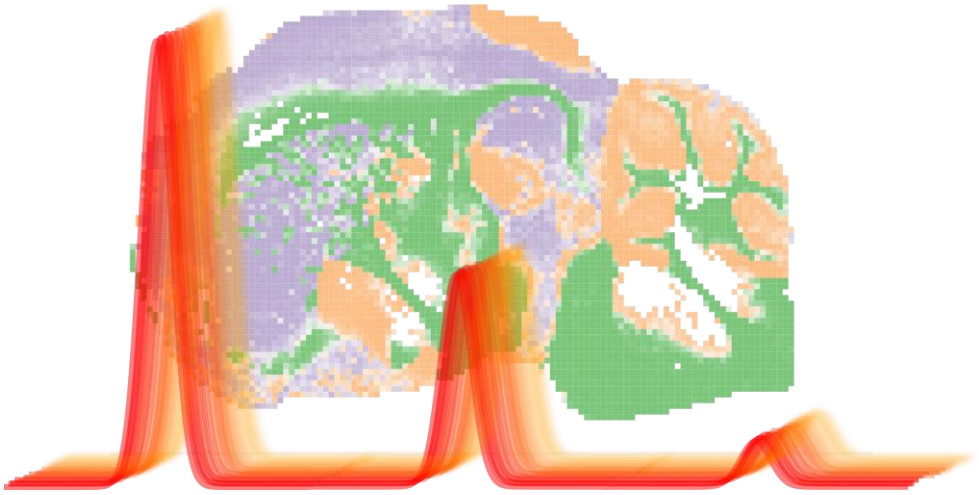
PEAK ANNOTATION AND DATA ANALYSIS SOFTWARE TOOLS FOR MASS SPECTROMETRY IMAGING

Lluc Sementé Fernández

UNIVERSITAT ROVIRA i VIRGILI

Peak annotation and data analysis software tools for mass spectrometry imaging

Lluc Sementé Fernández



DOCTORAL THESIS
2022

UNIVERSITAT ROVIRA I VIRGILI
PEAK ANNOTATION AND DATA ANALYSIS SOFTWARE TOOLS FOR MASS SPECTROMETRY IMAGING
Lluc Sementé Fernández

UNIVERSITAT ROVIRA I VIRGILI
PEAK ANNOTATION AND DATA ANALYSIS SOFTWARE TOOLS FOR MASS SPECTROMETRY IMAGING
Lluc Sementé Fernández

Lluc Sementé Fernández

PEAK ANNOTATION AND DATA ANALYSIS SOFTWARE TOOLS FOR 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
2022

UNIVERSITAT ROVIRA I VIRGILI
PEAK ANNOTATION AND DATA ANALYSIS SOFTWARE TOOLS FOR MASS SPECTROMETRY IMAGING
Lluc Sementé Fernández

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 "**Peak Annotation and Data Analysis Sofware Tools for Mass Spectrometry Imaging**", presented by **Lluc Sementé Fernández** 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, October 2022

Doctoral thesis supervisors:

Prof. Xavier Correig Blanchar

Dr. Pere Ràfols Soler

“La ciencia es el máximo control que podemos tener de la realidad.”

Gustavo Bueno Martínez

“Tiene la ciencia sus hipócritas, no menos que la virtud, y no menos es engañado el vulgo por aquéllos que por éstos. Son muchos los indoctos que pasan plaza de sabios.”

Benito Jerónimo Feijoo y Montenegro

“Cada planta, cada animal, incluso cada complejo minero, cada paisaje, tiene su razón de ser. No están a nuestro alcance por puro azar o capricho, sino que forma parte de nosotros mismos. El hombre no es un ovni venido de una lejana galaxia; el hombre es un poema tejido con la niebla del amanecer, con el color de las flores, con el canto de los pájaros, con el aullido del lobo o el rugido del león.”

Félix Rodríguez de la Fuente

Acknowledgments

L'elaboració d'aquesta tesis és, essencialment, un esforç col·lectiu, tal i com ho són valors com la convivència, l'harmonia social i l'empatia. Ha estat un exercici d'altruisme per part de moltíssimes persones cap a mi, que poca gent pot gaudir i el qual és impagable. Estic molt agraït pel meu camí i pels meus acompanyants.

A en Jesús, per obrim les portes de la casa. A en Xavier, per ser un pare en la ciència. A la María, per ser una mare en la ciència. A en Pere per ser un germà gran.

A l'Alex, la Carla, la Sonia, en Gerard i en Toufik, per ser companys de camí i batalles, exemples de treball i inspiració creativa.

A l'Esteban i en Nicolau per ser exemples a seguir i savis en el camí.

A tota la gent que forma o a format part de MiL@b, per la seva tasca crítica envers la meva feina, la seva generositat en l'ensenyança i l'amistat ràpida i afectuosa.

A la gent dels Països Baixos, per obrim la seva casa i ensenyem quan de diferent pot ser tot en un àmbit tant reduït com el nostre.

A la Júlia, el Dídac, el Salvador i la Gabriela, per ser els quatres pilars de la meva vida, sobre dels quals, en el seu suport, qualsevol cosa sembla minsa i passatgera.

A la família, tant natal com política, tant propera com llunyana, tant humana com animal, que amb el nostres noms i actes ens ubiquem sobre el món i ens reconeixen pel que som i actuem.

Als que no hi son, per ser-hi en el seu moment, el seu amor estarà sempre ben preservat en la nostra memòria.

A la meva terra, per ser el santuari de la calma que m'envaeix, la casa comú de la gent valuosa i paisatge dels meus somnis i fantasies.

Moltes gràcies a tots!

Lluc

Abstract

Spatial metabolomics is the discipline that studies the images of the distributions of low weight chemical compounds (metabolites) on the surface of biological tissues to unveil interactions between molecules. Mass spectrometry imaging (MSI) is currently the principal technique to get molecular imaging information for spatial metabolomics. MSI is a label-free molecular imaging technology that produces mass spectra preserving the spatial structures of tissue samples. This is achieved by ionizing small portions of a sample (a pixel) in a defined raster through all its surface, which results in a collection of ion distribution images (registered as mass-to-charge ratios (m/z)) over the sample. This thesis is aimed to develop computational tools for peak annotation in MSI and in the design of workflows for the statistical and multivariate analysis of MSI data, including spatial segmentation. The work carried out in this thesis can be clearly separated in two parts. Firstly, the development of an isotope and adduct peak annotation tool suited to facilitate the identification of the low mass range compounds. Secondly, the development of software tools for data analysis and spatial segmentation based on soft clustering for MSI data. All the developed algorithms have been implemented in software tools using the R platform, in continuation of the work of the group in the software packages rMSI and rMSIproc, since R is open and widely spread across biodata analysts. Nevertheless, we complement R code with C++ language to enable efficient memory control and faster execution of iterative algorithms. All the tools developed for this thesis are released under the general public license (GPL) to facilitate the exchange of ideas and collaboration between the MSI community.

The identification of the molecular formula and/or the chemical structure of the compounds in an MSI dataset is very challenging because usually there is only available the *m/z* exact mass. Therefore, new methods for the identifying and reporting molecular annotations in the low mass range for spatial metabolomics studies are required. To do so, we first aimed to localize which peaks were monoisotopic ions, as the searches on compounds libraries are based on the monoisotopic ion of the molecule. Additionally, we aimed to find groups of monoisotopic ions differing only with the adduct ion to determine neutral masses for the compounds, reducing the number of hits in compound libraries. However, we noticed that many compounds were missing in libraries like the Human Metabolome Database, and some of them are only included in minority libraries, like specific metabolites of algae. Therefore, developing a tool to match libraries into the data was not an option, and by that time the METASPACE tool, which successfully uses this strategy, already existed, so we did not want to reinvent the wheel. Instead, we opted for a completely new strategy which does not require searching on libraries at runtime. We developed a general rule of carbon-based isotropic patterns of the family of compounds under interest (metabolites and lipids) and compared it with spectral data. As a difference with LC-MS annotation methods, the higher number of observations (i.e., pixels) usually found in MSI gives statistical power to the results and allows to use the ratios between isotopic peaks as a key variable for monoisotopic peaks annotation. The result of this research was the development of the peak annotation tool rMSIannotation. rMSIannotation is useful for annotation of compounds and variable reduction strategies; and can be integrated in any low-weight compounds MSI data analysis workflows. The results show that rMSIannotation automatically extracts valuable information from both high (TOF) and ultra-high (FT-ICR) resolution spectrometers. The presented algorithm demonstrated a high performance and annotation confidence when compared to the established metabolomics MSI annotation platform: METASPACE and to the manual annotation approaches.

The huge number of ions in a MSI requires automatizing the report of ions of interest between different regions in spatial metabolomics studies. This is challenging as we had a big number of ions at each experiment and different spatial segmentation solutions, which resulted

in a big number of combinations of possible results. Additionally, as some ions have very low intensity values in some pixels of the clusters, classical parametric statistical tests failed. To overcome this we developed a workflow using nonparametric tests and the percentage of pixels in which a particular ion is not detected This work resulted in the publication of the R package rMSIKeyIon. The tool is very effective at discovering up or down-regulated ions between clusters using an unsupervised k-means procedure. The ions selected are the candidates that, subsequently, have to be identified. This package is a valuable tool for the untargeted analysis of MALDI images and is an important advance in this area because, at present, there are no tools available.

The state of the art of the segmentation techniques used in MSI considers that any pixel must be included only in one cluster (hard clustering). This is clearly a limitation, because in histology we find many transition regions between histological areas that are not captured by the clustering algorithms. Besides, it is known that it is very difficult to assess the performance of the hard clustering algorithms. These facts suggested the possibility of ranking the pixels in a cluster by similarity to the cluster prototype, with the objective of differentiating between pixels localized in homogenous regions and in transition regions from the point of view of histological areas. In this regard, we propose a soft/fuzzy clustering approach, a particular subset of clustering algorithms that could associate all clusters to a pixel in different degrees. We followed the trail of soft clustering in MSI, and we found that the Fuzzy c-means algorithm was not used in this context. Therefore, we researched the use of this soft clustering method as a possible way of ranking pixels for MSI data. From the study we conclude that fuzzy c-means brings additional information to MSI data analysis through the dimension of membership, which allows for new ways of interpreting the results compared with hard clustering results. In our case, the study of membership through the newly developed PFS (pixel fidelity score) , a score to compare membership distributions with different number of clusters, allowed easy selection of the pixels more related to a cluster, unveiled morphological regions more challenging to detect, and enhanced a tissue type classification workflow in multiple samples of a human head and neck cancer dataset.

In conclusion, the thesis covers the developments on three different directions of the data analysis and interpretation of MSI data: peak annotation, ion selection and soft clustering. We believe that the developed software tools together with the studies on soft clustering will have a positive impact on MSI for spatial metabolomics.

List of Publications

- del Castillo E, Sementé L, Torres S, Ràfols P, Ramírez N, Martins-Green M, Santafe M, Correig X. **rMSIKeyIon: An Ion Filtering R Package for Untargeted Analysis of Metabolomic LDI-MS Images.** Metabolites. 2019; 9(8):162. DOI: 10.3390/metabo9080162

- Iakab, S.A., Sementé, L., García-Altares, M. et al. **Raman2imzML converts Raman imaging data into the standard mass spectrometry imaging format.** BMC Bioinformatics 21, 448 (2020). DOI: 10.1186/s12859-020-03789-8

- Lluc Sementé, Gerard Baquer, María García-Altares, Xavier Correig-Blanchar, Pere Ràfols, **rMSIannotation: A peak annotation tool for mass spectrometry imaging based on the analysis of isotopic intensity ratios**, Analytica Chimica Acta, Volume 1171, 2021, 338669, ISSN 0003-2670, DOI: 10.1016/j.aca.2021.338669.

- Baquer, G, Sementé, L, Mahamdi, T, Correig, X, Ràfols, P, García-Altares, M. **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. DOI: 10.1002/mas.21794

- L. Sementé, G. Baquer, C. Bookmeyer, F. Avilés-Jurado, I. Vilaseca, P. Castillo, E. del Castillo, X. Correig, M. García-Altares, P. Ràfols **Fuzzy c-means improves the evaluation of segmentation processes in mass spectrometry imaging** (Submitted)

List of Congresses

- X. Correig, L. Sementé, G. Baquer, E. del Castillo, M. García-Altares, P. Ràfols **Library-free annotation of metabolites in Mass Spectrometry Imaging datasets using carbon-based isotopic patterns.** Congreso Anual de la Sociedad Española de Ingeniería Biomédica, CASEIB 22, (November 2022), Valladolid (Poster)

- X. Correig, L. Sementé, G. Baquer, E. del Castillo, M. García-Altares, P. Ràfols **Library-free annotation of metabolites in Mass Spectrometry Imaging datasets using carbon-based isotopic patterns.** 18th Annual Conference of the Metabolomics Society. METABOLOMICS 2022 (June 2022), Valencia, Spain. (Poster)

- L. Sementé, P. Ràfols, E. Del Castillo, J. Brezmes, O. Yanes, X. Correig. **New rMSIproc untargeted annotation engine highlights and ranks monoisotopic and adduct-related MS peaks within seconds.** Mass Spectrometry Imaging Society's seventh international conference. OurConVII. (October 2019) Saint Malo, France (Poster)

- L. Sementé, P. Ràfols, E. Del Castillo, J. Brezmes, O. Yanes, X. Correig. **A novel algorithm for low mass range automatic isotope & adduct annotation in MSI.** Mass Spectrometry Imaging Society's sixth international conference. OurConVI. (November 2018) Charleston, SC, United States of America. (Poster)

Table of Contents

CHAPTER 1

Introduction

1. Spatial Metabolomics

Spatial metabolomics is the discipline that studies the images of the distributions of low weight chemical compounds (metabolites) on the surface of biological tissues to unveil interactions between molecules.[1] The main objective of spatial metabolomics is to translate and expand the knowledge on metabolic pathways, disease dysregulations, microbiota interactions, and cell epigenetics over biological tissue surfaces preserving its morphology with the aim to isolate and understand the independent role of each morphological region over the global process.[2]

The precursors of spatial metabolomics are the classical microscopy technologies: histopathology[3] and immunohistochemistry.[4] These technologies are used to identify cell types and morphologies over tissue surfaces using different chemical labels to elaborate clinical diagnosis. Their main drawback is that they are only capable of representing the spatial distribution of a very limited number of compounds (not including metabolites) and they require the application of chemical labels.[5] To overcome these drawbacks, label-free molecular imaging methods are used.

