3302–10.

Hu, Qizhi, Robert J. Noll, Hongyan Li, Alexander Makarov, Mark Hardman, and R. Graham Cooks. 2005. “The Orbitrap: A New Mass Spectrometer.” *Journal of Mass Spectrometry: JMS* 40 (4): 430–43.

Iakab, Stefania-Alexandra, Gerard Baquer, Marta Lafuente, Maria Pilar Pina, José Luis Ramírez, Pere Ràfols, Xavier Correig-Blanchar, and María García-Altares. 2022. “SALDI-MS and SERS Multimodal Imaging: One Nanostructured Substrate to Rule Them Both.” *Analytical Chemistry* 94 (6): 2785–93.

Jurinke, Christian, Paul Oeth, and Dirk van den Boom. 2004. “MALDI-TOF Mass

Spectrometry." *Molecular Biotechnology* 26 (2): 147–63.

Knochenmuss, R., and R. Zenobi. 2003. "MALDI Ionization: The Role of in-Plume Processes." *Chemical Reviews* 103 (2): 441–52.

Nikolaev, Eugene N., Yury I. Kostyukevich, and Gleb N. Vladimirov. 2016. "Fourier Transform Ion Cyclotron Resonance (FT ICR) Mass Spectrometry: Theory and Simulations." *Mass Spectrometry Reviews* 35 (2): 219–58.

Notarangelo, Giulia, Jessica B. Spinelli, Elizabeth M. Perez, Gregory J. Baker, Kiran Kurmi, Ilaria Elia, Sylwia A. Stopka, et al. 2022. "Oncometabolite D-2HG Alters T Cell Metabolism to Impair CD8+ T Cell Function." *Science* 377 (6614): 1519–29.

Ovchinnikova, Katja, Lachlan Stuart, Alexander Rakhlin, Sergey Nikolenko, and Theodore Alexandrov. 2020. "ColocML: Machine Learning Quantifies Co-Localization between Mass Spectrometry Images." *Bioinformatics* 36 (10): 3215–24.

Palmer, Andrew, Prasad Phapale, Ilya Chernyavsky, Regis Lavigne, Dominik Fay, Artem Tarasov, Vitaly Kovalev, et al. 2016. "FDR-Controlled Metabolite Annotation for High-Resolution Imaging Mass Spectrometry." *Nature Methods* 14 (1): 57–60.

Ruttkies, Christoph, Emma L. Schymanski, Sebastian Wolf, Juliane Hollender, and Steffen Neumann. 2016. "MetFrag Relaunched: Incorporating Strategies beyond in Silico Fragmentation." *Journal of Cheminformatics* 8: 3.

Schramm, Thorsten, Alfons Hester, Ivo Klinkert, Jean Pierre Both, Ron M. A. Heeren, Alain Brunelle, Olivier Laprévote, et al. 2012. "ImzML - A Common Data Format for the Flexible Exchange and Processing of Mass Spectrometry Imaging Data." *Journal of Proteomics* 75 (16): 5106–10.

Schymanski, Emma L., Junho Jeon, Rebekka Gulde, Kathrin Fenner, Matthias Ruff, Heinz P. Singer, and Juliane Hollender. 2014. "Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence." *Environmental Science & Technology* 48 (4): 2097–98.

Sementé, Lluc, Gerard Baquer, María García-Altares, Xavier Correig-Blanchar, and Pere Ràfols. 2021. "rMSIannotation: A Peak Annotation Tool for Mass Spectrometry Imaging Based on the Analysis of Isotopic Intensity Ratios." *Analytica Chimica Acta* 1171 (August): 338669.

Tortorella, Sara, Paolo Tiberi, Andrew P. Bowman, Britt S. R. Claes, Klára Ščupáková, Ron M. A. Heeren, Shane R. Ellis, and Gabriele Cruciani. 2020. "LipostarMSI: Comprehensive, Vendor-Neutral Software for Visualization, Data Analysis, and Automated Molecular Identification in Mass Spectrometry Imaging." *Journal of the American Society for Mass Spectrometry* 31 (1): 155–63.

Wang, Fei, Jaanus Liigand, Siyang Tian, David Arndt, Russell Greiner, and David S. Wishart. 2021. "CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification." *Analytical Chemistry* 93 (34): 11692–700.

Wang, Lin, Dan Wang, Olmo Sonzogni, Shizhong Ke, Qi Wang, Abhishek Thavamani, Felipe Batalini, et al. 2022. "PARP-Inhibition Reprograms Macrophages toward an Anti-Tumor Phenotype." *Cell Reports* 41 (2): 111462.

Wishart, David S. 2016. "Emerging Applications of Metabolomics in Drug Discovery and Precision Medicine." *Nature Reviews. Drug Discovery* 15 (7): 473–84.

UNIVERSITAT ROVIRA I VIRGILI
COMPUTATIONAL TOOLS FOR THE ANNOTATION OF IN-SOURCE FRAGMENTS AND MATRIX-RELATED SIGNALS IN MALDI MASS SPECTROMETRY IMAGING
Gerard Baquer Gómez

UNIVERSITAT
ROVIRA i VIRGILI