Mass spectrometry imaging (MSI) is currently the principal technique to get molecular imaging information for spatial metabolomics.[2] MSI is a label-free molecular imaging technology that produces mass spectra preserving the spatial structures of tissue samples. This is achieved by ionizing small portions of a sample (a pixel) in a defined raster through all its surface, which results in a collection of ion distribution images (registered as mass-to-charge ratios (*m/z*)) over the sample. After the acquisition, the ions are annotated (putatively assigned to one molecule) and in some cases are identified using complementary techniques. Mass spectrometry imaging is the ground technique of this thesis, and therefore, in the next section there is a description of its process, summarizing all the steps and the elements of the whole workflow.

In recent years, other molecular imaging technologies are being used in combination with MSI to overcome some of their drawbacks and expand even more the field of spatial metabolomics. Vibrational spectroscopy imaging (VSI) technologies are probably the most notable case. VSI is based on the interaction between light beams of different wavelengths and a biological tissue. From this interaction, different vibration modes of the molecules are registered. The frequency of the vibrations registered at each sampling point (pixel) is transduced into molecular fingerprints, which can be used to study the abundance and structure of biomolecules like metabolites, lipids, proteins, and nucleic acids. The most commonly used VSI methods for spatial metabolomics are Fourier transform infrared spectroscopy (FTIR/IR), Raman spectroscopy, surface-enhanced Raman spectroscopy(SERS), and fluorescence spectroscopy.[6] The interest in combining both sources of spatial metabolomic information has found the appearance of a new line of research known as multimodal imaging, the aim of which is to directly fusion the data coming from MSI and VSI to compensate for the shortcomings of one with the capabilities of the other.[7] One of the most common approaches to multimodal imaging is scanning completely a tissue sample with an MSI technique and combining it with a “zoom in” scan of a morphological element of interest using a VSI technique, provided that these techniques allows much higher spatial resolution in the range of optical microscopy.

Apart from the technologies involved, spatial metabolomics experiments are usually classified into two groups according to the prior knowledge the scientist has of expected compounds in the sample[8]: (1) targeted experiments, where the compounds of interest are previously known and the goal is to study their spatial distribution, concentration, and interaction between them, like in pharmacology; and (2) untargeted experiments, where the goal of the study is to unveil the distribution of as many compounds as possible in a tissue including chemical unknowns, as a consequence of a biological experiment.

Finally, metabolomics is just one of the many pieces of the big puzzle that represents systems biology. Therefore, it is common for metabolomics to be combined with other omic sciences like proteomics, transcriptomics, and genomics in their classical and spatial approaches to answer a biological question.[9,10]

2. Mass spectrometry imaging

2.1. Technology introduction

MSI is an analytical technique capable of localizing mass spectra over sample surfaces[11]. MSI comprehends a wide range of technologies, each of them targets more efficiently specific molecular classes and resolves the sample in different spatial resolutions. MSI instruments consist mainly of two parts: the ion source and the mass analyzer. The ion source ablates small parts of the sample (pixels) and promotes the generation of ions. The mass analyzer measures the *m/z* of the ions and composes the spectrum at each sampling point by registering the intensity of the *m/z* measure, which can consist of directly counting ion impacts or measuring the magnitude of orbit frequencies. MSI technologies are usually classified according to the ion source mounted on the instrument, as ion sources are responsible for the generation of the molecular ions and its laser (or ion beam) and optical arrangements determines the spatial resolution of the images (the pixel size of the image), two of the most crucial elements that any scientist takes into consideration before planning an MSI experiment. The most mature MSI ionization methods are Desorption Electrospray Ionization (DESI), Laser Ablation Inductively Coupled Plasma (LA-ICP), Secondary Ion Mass Spectrometry (SIMS), and Matrix-Assisted Laser Desorption/Ionization (MALDI).[12] The work developed on this thesis was mainly achieved using MALDI-MSI data.

2.2. MALDI-MSI

Between them, MALDI-MSI is the most extended MSI technology due to the flexibility of their acquisition methods, which are capable of ionizing a wide range of compounds as metabolites, lipids, peptides and proteins, with spatial resolutions that can go down to 10 µm and in some specific cases even further.[13,14] MALDI-MSI consists of a MALDI ion source, which samples the surface of the sample with a pulsating laser (usually UV or IR) generally in a vacuum chamber. The ionization process of MALDI sources is considered to be soft, as most of the molecules remain intact after the LDI process. Still, in-source fragmentation occurs, producing peaks of the ion fragments which overlap with other low weight compound peaks. This is especially problematic in spatial metabolomics, as fragments tend to overlap in the mass range of the metabolites.[15,16] Apart from the instrument itself, MALDI sources require the choice and application of a matrix, usually small organic compounds, that assist in the desorption and ionization of the analytes from the tissue and are deposited over the tissue by spraying and more recently by sublimation[17]. Figure 1 shows a complete schema of the MALDI experiment.

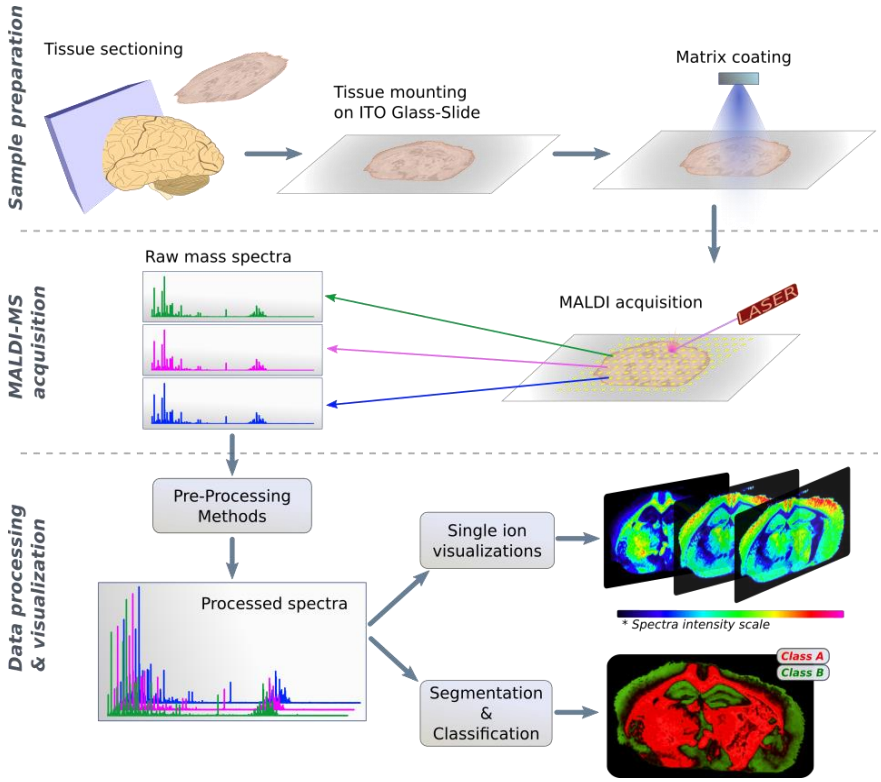


Figure 1. MALDI-MSI experiment overview. The process starts by sectioning a tissue, mounting it over an ITO glass-slide and coating it with a matrix. Later, the MALDI instruments samples multiple spots with a laser to generate the mass spectra. Finally, the spectra are processed, and the resulting data can be visualized and analyzed. Original image extracted from *Rafols. et al.*[18]

Depending on the kind of matrix, differences in the ionization process will occur including the promotion of some adduct elements, different ionization efficiency for different compound classes and apparition of clusters of matrix signals or matrix-related fragments in the spectra. These effects are determinant in the MSI experiment as they completely condition the identification of compounds, especially in the low mass range. To overcome these, in recent years inorganic matrices have been proposed as they produce a cleaner background spectrum.[18]

The most commonly used MALDI-MSI mass spectrometers are time-of-flight (TOF), a Fourier transform ion cyclotron resonance (FTICR), or an Orbitrap mass analyzer, which receive the ionic plume and compose the mass-to-charge (*m/z*) spectra at each sampling point. TOF mass analyzers are the most extended for MALDI-MSI and are used in all kinds of experiments as they can detect molecular ions in a very extended mass range (from low weight chemical compounds to proteins). Additionally, they have higher scan rates and are less expensive compared to Orbitrap and FTICR. Since the resolution of the TOF detectors increases with the mass, it is most used in peptidomics and proteomics studies. At low mass range, the poor resolution of the TOF spectrometers, together with the elevated number of ions from in-source fragmentation and the ions coming from the matrix, impedes the identification of compounds, since the peaks in the low mass range are highly overlapped. Orbitrap and FTICR mass analyzers produce spectra with ultra-high mass resolution, and they perform better in the low mass range as their mass resolution increases linearly as the sampled mass decreases, which allows them to distinguish better between low mass compounds like metabolites. For more details, a complete comparison between mass analyzers and ion sources can be found at Chapter 2.

2.3. Data processing

The steps in charge of transforming and organizing the raw data before the analysis and the identification of molecular species are known as data processing. The main objective behind data processing is to get an ion peak matrix, consisting of a matrix in which the observations (pixels) are organized in rows and the ions in columns. The matrix is calculated after several processing steps that minimize the undesirable effects that the instrument, the acquisition process, or other external factors may introduce to the data. The starting point of data processing is the acquisition of the raw data and particularly, the file format in which it is stored. Nowadays, the data produced by most of the instruments can be converted to the imzML standard[19,20], an open standard promoted by the MSI scientific community to facilitate the exchange and processing of mass spectrometry imaging data. This is done either by exporting the data directly from the instrument into the imzML format, or by using a data format converter. The imzML standard defines two structures for the spectral data: the continuous data format, in which all the raster positions (pixels of the image) share the same mass axis, and the processed data format, in which each pixel has a different mass axis, intended for storing already processed data.

Assuming that the data is stored without any kind of processing, the data processing workflow starts with the smoothing and the baseline reduction of all the spectra. Smoothing consists of removing high frequency noise from the spectra, while baseline reduction attempts to remove the very low frequencies. A common approach for smoothing is using the Savitzky-Golay filter[21], and for baseline reduction the TopHat algorithm.[22] Later on, all the mass spectra are aligned to compensate for possible peak mass variations during the acquisition between pixels. Alignment usually consists of shifting the spectra to minimize the mass error to a reference spectrum of calibrants. Additionally, it can be done without any calibrant spectrum by maximizing the spectral correlation between pixels or using more advanced methods based capable of shifting, contracting and expanding the spectrum.[23] Once all the spectra are well aligned, a general mass calibration procedure is applied using reference masses to increase the overall mass accuracy. Following this step, a peak detection algorithm is used to drastically reduce the spectral data to a peak list where only the mass peaks information is retained: *m/z* centroid position, peak intensity, peak shape integrated area and the calculated signal-to-noise ratio (SNR). Next, a common *m/z* axis can be obtained by binning together all the *m/z* centroids for the detected peaks in the spectra. Later, the peaks that were not detected in a pixel are usually filled with the integrated value of the processed spectrum in the surroundings of the *m/z* centroid of the peak. All these transformations shape the data in a matrix-like structure in which all pixels have an intensity value registered for all the *m/z* peaks, which allows the composition of *m/z* images using the coordinates of the pixels. Finally, the intensity of all the spectra is normalized to share a common intensity scale. This accounts for differences in ablation power between pixels during the acquisition and tissue inhomogeneities. The most used normalization methods are the Total Ion Count (TIC), the Root Mean Square (RMS) and using the logarithm of peak intensities.[24] It is important to say that the order or even the presence of each processing step described in this section is not mandatory. The workflow described in this section is the one that our team follows, which comprehends most of the universally followed steps.

2.4. Data analysis

Once the data has been processed starts the genuine spatial metabolomic data analysis of the tissue. The principal objective of the data analysis is to study the sample metabolism, unveiling regions of interest (ROI) over the sample surface and finding which are the key *m/z* ions involved in it. Various statistical tools and multivariate analysis techniques have been used to

confront this challenge,[25,26] but depending on the experimental approach, two analysis schemes are followed: targeted analysis and untargeted analysis.

If the MSI experiment is conceived as a targeted analysis, the first step usually consists of directly localizing the molecules of interest in the data that have been characterized and identified previously. This can be done manually manipulating the data searching for the specified *m/z* peaks or can be attempted with automatic peak annotation tools. If the MSI experiment is conceived as an untargeted analysis, the first step usually consists of applying a dimensionality reduction technique, as the number of *m/z* features under analysis tends to be too large. One way of approaching it is combining *m/z* features based on statistical criteria. This is achieved using machine learning algorithms like Principal Component Analysis (PCA)[27], Non-Negative Matrix Factorization[28], or t-distributed Stochastic Neighbor Embedding (t-SNE)[29], which transform the multidimensional *m/z* space into a smaller components number space in which the resulting components facilitate the identification of spatial morphologies over the sample. The main disadvantage of these methods is that once the new components are computed it can be difficult to identify the key *m/z* features, especially when using non-linear methods like the t-SNE.

After having a reduced number of *m/z* features under study, their spatial distributions are evaluated in both analysis schemes. Depending on the experimental design, the *m/z* features are compared in different regions of interest (ROI's) within the same tissue, whole tissue sections according to its experimental condition classification, or a combination of both approaches. The regions in which the intensity levels of the *m/z* features are compared, can be determined in a supervised manner, either directly over the MSI data or using an orthogonal imaging technique like histopathology[30]; or using unsupervised methods like spatial segmentation algorithms. The diverse clustering algorithms used for image segmentation are of utmost importance for this task. Clustering algorithms form data groups according to similarity scores to reveal spatial structures over the tissue. Most clustering workflows in MSI include variations of the k-means algorithm and hierarchical clustering.[26] These algorithms optimize objective functions using distance metrics like the euclidean or the cosine, and use a predefined number of clusters (k-means) or develop a complete dendrogram with different numbers of clusters (hierarchical clustering).

Finally, once specified the ROIs in the form of pixels with different labels, the following step is comparing the intensity distribution of all the *m/z* features of interest. This can be done visually, by exploring the distribution of the *m/z* features over the ROIs; using discriminant analysis like univariate tests (t-test, Wilcoxon rank sum test) and Fold change; or multivariate methods like partial least squares-discriminant analysis (PLS-DA).[31] Once discovered the differences between *m/z* peaks over the ROIs only lacks, for untargeted analysis, the identification of the *m/z* peaks.

2.5. Peak annotation and identification

The identification process consists of confidently assign a molecular formula and/or the chemical structure to a group of mass peaks. In MSI this is a very challenging process because the data only consist of the mass of the peaks. Furthermore, due to the lack of a chromatographic step, all compounds ionizes and travel to the mass analyzer simultaneously at each sampling point, producing overlapped ion peaks in the spectra. For instance, in liquid chromatography-mass spectrometry (LC-MS), the chromatography establishes an order in the ionization of compounds by affinity to the chromatographic column, developing an orthogonal dimension usually referred to as retention time (RT). Using the RT helps in distinguishing between compounds with a similar molecular weight, however ion fragmentation is required for absolute confidence.[32,33]

Another limitation is the lack of a controlled and efficient ion fragmentation process in MSI instruments. Ion fragmentation is used to break down the molecule in small parts that can be associated with common molecular building blocks, like functional groups, due to the energy required to break their bounds. There are very few instruments and experimental setups that can combine MSI and tandem mass spectrometry (MS/MS) in the same sample. Moreover, as all the compounds are ionized at the same time, the MS/MS spectra can get crowded with peaks even filtering the ions with a mass window of 1 Da. Additionally, the concentration of the precursor is very low compared to most MS/MS methods, as it is limited to the precursors localized on the scan area (pixel) and not on a complete sample like LC-MS. Therefore, the compound identification using MS/MS is usually done with other samples to later be searched in the MSI dataset.

Some software tools have been developed to annotate peaks, which is a previous step of molecular identification. In this regard, it is important to highlight the METASPACE[34] annotation platform, an online MSI annotation resource that produces molecular annotations and contains the biggest repository of MSI datasets with annotations that most of them are downloadable. Chapter 2 consists of a complete state of the art of the identification and annotation of ion peaks, a revision of the software tools and experimental methods.

3. Thesis motivation and objectives

The work presented in this thesis is the result of the research carried out in the Metabolomics Interdisciplinary Laboratory (MiL@b) group. MiL@b is a research group from the Department of Electronic, Electrical, and Automation Engineering (DEEEA) at the Universitat Rovira i Virgili (URV), and the Metabolomics Platform. The Metabolomics Platform is part of the Pere Virgili Health Research Institute (IISPV) and CIBER of Diabetes and Metabolic Diseases (CIBERDEM). The principal lines of research of MiL@b are a) signal processing for Nuclear Magnetic Resonance (NMR) metabolomics, b) LC-MS and GC-MS metabolomics data processing and metabolite identification, c) Toxicology & environmental metabolomics and d) Label-free spectrometry imaging for biological applications. This thesis has been developed in the context of this latter research area.

This group delves into all steps of the MSI workflow: sample preparation and treatment, spectra processing, data analysis and biomedical applications. We specialize in matrix dry-deposition techniques such as sputtering and thin layer thermal evaporation to apply ionization promoters on biological samples. We also design nanostructured surfaces for MSI that are compatible with Raman Imaging. In terms of instrumentation, the laboratory is equipped with a Spectroglyph MALDI source coupled with an Orbitrap Exploris 120, and we have access to a Bruker ultrafleXtreme MALDI-TOF. Additionally, we have a cryostat to cut tissue slides and a thermal evaporation system able to deposit matrices to tissues by sublimation.

This thesis started as a continuation of the work in MSI data processing and analysis. This work is based on the previous R packages released by the group: rMSI[35] and rMSIproc[36]. These packages allow the exploratory analysis and visualization of MSI data and process raw MSI spectra, respectively. rMSIproc data processing workflow comprehends all the steps described in Section 1.2.3. rMSIproc is compatible with the imzML data format and includes all the steps in the spectra processing workflow (noise reduction, label-free alignment, mass calibration, peak detection, peak binning, and normalizing). The output of the rMSproc package is a data matrix in which the columns represent the *m/z* peaks and the rows the pixels of an image, which we refer as peak matrix. Figure 2 illustrates the different steps of the general MSI pipeline followed by rMSI and rMSIproc. rMSI and rMSIProc packages act as a starting point for this thesis, with the aim to increase the functionality of them with the packages developed in this thesis.

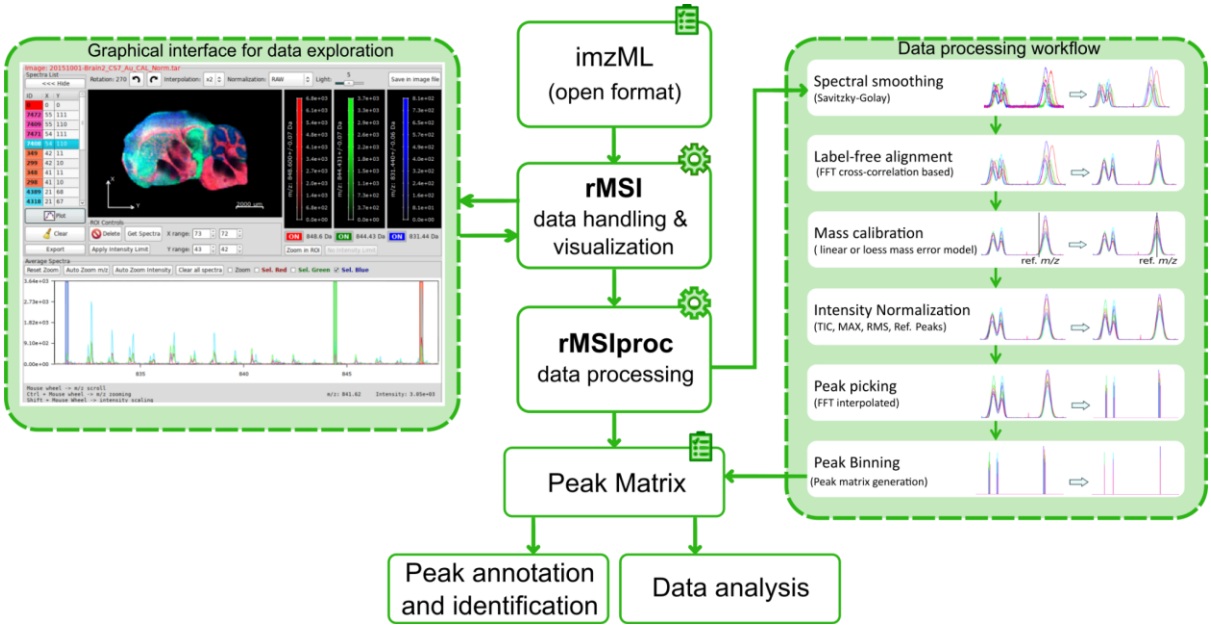


Figure 2. MSI data flow diagram covered by rMSI and rMSIproc environment. Peak annotation and identification, and data analysis strategies are the main objective of the thesis.

This thesis is aimed to develop computational tools for peak annotation in MSI and in the design of workflows for the statistical and multivariate analysis of MSI data, including spatial segmentation. The main objectives of this thesis are:

1. To develop and implement an automatic workflow for the statistical analysis of ion abundances distribution in MSI datasets.

This first objective originated from the necessity of automatizing the report of ions of interest between different regions within and between images in spatial metabolomics studies. This was challenging as we had a big number of ions at each experiment and different spatial segmentation solutions, which resulted in a big number of combinations of possible results. Additionally, as some ions had very low intensity values in some pixels of the clusters, classical parametric statistical tests failed. To overcome this, we developed a workflow using nonparametric tests and the percentage of pixels in which a particular ion is not detected This work is presented in the third chapter of this thesis and resulted in the publication of the R package rMSIKeyIon[37].

2. To develop an algorithm for isotopic and adduct ion annotation for MSI datasets in the low mass range without using libraries and implement the algorithm in a software tool compatible with rMSIproc.

The second objective is aimed at identifying and reporting molecular annotations in the low mass range for spatial metabolomics studies. To do so, we first aimed to find the isotopic patterns of a molecule to localize which peaks were monoisotopic ions, as the searches on compounds libraries are based on the monoisotopic ion of the molecule. Additionally, we aimed to find groups of monoisotopic ions differing only with the adduct ion to determine neutral masses for the compounds, reducing the number of hits in compound libraries. However, we noticed that many compounds were hard to find in libraries like the Human Metabolome Database as they are based on LC-MS data, and some compounds depending on

the study might not be included in some libraries, like specific metabolites of algae. Therefore, developing a tool to match libraries into the data was not an option, and by that time the METASPACE tool, which successfully uses this strategy, already existed, so we did not want to reinvent the wheel. Instead, we opted for a completely new strategy which does not require searching on libraries at runtime. We developed a general rule of carbon-based isotropic patterns of the family of compounds under interest (metabolites and lipids) and compared it with spectral data. As a difference with LC-MS annotation methods, the higher number of observations (i.e., pixels) usually found in MSI gives statistical power to the results and allows to use the ratios between isotopic peaks as a key variable for monoisotopic peaks annotation. The result of this research was the development of the peak annotation tool rMSIannotation[38], presented in chapter 4. Moreover, we reasoned that having redundant peaks (isotopic peaks) in our peak matrices was detrimental to upcoming statistical analysis, as including many strongly correlated peaks could bias clustering procedures and dimensionality reduction techniques by overrepresenting their own morphological features.

3. To implement and evaluate the performance of the fuzzy c-means algorithm for the spatial segmentation of MSI datasets.

The third objective originated from the state of the art of the segmentation techniques used in MSI, which considers that any pixel must be included only in one cluster (hard clustering). This is clearly a limitation, because in histology we find many transition regions between histological areas that are not captured by the clustering algorithms. Besides, it is known that it is very difficult to assess the performance of the hard clustering algorithms. These facts suggested the possibility of ranking the pixels in a cluster by similarity to the cluster prototype, with the objective of differentiating between pixels localized in homogenous regions and in transition regions from the point of view of histological areas. In this regard, we propose a soft/fuzzy clustering approach, a particular subset of clustering algorithms that could associate all clusters to a pixel in different degrees. We followed the trail of soft clustering in MSI, and we found that the Fuzzy c-means algorithm was not used in this context. Therefore, we researched the use of this soft clustering method as a possible way of ranking pixels for MSI data. The results of this research are presented in chapter five.

Finally, all the developed algorithms have been implemented in software tools using the R platform, in continuations of rMSI and rMSIproc, since R is open and widely spread across biodata analysts. Nevertheless, we complement R code with C++ language to enable efficient memory control and faster execution of the iterative algorithm. All the tools developed for this thesis are released under the general public license (GPL) to facilitate the exchange of ideas and collaboration between the MSI community.

4. Organization of the document

The thesis is divided into six chapters, which comprehend this introduction, the compendium of articles that cover the goals of this thesis, and a final discussion with conclusions. Chapter 1 contains a general introduction to spatial metabolomics, a more in-depth introduction to MSI, and the motivations, the context, and the objectives of the thesis. Chapter 2 contains a review article with the state of the art of the identification and annotation of compounds in MSI published in Mass Spectrometry Reviews[12]. This chapter covers the whole process of identification using MSI data with special emphasis on the available software and the strategies they follow to achieve the identification. Chapter 3 contains the results of the first objective. In it the tool rMSIKeyIon, an R package for the automatic filtering of ions based on the intensity distribution differences over clusters is presented. Chapter 4 covers the second

objective of the thesis and describes the algorithms used for isotope and adduct annotation and presents the rMSIannotation software that implements it. Chapter 5 covers the third objective exploring the use of fuzzy c-means for MSI data as a criterion for the evaluation of clustering results using the information of the pixel membership to the clusters. Finally, Chapter 6 contains a discussion and the final conclusions on the milestones achieved during this thesis over the three main objectives and some perspectives on future work based on them.

5. References

1. Petras, D., Jarmusch, A. K. & Dorrestein, P. C. From single cells to our planet—recent advances in using mass spectrometry for spatially resolved metabolomics. *Current Opinion in Chemical Biology* vol. 36 24–31 (2017).
2. Alexandrov, T. Spatial Metabolomics and Imaging Mass Spectrometry in the Age of Artificial Intelligence. *Annu Rev Biomed Data Sci* **3**, 61–87 (2020).
3. Lavis, L. D. Histochemistry: live and in color. *J. Histochem. Cytochem.* **59**, 139–145 (2011).
4. Ramos-Vara, J. A. & Miller, M. A. When tissue antigens and antibodies get along: revisiting the technical aspects of immunohistochemistry--the red, brown, and blue technique. *Vet. Pathol.* **51**, 42–87 (2014).
5. Alturkistani, H. A., Tashkandi, F. M. & Mohammedsaleh, Z. M. Histological Stains: A Literature Review and Case Study. *Global Journal of Health Science* vol. 8 72 (2015).
6. Neumann, E. K., Djambazova, K. V., Caprioli, R. M. & Spraggins, J. M. Multimodal Imaging Mass Spectrometry: Next Generation Molecular Mapping in Biology and Medicine. *J. Am. Soc. Mass Spectrom.* **31**, 2401–2415 (2020).
7. Iakab, S. A., Ràfols, P., Correig-Blanchar, X. & García-Altares, M. Perspective on multimodal imaging techniques coupling mass spectrometry and vibrational spectroscopy: Picturing the best of both worlds. *Anal. Chem.* **93**, 6301–6310 (2021).
8. Roberts, L. D., Souza, A. L., Gerszten, R. E. & Clish, C. B. Targeted metabolomics. *Curr. Protoc. Mol. Biol.* **Chapter 30**, Unit 30.2.1–24 (2012).
9. Vailati-Riboni, M., Palombo, V. & Loor, J. J. What Are Omics Sciences? in *Periparturient Diseases of Dairy Cows* 1–7 (Springer International Publishing, 2017).
10. Boja, E. S., Kinsinger, C. R., Rodriguez, H. & Srinivas, P. Integration of omics sciences to advance biology and medicine. *Clin. Proteomics* **11**, 45 (2014).
11. McDonnell, L. A. & Heeren, R. M. A. Imaging mass spectrometry. *Mass Spectrom. Rev.* **26**, 606–643 (2007).
12. Baquer, G. *et al.* What are we imaging? Software tools and experimental strategies for annotation and identification of small molecules in mass spectrometry imaging. *Mass Spectrom. Rev.* e21794 (2022).
13. Hansen, R. L. & Lee, Y. J. Overlapping MALDI-Mass Spectrometry Imaging for In-Parallel MS and MS/MS Data Acquisition without Sacrificing Spatial Resolution. *J. Am. Soc. Mass Spectrom.* **28**, 1910–1918 (2017).
14. Wäldchen, F., Mohr, F., Wagner, A. H. & Heiles, S. Multifunctional Reactive MALDI Matrix Enabling High-Lateral Resolution Dual Polarity MS Imaging and Lipid C═C Position-Resolved MS Imaging. *Anal. Chem.* **92**, 14130–14138 (2020).
15. Baquer, G. *et al.* rMSIcleanup: an open-source tool for matrix-related peak annotation in mass spectrometry imaging and its application to silver-assisted laser desorption/ionization. *J. Cheminform.* **12**, 45 (2020).
16. Garate, J. *et al.* Influence of Lipid Fragmentation in the Data Analysis of Imaging Mass Spectrometry Experiments. *J. Am. Soc. Mass Spectrom.* **31**, 517–526 (2020).

17. Gemperline, E., Rawson, S. & Li, L. Optimization and comparison of multiple MALDI matrix application methods for small molecule mass spectrometric imaging. *Anal. Chem.* **86**, 10030–10035 (2014).

18. Ràfols, P. *et al.* Assessing the potential of sputtered gold nanolayers in mass spectrometry imaging for metabolomics applications. *PLoS One* **13**, e0208908 (2018).

19. Römpp, A. *et al.* imzML: Imaging Mass Spectrometry Markup Language: A common data format for mass spectrometry imaging. *Methods Mol. Biol.* **696**, 205–224 (2011).

20. Schramm, T. *et al.* imzML — A common data format for the flexible exchange and processing of mass spectrometry imaging data. *Journal of Proteomics* vol. 75 5106–5110 (2012).

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

22. van Herk, M. A fast algorithm for local minimum and maximum filters on rectangular and octagonal kernels. *Pattern Recognit. Lett.* **13**, 517–521 (1992).

23. Ràfols, P., Castillo, E. D., Yanes, O., Brezmes, J. & Correig, X. Novel automated workflow for spectral alignment and mass calibration in MS imaging using a sputtered Ag nanolayer. *Anal. Chim. Acta* **1022**, 61–69 (2018).

24. Deininger, S.-O. *et al.* Normalization in MALDI-TOF imaging datasets of proteins: practical considerations. *Anal. Bioanal. Chem.* **401**, 167–181 (2011).

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

26. Alexandrov, T. MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC Bioinformatics* **13 Suppl 16**, S11 (2012).

27. Klerk, L. A., Broersen, A., Fletcher, I. W., van Liere, R. & Heeren, R. M. A. Extended data analysis strategies for high resolution imaging MS: New methods to deal with extremely large image hyperspectral datasets. *Int. J. Mass Spectrom.* **260**, 222–236 (2007).

28. Leuschner, J. *et al.* Supervised non-negative matrix factorization methods for MALDI imaging applications. *Bioinformatics* **35**, 1940–1947 (2019).

29. Abdelmoula, W. M. *et al.* Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 12244–12249 (2016).

30. Schwamborn, K. The Importance of Histology and Pathology in Mass Spectrometry Imaging. *Adv. Cancer Res.* **134**, 1–26 (2017).

31. Pérez-Guaita, D., Quintás, G. & Kuligowski, J. Discriminant analysis and feature selection in mass spectrometry imaging using constrained repeated random sampling - Cross validation (CORRS-CV). *Anal. Chim. Acta* **1097**, 30–36 (2020).

32. Tada, I. *et al.* Creating a Reliable Mass Spectral-Retention Time Library for All Ion Fragmentation-Based Metabolomics. *Metabolites* **9**, (2019).

33. Schymanski, E. L. *et al.* Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ. Sci. Technol.* **48**, 2097–2098 (2014).

34. Alexandrov, T. *et al.* METASPACE: A community-populated knowledge base of spatial metabolomes in health and disease. *bioRxiv* (2019) doi:10.1101/539478.

35. Ràfols, P. *et al.* rMSI: an R package for MS imaging data handling and visualization. *Bioinformatics* **33**, 2427–2428 (2017).

36. Ràfols, P. *et al.* rMSIproc: an R package for mass spectrometry imaging data processing. *Bioinformatics* **36**, 3618–3619 (2020).

37. Del Castillo, E. *et al.* rMSIKeyIon: An Ion Filtering R Package for Untargeted Analysis of Metabolomic LDI-MS Images. *Metabolites* **9**, (2019).

38. Sementé, L., Baquer, G., García-Altares, M., Correig-Blanchar, X. & Ràfols, P. rMSIannotation: A peak annotation tool for mass spectrometry imaging based on the analysis of isotopic intensity ratios. *Anal. Chim. Acta* **1171**, 338669 (2021).

CHAPTER 2

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

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.

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/), at the moment 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 Supplementary 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 adaptation of the 5 level system to report identification confidence in MSI experiments as described below. The strategies to achieve the different confidence levels mentioned in this section are described in detail in sections III and IV.

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). The discrimination of isobaric and isomeric species is one of the major challenges of MSI, since it cannot rely of the chromatographic separation of these species. As described in a recent review (Bednařík *et al.*, 2022), there are several strategies to discriminate isomers in MSI, including ion activation and chemical derivatization, ion mobility spectroscopy, and tandem MS analysis directly on tissue, among others.

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 -80°C 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.

Since the reactions for protein digestion are performed in solution, the tissues need to be covered by solvents containing the digestive enzymes, which can lead to the delocalization of the peptide products. Solvent-free solutions can avoid peptide delocalization, for instance by the use of plasmonic thermal decomposition/digestion (Zhou and Basile, 2017). This process uses continuos wave laser excitation and gold nanoparticles to decompose proteins at known locations (C-terminus of aspartic acid and at the N-terminus of cysteine) and since this is a dry technique, product peptides retain their original location on the tissue.

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). Additionally, the induced epoxidation of peracetic acid has been used to localize the C=C bonds in unsaturated fatty acids, allowing the discrimination of isomeric fatty acid (H. Zhang *et al.*, 2021); and Giard's reagent P has been used to label N-glycans in FFPE tissue samples, increasing the sensitivity of the tissue samples characterization (H. Zhang *et al.*, 2020).

Chemical reagents can also be used to promote a specific adduct of relevant biological molecules that are present in low concentration in tissues. For instance, *Duncan et al.* added silver ions to the solvent for nanospray desorption electrospray ionization MSI to enhance the ionization of prostaglandins as silver adducts, which allowed their monitoring directly on mice tissues (Duncan *et al.*, 2018).

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 glycerophosphoinositol 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. Reactive matrices can also aid discriminating between isomeric compounds, such as the reactive matrix benzophenone, that serves both as ionization promoter and as derivatization reagent to selectively functionalize unsaturated phospholipids (Wäldchen, Spengler and Heiles, 2019). 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 1 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 amount 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; G. 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.

All mass analyzers are often calibrated before acquisition to obtain accurate *m/z* measurements. The calibration consists of tuning the electronic parameters of the instrument to modify the *m/z* axis according to different calibration curves build upon measured calibration standards. The calibration standards are liquid mixtures of highly purified molecules designed for positive and/or negative mode in a particular mass range. Different strategies to obtain calibration curves are used depending on the time of the standard application and the mass analyzers (Smith *et al.*, 2012). In terms of standard application, internal calibration consists of mixing calibration standards directly with the sample of interest, while external calibration consists of measuring the standards alone (Muddiman and Oberg, 2005). When FT-ICR or

Orbitrap instruments are used, it is convenient to use an abundance dependent calibration curve, as different amount of ions inside the ICR cell produce varying frequency shifts, resulting in different mass errors at each sampling point (Easterling, Mize and Amster, 1999; Zhang *et al.*, 2005; Gorshkov *et al.*, 2010).

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). Tandem MS on tissue can help discriminating isomers of lipids due to their differential fragmentation. For instance,

(Takeo *et al.*, 2019) used tandem MS (MS^3) to discriminate between structural isomers of some steroids, after applying on-tissue chemical derivatization techniques to enhanced their ionization efficiency.

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 (unless other strategies to promote their ionization are considered).

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 amount 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) has to 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.). An alternative technique to LCM is liquid extraction surface analysis mass spectrometry (LESA-MS), which combines micro-extraction in the liquid phase directly from the tissue with nano-electrospray MS. While it is considered a low spatial resolution technique (generally aprox. 1mm), it is a valuable complement to high spatial resolution techniques (like MALDI-MSI), since it provides additional information for identification of compounds *in-situ* without the need to homogenate the sample. This technique has been successfully applied for instance to monitor drugs and drug metabolism on mice organs (Eikel et al. 2011; Swales et al. 2015).

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 MSIReader (Robichaud *et al.*, 2013; Bokhart *et al.*, 2018), 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 2 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, taking into account 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 proposed and compared with cosine similarity, obtaining better results in library matching fragmented molecules.

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 2210 ion images ranked by similarity by 42 MSI 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) or spatial chaos (Palmer *et al.*, 2016; Tortorella *et al.*, 2020). Additionally, the notion of False Discovery Rate (FDR) has been used to estimate the confidence of annotations using a target-decoy approach in which the resulting molecular formulas are compared with impossible adduct formations (Palmer et al. 2016; Guo et al. 2021).

There is no consensus on the relationship between different scores or how to unify them. Typically, when multiple scores are available, each score is scaled to fit a range of 0 to 1 and the product of all scores is taken as a single metric (Baquer et al. 2020; Sementé et al. 2021; Tortorella et al. 2020; 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.

4.2.6 In silico libraries

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 LipidBlast (Kind *et al.*, 2013), 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.

One of the main limitations of these libraries for MSI is that they tend to be trained with experimental data obtained by LC-MS/MS experiments, where most parental ions are fragmented as protonated adducts. Therefore, most common *in silico* libraries only include fragmentation predictions of protonated molecules, which do not represent how other adducts fragment (Al-Saad et al. 2003). In MSI, it is common to obtain different adducts such as sodium and potassium, depending on the sample preparation and acquisition, and for some species the protonated adduct may not even be detected (Garate et al. 2020).

4.2.7 Peak Filtering

Peak annotation results can be used to filter out redundant or not 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).

4.2.8. 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 are able to 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. 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 taken into account 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). With this transformation, the mass of the reference molecule -which is usually a repeating block in a bigger structure like CH_2 in lipids- does not contribute to the decimal part of the KM, which contains only information of the other elements of the molecular structure. The KMD is later obtained by subtracting the rounded KM from the KM. (Kune *et al.*, 2019) 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 allows the prediction of 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 the clustering algorithm bisecting k-means and based on the correlation of their spatial distribution. The output of rMSIcleanup is an R data frame that can be exported in Rdata or CSV formats. Additionally, the package produces an informative visual report in PDF with all the patterns detected, ion images, and matrix-related annotations.

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 the 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 at the moment, 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. Polypeptides 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 Supplementary 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.

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 terms 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 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 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 2 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. In particular, 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 3 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).

mzTab-M is in plain text, making it visually easy to read and understand. Additionally, its tab-separated format, similar to the CSV format, is natively supported in Excel and other spreadsheet software. It is therefore a viable alternative to excel and CSV files used in publications and statistical programming languages like R.

The main drawback of mzTab-M is that it relies on a custom structure defined by its own specification. We consider that using Extensible Markup Language (XML), a ubiquitous file format in all fields of computer science, would offer several advantages. All major programming languages and platforms have plenty of reliable tools to read, write and validate XML. And its well-defined structure makes it extensible. In the long run, adapting mzTab-M to 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. To ensure ease of access to the annotations and identifications by researchers with a lack of coding background the community should develop a converter to the original mzTab-M tab-separated 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 take into account 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 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 by addressing the hot topic on every researcher's lips: 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 (W. 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?".

7. Figures and tables

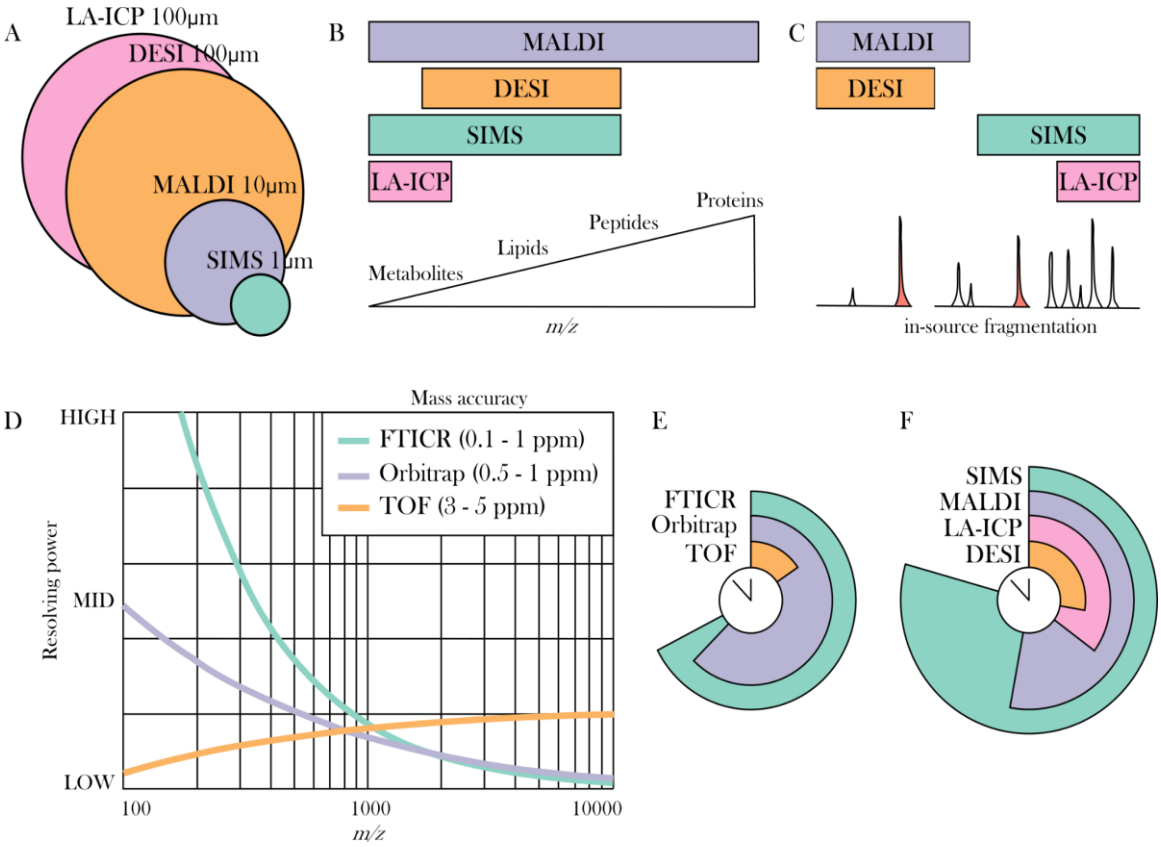


Figure 1. General comparison of the most widely used ion sources and mass analyzers for MSI. A: Spatial resolution, B: Mass range, and C: In-source fragmentation of the four most common ion sources. D: Resolving power and mass accuracy and E: Acquisition time of the three most common mass analyzers. 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/.

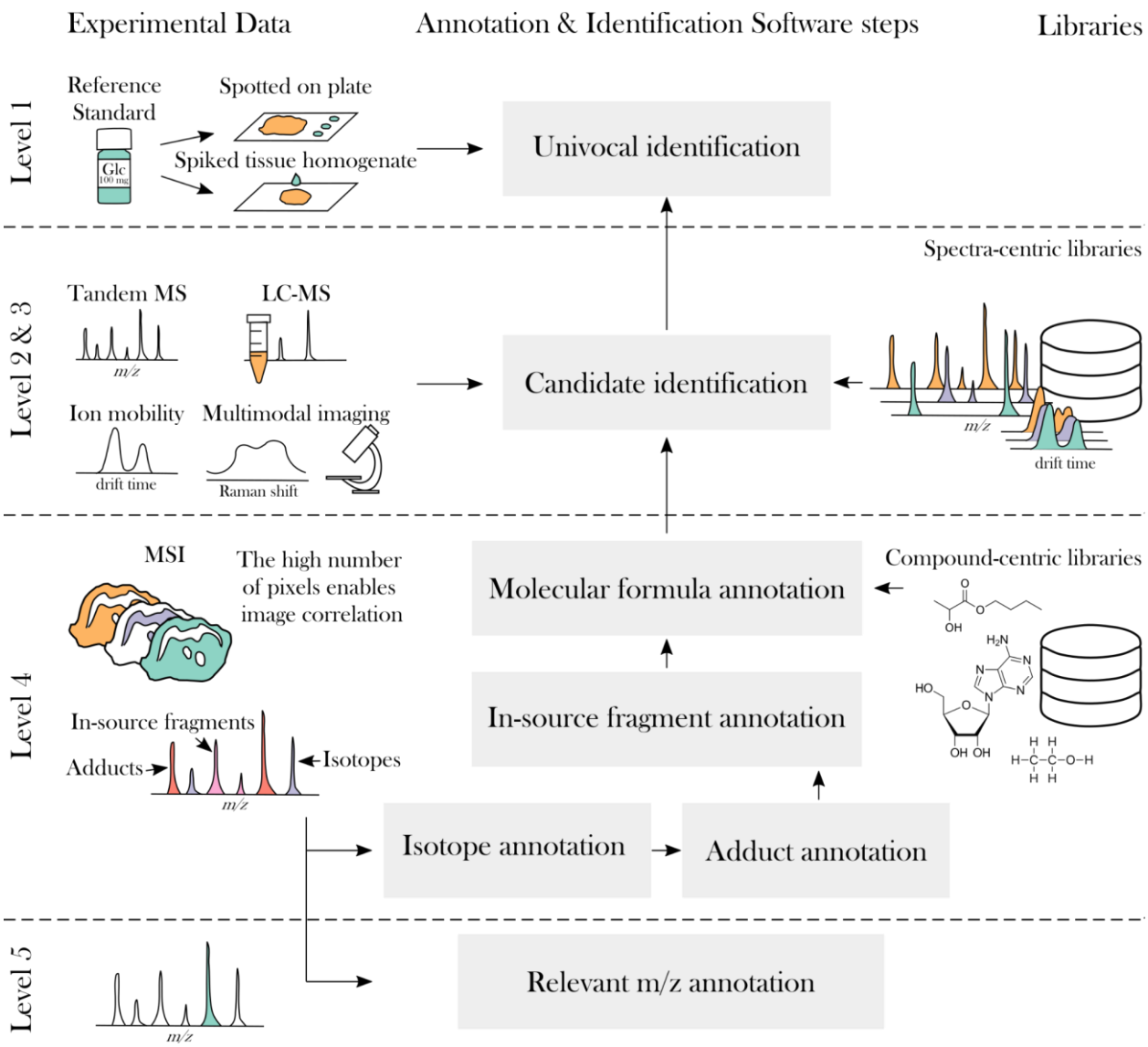


Figure 2. General steps in software annotation & identification in MSI experiments

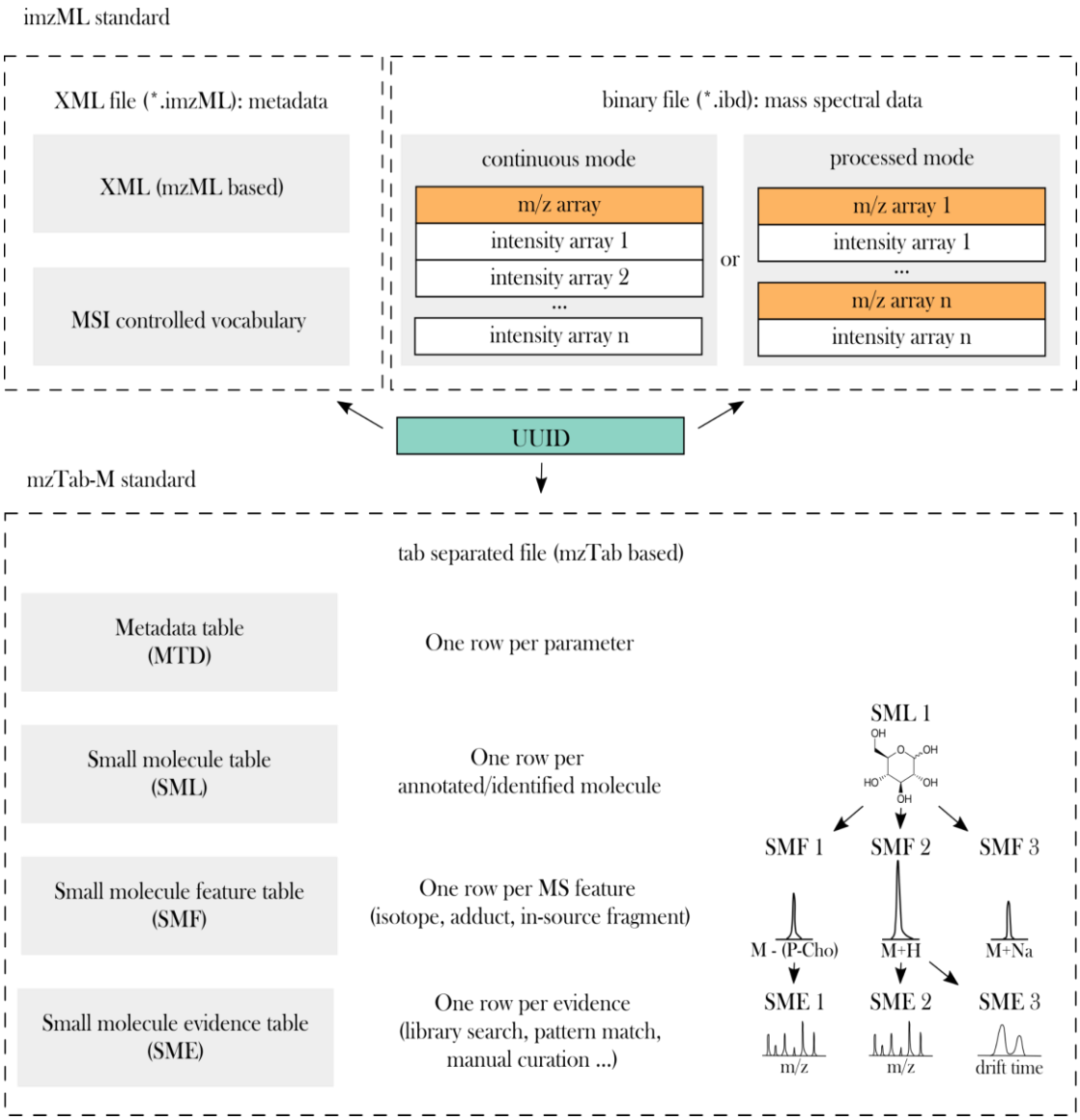


Figure 3. Adaptation of mzTab-M format to be compatible with imzML. A list of Unique Universally Identifiers (UUIDs) would link multiple imzML files from the same study to a single mzTab-M file containing annotations and identifications. Adapted with permission from (Schramm *et al.*, 2012) and (Hoffmann *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 ranges 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 ranges 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 ranges 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** **- Ranking scores**	**- 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) - (1) Matched fragments percentage	- 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 windows with tables and spectra. - (1) Matched fragment count (2) Sum of fragment intensities	- 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 - (1) Matched peaks percentage (2) RMS spectral error (3) Mass error (4) FDR (Target-Decoy)	- 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 windows with tables and spectra - (1) Matched fragment percentage (2) TIC ratio (3) Spatial chaos (4) Mass error	- Not specified - Download and install a standalone package. - Private Software	(Tortorella et al., 2020)

Name	**Confidence level**	**- Annotation type** **- Target features**	**-Approach** **- Library**	**-Input data format** **-Output** **- Ranking scores**	**- Programming language** **- Installation** **- License**	**Ref.**
mass2adduct	4	- Specific to metabolite-matrix adducts. - Isotopes, adducts, and matrix adducts	- Feature-centric	- imzML - List of adduct masses in the R environment - (1) Pearson correlation	- 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 - Tables in CSV format - None	- 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. - (1) KMD	- 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 - None	- 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 - Tables in CSV format with annotations and FDR level of confidence. - (1) Weighted Pearson Correlation (2) Average difference of normalized intensities (3) Spatial chaos (4) FDR	- Python - Install from github/ Built-in functionality in METASPACE - Apache 2.0	(Palmer et al., 2016)

Name	**Confidence level**	- **Annotation type** - **Target features**	-**Approach** - **Library**	-**Input data format** -**Output** **- Ranking scores**	- **Programming language** - **Installation** - **License**	**Ref.**
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 - (1) 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. - (1) Linear Regression R^2 (2) M+0/M+1 ratio difference (3) Mass error	- 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 - (1) Weighted Pearson's Correlation (2) Exponential of Euclidean distance	- R - Install from github - GNU GPL v3.0	(Baquer et al., 2020)

8. References

Abdelhamid, H.N. (no date) 'Nanoparticle assisted laser desorption/ionization mass spectrometry for small molecule analytes'. 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.

Al-Saad, K.A. *et al.* (2003) 'Structural analysis of phosphatidylcholines by post-source decay matrix-assisted laser desorption/ionization time-of-flight mass spectrometry', *Journal of the American Society for Mass Spectrometry*, pp. 373–382. doi:10.1016/s1044-0305(03)00068-0.

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.

Bednařík, A. *et al.* (2022) 'Mass Spectrometry Imaging Techniques Enabling Visualization of Lipid Isomers in Biological Tissues', *Analytical chemistry*, 94(12), pp. 4889–4900. doi:10.1021/acs.analchem.1c05108.

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.

Bokhart, M.T. *et al.* (2018) 'MSiReader v1.0: Evolving Open-Source Mass Spectrometry Imaging Software for Targeted and Untargeted Analyses', *Journal of the American Society for Mass Spectrometry*, 29(1), pp. 8–16. doi:10.1007/s13361-017-1809-6.

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.

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.

Duncan, K.D. *et al.* (2018) 'Quantitative Mass Spectrometry Imaging of Prostaglandins as Silver Ion Adducts with Nanospray Desorption Electrospray Ionization', *Analytical chemistry*, 90(12), pp. 7246–7252. doi:10.1021/acs.analchem.8b00350.

Easterling, M.L., Mize, T.H. and Amster, I.J. (1999) 'Routine Part-per-Million Mass Accuracy for High- Mass Ions: Space-Charge Effects in MALDI FT-ICR', *Analytical chemistry*, 71(3), pp. 624–632. doi:10.1021/ac980690d.

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.

Eikel, D. *et al.* (2011) 'Liquid extraction surface analysis mass spectrometry (LESA-MS) as a novel profiling tool for drug distribution and metabolism analysis: the terfenadine example', *Rapid communications in mass spectrometry: RCM*, 25(23), pp. 3587–3596. doi:10.1002/rcm.5274.

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.

Gorshkov, M.V. *et al.* (2010) 'Calibration function for the Orbitrap FTMS accounting for the space charge effect', *Journal of the American Society for Mass Spectrometry*, 21(11), pp. 1846–1851. doi:10.1016/j.jasms.2010.06.021.

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. *et al.* (2013) 'LipidBlast in silico tandem mass spectrometry database for lipid identification', *Nature methods*, 10(8), pp. 755–758. doi:10.1038/nmeth.2551.

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.

Muddiman, D.C. and Oberg, A.L. (2005) 'Statistical evaluation of internal and external mass calibration laws utilized in fourier transform ion cyclotron resonance mass spectrometry', *Analytical chemistry*, 77(8), pp. 2406–2414. doi:10.1021/ac048258l.

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.

Robichaud, G. *et al.* (2013) 'MSiReader: an open-source interface to view and analyze high resolving power MS imaging files on Matlab platform', *Journal of the American Society for Mass Spectrometry*, 24(5), pp. 718–721. doi:10.1007/s13361-013-0607-z.

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.

Smith, D.F. *et al.* (2012) 'Advanced mass calibration and visualization for FT-ICR mass spectrometry imaging', *Journal of the American Society for Mass Spectrometry*, 23(11), pp. 1865–1872. doi:10.1007/s13361-012-0464-1.

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.

Swales, J.G. *et al.* (2015) 'Mapping drug distribution in brain tissue using liquid extraction surface analysis mass spectrometry imaging', *Analytical chemistry*, 87(19), pp. 10146–10152. doi:10.1021/acs.analchem.5b02998.

Takeo, E. *et al.* (2019) 'Tandem Mass Spectrometry Imaging Reveals Distinct Accumulation Patterns of Steroid Structural Isomers in Human Adrenal Glands', *Analytical chemistry*, 91(14), pp. 8918–8925. doi:10.1021/acs.analchem.9b00619.

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.

Wäldchen, F., Spengler, B. and Heiles, S. (2019) 'Reactive Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry Imaging Using an Intrinsically Photoreactive Paternò-Büchi Matrix for Double-Bond Localization in Isomeric Phospholipids', *Journal of the American Chemical Society*, 141(30), pp. 11816–11820. doi:10.1021/jacs.9b05868.

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, H. *et al.* (2020) 'On-Tissue Derivatization with Girard's Reagent P Enhances N-Glycan Signals for Formalin-Fixed Paraffin-Embedded Tissue Sections in MALDI Mass Spectrometry Imaging', *Analytical chemistry*, 92(19), pp. 13361–13368. doi:10.1021/acs.analchem.0c02704.

Zhang, H. *et al.* (2021) 'Quantification and molecular imaging of fatty acid isomers from complex biological samples by mass spectrometry', *Chemical science* , 12(23), pp. 8115–8122. doi:10.1039/d1sc01614h.

Zhang, L.-K. *et al.* (2005) 'Accurate mass measurements by Fourier transform mass spectrometry', *Mass spectrometry reviews*, 24(2), pp. 286–309. doi:10.1002/mas.20013.

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, R. and Basile, F. (2017) 'Plasmonic Thermal Decomposition/Digestion of Proteins: A Rapid On-Surface Protein Digestion Technique for Mass Spectrometry Imaging', *Analytical chemistry*, 89(17), pp. 8704–8712. doi:10.1021/acs.analchem.7b00430.

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.

CHAPTER 3

rMSIKeyIon: an ion filtering R package for untargeted analysis of metabolomic LDI-MS images

Abstract: Many MALDI-MS imaging experiments make case vs. control studies of different tissue regions in order to highlight significant compounds affected by the variables of study. This is a challenge because the tissue samples to be compared come from different biological entities and therefore exhibit high variability. Moreover, the statistical tests available cannot properly compare ion concentrations in two regions of interest (ROIs) within or between images. The high correlation between the ion concentrations due to the existence of different morphological regions in the tissue means that the common statistical tests used in metabolomics experiments cannot be applied. Another difficulty with the reliability of statistical tests is the elevated number of undetected MS ions in a high percentage of pixels.

In this study, we report a procedure for discovering the most important ions in the comparison of a pair of ROIs within or between tissue sections. These ROIs were identified by an unsupervised segmentation process, using the popular k-means algorithm. Our ion filtering algorithm aims to find the up or down-regulated ions between two ROIs by using a combination of three parameters: a) the percentage of pixels in which a particular ion is not detected, b) the U Mann Whitney ion concentration test, and c) the ion concentration fold-change. With this methodology we found the important ions between the different segments of a mouse brain tissue sagittal section and determined some lipid compounds (mainly triacylglycerols and phosphatidylcholines) in the liver of mice exposed to thirdhand smoke.

1. Introduction

Mass Spectrometry Imaging (MSI) is a label-free analytical technique that can locate chemical compounds (metabolites, peptides, lipids or proteins) directly in a biological sample and give their concentration for every pixel. The most common analytical strategy is matrix-assisted laser desorption ionization (MALDI) due to its soft ionization, fast analysis, high throughput, versatility, and selectivity [1]. Other techniques like desorption electrospray ionization (DESI) are becoming more popular because of the simplicity of their sample preparation [2]. MSI is currently used in the fields of drug discovery and toxicology [3], [4]. In most experiments, researchers use a targeted strategy, which consists of visualizing and (sometimes) quantifying the concentration of a particular compound, or a reduced set of compounds throughout the tissue. Many MSI software packages have been released [5]. Even though, none of them provides an automated workflow for untargeted MSI applications since the end-user must approach each MSI experiment data analysis in its unique manner.

Besides annotating and identifying the MS ions, one of the main challenges in untargeted MSI analysis is to determine the statistically differentiating ions in different ROIs of the same tissue section or in different tissues of case vs. control experiments. These key ions could be associated with biomarker candidates of disease or treatment efficacy. Previous studies have successfully used segmentation processes to find these key ions between clusters [6], [7]. Most of these studies identify the key ions associated with a certain region by analyzing the ions that most influence the segmentation process. In [8], the authors applied a Nonnegative Matrix Factorization multivariate analysis to select a reduced group of lipid MS signals associated with the metabolite profile of each component. The t-test associated with segmentation with Spatial Shrunken Centroids can find the enriched and absent MS peaks for a particular region in a segmented image [9], [10]. A technique based on deep unsupervised neural networks and parametric t-SNE was used to detect metabolic hidden sub-regions [11]. The same algorithm, linked to a significance analysis of microarrays (SAM), detected the protein subpopulations that can differentiate between t-SNE segments in a dataset of breast cancer samples; interestingly, they used the selected ions for a kNN second segmentation step [12]. Gorzolka et al. [13] studied the space-time profiling of the barley germination process by carrying out an unsupervised joint segmentation of a high number of images and found the ion-associated

profile for every segment. The Algorithm for MSI Analysis by Semi-supervised Segmentation (AMASS) was used to segment leech embryo samples [14] and there is a complete analysis of the ions associated to every region according to its weighting factors. In all these references, no statistical significance test was conducted on the key ions found.

Another common strategy in MSI data analysis is to manually define the ROIs to be compared, guided by an annotated histology image [15]–[18]. In general, the ions are selected by means of statistical hypothesis testing and the fold change calculation of the ion concentrations between ROIs. These parameters are usually represented as volcano plots. By way of example, Hong et al. [19] studied the global changes of phospholipids in brain samples from a mouse model of Alzheimer disease by performing ANOVA tests of ion concentrations in regions of interest. A common problem that MSI has in calculating statistical significance is that the p-values are generally extremely low [16]. This is because there are a large number of pixels within each ROI, which gives this parameter a low discrimination power.

Additionally, the statistical hypothesis testing (such as the t-test) fails when applied to compare the concentration of an ion between ROI's. The existence of morphological areas in the images is responsible for a high pixel autocorrelation. This violates the assumption of observation independence necessary for statistical hypothesis testing. To find statistically significant ions between ROIs, Conditional Autoregressive (CAR) models, which consider the autocorrelated nature of ion distribution concentration in MS image ROIs, are calculated to correct the p-values [20]. Nevertheless, the difficulty of calculating the autocorrelation models and the complexity of the computational approach hampers the inclusion of this strategy in a MSI workflow.

Another common situation in MS imaging is the elevated intensity differences of the ion's concentration between pixels, due to the existence of several morphologic regions with different metabolic profiles [21] and the ion shielding phenomena which takes place in MSI. It is also common to find a high proportion of pixels where a certain ion is not detected, for a given signal to noise ratio. This influences to a large extent the calculation of the p-values and the fold change.

In this study, we describe the development of an ion filtering algorithm that is used in a workflow for the untargeted analysis of metabolomic MALDI-MS images. The workflow consists of a segmentation step, followed by the ion filtering procedure, independent of the segmentation process, that detects the up/down regulated ions between image segments. Our algorithm calculates and combines three parameters: a) the Mann-Whitney statistical test of the ion concentration between segments [22]; b) the fold change in the ion concentration between segments; and c) a new parameter that accounts for the proportion of pixels with undetected ions between segments. With this methodology, we can find the key ions between any segment pair in MSI datasets, from single or multiple tissue sections. We successfully applied this workflow to the analysis of mouse brain tissue samples and to study fatty liver disease in mice liver tissue samples.

2. Results

The rMSIKeyIon package, written in R, is able to find the key ions in a pair of ROI's within or between images. The ions are selected according to the similarity parameters calculated in Appendix A and ordered following the contrast parameter, described in Appendix B. In the next section we will describe the results of the package in the analysis of a sagittal brain mouse sample, which has been segmented by k-means algorithm (section 2.1). In particular, we will illustrate the up or down regulated ions resulting from the comparison of two clusters and the up/down regulated ions when comparing one cluster with the rest.

In section 2.2 we will apply the package in the identification of the fat areas in control liver samples and liver samples exposed to third hand smoke (THS).

2.1 Results of the brain mouse sample

Figure 1 shows the number of up- and down-regulated ions associated with the comparison of one cluster with each of the others (columns 1 to 7) in the segmented image of the brain slice tissue of C57BL/6 mouse using the k-means algorithm (n = 7 clusters). In column "All" appear the ions that are up-regulated (or down-regulated) in a cluster because of the comparison between this cluster and the rest of clusters, called "absolutely up-regulated ions" (or "absolutely down-regulated ions")

For each cluster comparison, an associated figure gives information about the resulting up- or down-regulated ions, and the number of null and non-null parameters defined in the section Ion analysis and filtering (see below). The ions on the list are ordered in terms of the value of the "contrast parameter", calculated with Equation B1 in Appendix B.

Cluster Index	1	2	3	4	5	6	7	All
1	Up / Down	58 / 78	39 / 3	0 / 17	0 / 23	0 / 28	125 / 0	—
2	79 / 57	Up / Down	3 / 0	47 / 60	48 / 69	59 / 76	127 / 0	2 / 12
3	2 / 47	0 / 3	Up / Down	1 / 47	1 / 55	6 / 61	104 / 5	0 / 1
4	17 / 13	60 / 52	46 / 6	Up / Down	0 / 1	0 / 9	129 / 9	—
5	23 / 13	72 / 51	55 / 7	2 / 0	Up / Down	0 / 4	131 / 9	—
6	28 / 22	77 / 71	63 / 22	8 / 10	3 / 10	Up / Down	121 / 15	2 / 6
7	0 / 165	0 / 149	0 / 142	0 / 166	0 / 168	0 / 155	Up / Down	0 / 123

Figure 1. Number of up or down-regulated ions associated with the comparison of one particular cluster with each of the others (columns 1 to 7) and the ions that are up-regulated (or down-regulated) in a cluster as a result of the comparison between this cluster and the rest of clusters, called "absolutely up-regulated ions" (or "absolutely down-regulated ions"). The image is composed of 6898 pixels and the detected ion number is 277. The percentile value used for the selection of the ions is 8 % for the null concentration parameter and 10 % for the U Mann-Whitney test and for the concentration fold change (FC). The intensity threshold for the ions is $1 \cdot 10^{-3}$ over the normalized spectra matrix. The lack of symmetry observed in the table is a consequence of the lack of symmetry in the distributions considered.

Comparison of C2&C6

By way of example, the comparison of clusters C2&C6 showed 59 up-regulated ions in C2 vs. C6 and 76 down-regulated ions in C2 vs. C6. As an example, Figure S1 shows the volcano plot of the ions resulting from the comparison of C2 & C6. The ions at the top right and top left are selected by the ion filtering algorithm (see the caption to Figure S1 for more details). Figure S2.a shows the histogram of the concentration of the up-regulated ion with the highest contrast parameter (m/z 198.076) in C6 and Figure S2.b shows the histogram of the up-

regulated ion (m/z 848.636) in C2 also with the highest contrast parameter. Figure 2a shows the segmented brain image (n=7) and Figures 2b and c the concentration intensity plot of the ions mentioned above. In these intensity maps, the contrast intensity between both ions and clusters is clear, and the intensity of m/z 848.636 is much higher in C2 than in C6 and vice-versa for m/z 198.076.

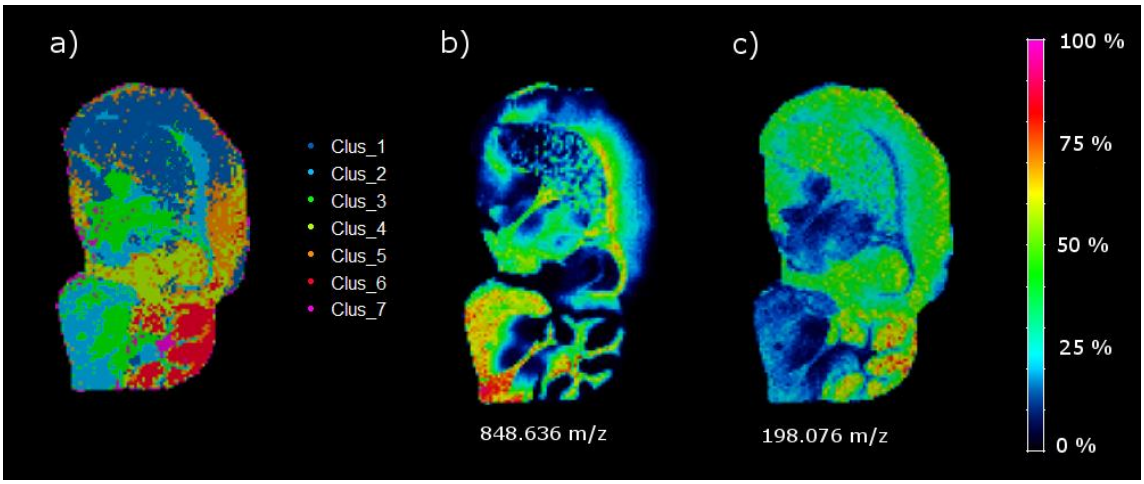


Figure 2. (a) Mouse brain segmentation using k-means (n = 7 clusters), (b) Intensity map of ion m/z 848.636 (the up-regulated ion in C2 vs C6 with the highest contrasting parameter extracted from the null concentration parameter) and (c) Intensity map of ion m/z 198.076, the down-regulated ion with the highest contrast parameter after comparing C2 & C6, extracted from the volcano plot.

Absolutely up and down-regulated ions in brain

According to the results in Figure 1, there are two absolutely up-regulated ions in C2, and 123 absolutely down-regulated ions in C7. Figure 3 shows the concentration intensity plot of the two up-regulated ions (m/z 832,644; m/z and m/z 834,654) in C2 and Figure 4 shows the three down-regulated ions (m/z 274,792; m/z 298,811 and m/z 258,822) in C7 with the highest contrast parameter. There is an evident similarity between the images of the two up-regulated ions for one hand and three down-regulated ones for the other one. A comparison of the images in Figure 3 with the distribution of C2 in the brain are clearly similar. And the same is true of a comparison of the images in Figure 4 with the distribution of C7 in the brain.

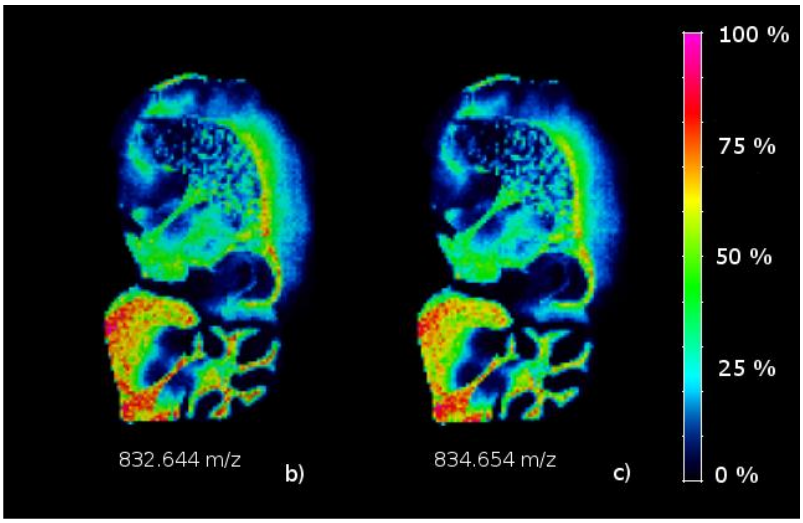


Figure 3. Concentration images of the two absolutely up-regulated ions in C2

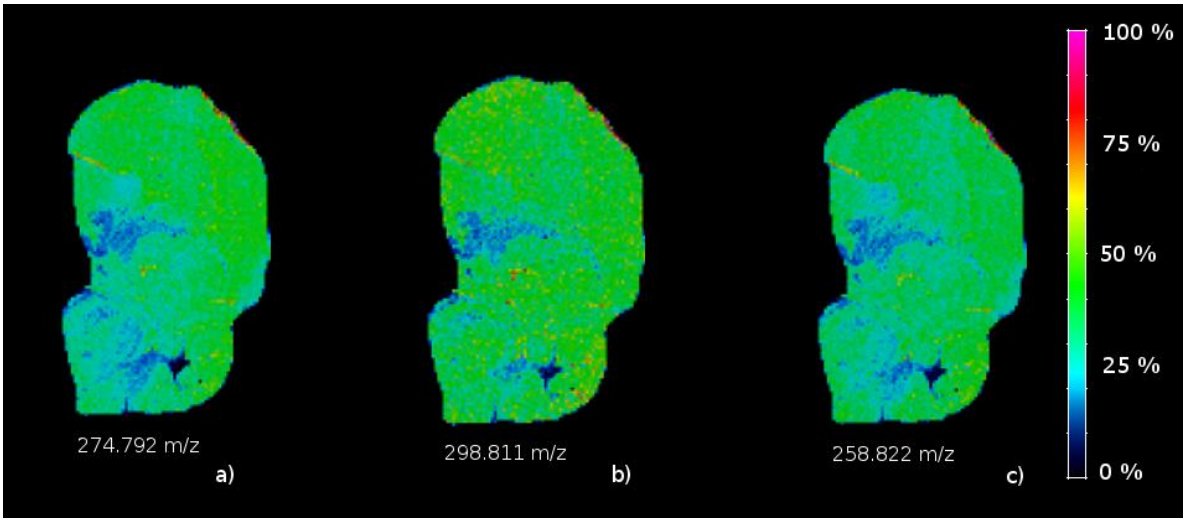


Figure 4. Concentration images of three absolutely down-regulated ions in C7

2.2 Results of the liver samples

The methodology used in this article has been applied to the study of non-alcoholic fatty liver disease in mice exposed to thirdhand tobacco smoke [23]. We have taken a total of 6 images from the liver samples (three from a control mouse and three from a THS-exposed mouse). The images have been segmented using the k-means algorithm (n = 6 clusters).

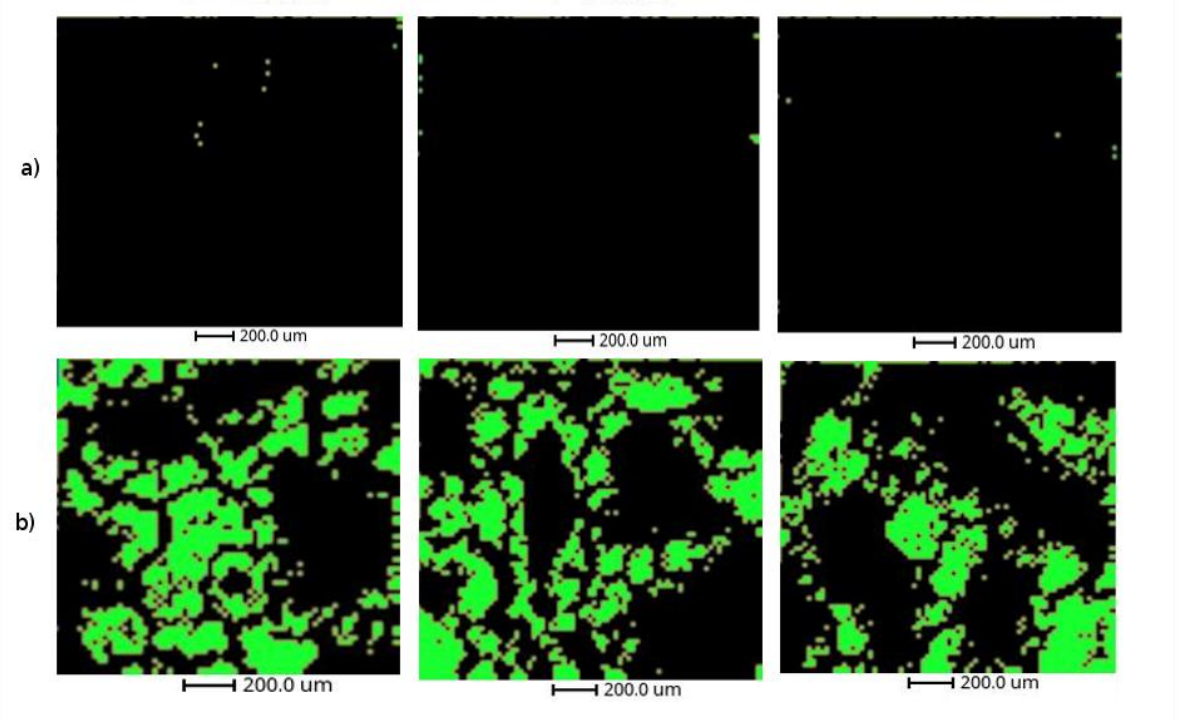


Figure 5. Representation of cluster 2 of the 6 liver samples: a) the three analytical replicates of a control mouse and b) the three replicates of a THS-exposed mouse.

The results of rMSIKeyIon algorithm showed that cluster 2 (C2) has an elevated number of ions in the lipid mass range, that are absolutely up-regulated and we hypothesised that this

cluster represents the lipid droplet areas characteristic of the fatty livers (see Figure 5). The THS exposed mouse has the largest area while the control animals have the smallest, in accordance with Martins-Green et al. [23].

Table S1 shows the compounds in C2 putatively identified after a manual curation process. As can be observed, most of them are putatively identified as TG or phosphatidylcholine. In Figure S3 there is the intensity map of the triacylglycerol (50:30), which is highly similar to the geometry of C2

3. Discussion

Here we developed a new methodology for the untargeted analysis of MS images that can be used coupled with any segmentation process and an ion filtering algorithm based on the combination of three parameters: a) The ratio of ions with a null concentration between the regions, b) the U Mann-Whitney Test, calculated by segregating the non-detected ions from the distribution, and c) the fold change between the medians of the distribution (the non-detected ions were also segregated from the distribution). This methodology has proved to be efficient at finding the up/down-expressed ions in an intra-image analysis or in the comparative analysis of groups of images. The presented workflow is different to previously released software tools due to two main reasons: a) it is flexible and independent to the segmentation process, so the ion selection process can be applied to any clustering algorithm or manually drawn ROI's. b) Our methodology provides a completely automated ion filtering approach enabling the fast detection of a morphological region characteristic ions.

The results on the sagittal mouse brain sample show that an unsupervised clustering process followed by the rMSIKeyIon algorithm can select the (possible) up/down-regulated ions between any pair of clusters, in a holistic approach, and between one cluster and the rest. The concentration maps of the selected ions, ordered by the contrast parameter, depicts faithfully the morphology of the brain. These ions are probably biologically relevant and could be interesting to identify.

Using the described methodology, we have been able to detect the regions containing the lipid droplets in the liver samples from mice exposed to THS. The putative identification of the key up-regulated ions in the cluster 2, mainly triglycerides and phosphatidylcholines, confirm that THS exposure conducts to the apparition of fatty liver disease in mice [23].

Untargeted metabolomics data analysis workflows are associated with standard analytical platforms (LC-MS, GC-MS and NMR) [24]. These analyses compare the concentrations of chemical compounds in a CASE and a CONTROL group in order to discover features that they express differently, and which could be used as biomarkers or in biological pathway analysis. In general, the number of samples (n) of each experimental group are similar, the distribution is normal (for large n values) and the principle of independent measures is assumed. However, in spatial metabolomics, the number of samples in every group (i.e. the number of pixels in a ROI) is not determined a priori, as in metabolomics studies.

Untargeted image analysis has two main applications:

a) The comparison of two regions inside the same tissue section (intra-image analysis) to find the relevant ions. This could be used to discover cancer biomarkers by comparing the ion profile of the tumorous area with a non-tumorous area from the same sample. In general, the areas to be compared are determined by a histopathologist annotating a consecutive tissue section. The size of the ROIs in which we will compare the ions is determined manually.

b) For several reasons the analysis of morphologically equivalent regions in different tissues in a case-control experiment is much more complicated. First, the tissue samples to be compared between groups are equivalent but not similar because of the biological differences between the animals and the intrinsic difficulty of achieving identical tissue sections.

Consequently, it is not straightforward to delimit the areas to be compared. The ROIs to be compared can be determined by histological annotation (supervised process), or automatically by means of a segmentation process (unsupervised process). In both cases, there are not established rules, and the following steps in the statistical analysis of the ions between ROI's can be highly affected by this fact.

In both cases, it is very common to find skewed ion distributions and a high percentage of null values, a high degree of autocorrelation between pixels and a very high number of observations (pixels). This leads to extremely low p-values when classical parametric or non-parametric statistical tests are used [25] so these tests are not appropriate for this kind of analysis. All the above reasons make the untargeted analysis of images. However, the results shown by rMSIKeyIon R package have been revealed to be very useful to find the most differential ions between ROI's. The biological relevance of these ions has been validated in a fatty liver study with animal models.

4. Materials and methods

4.1 Materials

Indium tin oxide (ITO)-coated glass slides were obtained from Bruker Daltonics (Bremen, Germany). The gold target used for sputtering coating was obtained from Kurt J. Lesker Company (Hastings, England) with a purity grade higher than 99.995%. HPLC grade xylene was supplied by Sigma-Aldrich (Steinheim, Germany) and ethanol (96% purity) was supplied by Scharlau (Sentmenat, Spain).

4.2 Methods

4.2.1 Sample preparation

Mice models were developed at the Department of Molecular, Cell and Systems Biology at the University of California Riverside [23]. Animal experimental protocols were approved by the University of California, Riverside, Institutional Animal Care and Use Committee (IACUC). The suitability of the workflow presented here to determine significant ions between ROIs from the same tissue was tested in a brain sample from a 6-month-old C57BL/6 mouse feed with a standard chow diet (percent calories: 58% carbohydrates, 28.5% protein, and 13.5% fat). To test the suitability of the method in different tissue sections in a case vs. control experiment, we used liver samples from mice exposed to thirdhand tobacco smoke (THS) – the residual particles and gases from tobacco smoke that remain in dust and surfaces – from weaning (three weeks of age) to 24 weeks, without exposure to SHS at any time during the study, and compared them with liver samples of mice that had not been exposed to THS (control group) [26]. Brain and liver samples were snap frozen at -80°C after collection and stored and shipped at this temperature until analysis.

For MSI acquisition, the tissues were sectioned at -20°C in slices 10 µm thick using a Leica CM-1950 cryostat (Leica Biosystems Nussloch GmbH) located at the Centre for Omics Sciences (COS) of the Rovira i Virgili University and mounted on indium-tin oxide-coated (ITO) slides by directly placing the glass slide onto the section at ambient temperature. To remove residual humidity, samples were dried in a desiccator under vacuum for 15 minutes after tissue mounting.

4.2.2 Deposition of Au nanolayers for LDI-MS imaging

Gold nanolayers were deposited on the 10 µm tissue sections using an ATC Orion 8-HV sputtering system (AJA International, N. Scituate, MA, USA) [27]. Briefly, an argon atmosphere with a pressure of 30 mTor was used to create the plasma in the gun. The working distance of the plate was set to 35 mm. Sputtering conditions for MS were ambient temperature, and RF mode at 60 W for 50 s. The argon ion current was adjusted to 20 mL min^{-1}.

4.2.3 LDI-MS acquisition

One image of a sagittal brain tissue section and six liver tissue sections (3 slices from a control animal and 3 sections from a THS-exposed animal) were acquired using a MALDI TOF/TOF UltrafleXtreme instrument with SmartBeam II Nd:YAG/355 nm laser from Bruker Daltonics, also at the COS facilities. Raster sizes of 80 and 20 µm were used for the brain and liver tissue sections, respectively. The TOF spectrometer operated in reflectron positive mode with the digitizer set at a sample rate of 1.25 GHz in a mass range between 70 and 1.200 Da. The spectrometer was calibrated prior to tissue image acquisitions using $[Au]^+$ cluster MS peaks as internal mass references [27].

4.2.4 MSI data processing and image segmentation

The MSI data acquired with Bruker's FlexImaging 3.0 software was exported to XMASS data format using instrument manufacturer software packages (FlexImaging and Compass export). The raw data was loaded using the in-house rMSI package [28]. This package provides a data storage format based on segmented matrices and optimized for processing large MSI datasets in R language. Next, we applied our complete MSI pre-processing workflow consisting of spectral smoothing, alignment, mass recalibration, peak detection, and peak binning [29] with the default parameters: Savitzky-Golay kernel size of 7, peak detection threshold SNR of 5 and peak binning tolerance of 6 scans with 5% filter. At this point we obtained a peak matrix object of each MSI dataset: the brain tissue sagittal section and the liver tissue sections. These peak matrix objects are highly reduced, robust and accurate representations of all the MSI data, and can be used to perform complex statistical analyses on the huge amount of data generated in the MSI experiment. ROIs were generated by means of a k-means process. Finally, we applied the rMSIKeyIon workflow using the peak matrices as the input data.

4.2.5 Ion analysis and filtering

The procedure used for identifying statistically different ions compared the concentration distributions of the ions in all possible pairs of ROIs in which the tissue (or tissues) had been segmented.

In general, the total number of pixels in each ROI is different and the probability density function of the ion concentrations is not normal. We used the U Mann-Whitney test [30] because it can test the null hypothesis (both sets of samples come from the same distribution) of two non-normal distributions that have a different number of observations.

In addition, in non-normal distributions of different sample sizes, there is usually a singular element: in some ROIs there is a considerable possibility that the distribution of some ions will have small concentration values. Fig. S4 represents the percentage of non-detected ions in the segmented brain image, using the k-means algorithm with n= 7 clusters. It can be observed that for some clusters (i.e cluster 7) the percentage is very high.

For purposes of illustration, Figure S5 shows two synthetic histograms with samples taken from normal distributions, with different average values, to which significant amounts of null

values have been added. In total, there are 200 samples for both cases. Both distributions appear to be very different, and the Mann-Witney test yields a very high p-value (0.38). The idea we have worked on here is to segregate the values obtained from non-detected ions (nulls) from the rest of the distribution so that they can be treated separately. Thus, we obtain a very small p-value (of the order of 1e-43). On the other hand, the % of null values in each ROI also provides valuable information. For these reasons, we decided to segregate the null values from the ion matrix and use them to calculate a parameter (null concentration parameter), as will be explained below.

The calculation of the null concentration parameter, and the non-null parameters (U Mann Whitney distribution and fold change) are described in Appendix A.

Once the ions had been selected using the two procedures described above, they were ordered in terms of the contrast generated by every ion between one ROI and the set of other ROIs. The procedure is described in Appendix B.

The ion filtering algorithm described in this section has been implemented in the R package named rMSIkeyIons, accessible at (https://github.com/LlucSF/rMSIKeyIon). The software's source code was written in C++ and requires the GSL library. Later, it was ported to R using the Rcpp R package. As input, the function requires an rMSIproc peak matrix, a previously calculated segmentation, and the percentiles for each parameter, and as output, the function returns a list containing the ions for each comparison between all pair of clusters and the data related with those ions.

4.2.6 Metabolite identification

The obtained list of up regulated lipids for mice liver samples in Cluster 2 was matched with the HMDB 4.0 [31] database within a tolerance of 20 ppm and the possible ion adducts: H, Na, K and NH4. Results were filtered using the biological information of molecules provided by the HMDB, thus metabolites with no biological origin or not likely to be found in the liver were discarded.

5. Conclusions

In this study we have developed the ion filtering R package rMSIkeyIon. It is open source, publicly available and based on the combination of three parameters: the non-detected ion concentration ratio, the Mann-Whitney ion concentration test and the fold change in the ion concentration. We demonstrated that our tool is very effective at discovering up or down-regulated ions between clusters using an unsupervised k-means procedure. The ions selected are the candidates that subsequently have to be identified. This package is a valuable tool for the untargeted analysis of MALDI images and is an important advance in this area because, at present, there are no tools available.

6. References

[1] M. Karas and F. Hillenkamp, "Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons," *Anal. Chem.*, vol. 60, no. 20, pp. 2299–2301, Oct. 1988.

[2] J. M. Wiseman, D. R. Ifa, Q. Song, and R. G. Cooks, "Tissue Imaging at Atmospheric Pressure Using Desorption Electrospray Ionization (DESI) Mass Spectrometry," *Angew. Chemie Int. Ed.*, vol. 45, no. 43, pp. 7188–7192, Nov. 2006.

[3] L. Morosi, M. Zucchetti, M. D'Incalci, and E. Davoli, "Imaging mass spectrometry: challenges in visualization of drug distribution in solid tumors.," *Curr. Opin. Pharmacol.*, vol. 13, no. 5, pp. 807–12, Oct. 2013.

[4] T. Greer, R. Sturm, and L. Li, "Mass spectrometry imaging for drugs and metabolites," *J. Proteomics*, vol. 74, no. 12, pp. 2617–2631, Nov. 2011.

[5] P. Ràfols *et al.*, "Signal preprocessing, multivariate analysis and software tools for MA(LDI)-TOF mass spectrometry imaging for biological applications," *Mass Spectrom. Rev.*, vol. 37, no. 3, pp. 281–306, May 2018.

[6] T. Alexandrov, "MALDI imaging mass spectrometry: statistical data analysis and current computational challenges.," *BMC Bioinformatics*, vol. 13 Suppl 1, no. Suppl 16, p. S11, 2012.

[7] E. A. Jones, S.-O. Deininger, P. C. W. Hogendoorn, A. M. Deelder, and L. A. McDonnell, "Imaging mass spectrometry statistical analysis.," *J. Proteomics*, vol. 75, no. 16, pp. 4962–89, Aug. 2012.

[8] D. Y. Lee *et al.*, "Resolving brain regions using nanostructure initiator mass spectrometry imaging of phospholipids.," *Integr. Biol. (Camb).*, vol. 4, no. 6, pp. 693–9, Jun. 2012.

[9] K. D. Bemis *et al.*, "Probabilistic Segmentation of Mass Spectrometry (MS) Images Helps Select Important Ions and Characterize Confidence in the Resulting Segments," *Mol. Cell. Proteomics*, vol. 15, no. 5, pp. 1761–1772, 2016.

[10] K. D. Bemis *et al.*, "Cardinal: an R package for statistical analysis of mass spectrometry-based imaging experiments," *Bioinformatics*, vol. 31, no. 14, pp. 2418–2420, Jul. 2015.

[11] P. Inglese *et al.*, "Deep learning and 3D-DESI imaging reveal the hidden metabolic heterogeneity of cancer," *Chem. Sci.*, vol. 8, pp. 3500–3511, 2017.

[12] W. M. Abdelmoula *et al.*, *Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data*, vol. 113, no. 43. 2016.

[13] K. Gorzolka, J. Kölling, T. W. Nattkemper, and K. Niehaus, "Spatio-Temporal metabolite profiling of the barley germination process by MALDI MS imaging," *PLoS One*, vol. 11, no. 3, pp. 1–25, 2016.

[14] J. Bruand *et al.*, "AMASS: algorithm for MSI analysis by semi-supervised segmentation.," *J. Proteome Res.*, vol. 10, no. 10, pp. 4734–43, Oct. 2011.

[15] E. Moreno-Gordaliza *et al.*, "Lipid imaging for visualizing cilastatin amelioration of cisplatin-induced nephrotoxicity," *J. Lipid Res.*, vol. 59, no. 9, pp. 1561–1574, 2018.

[16] Y. Yajima *et al.*, "Region of Interest analysis using mass spectrometry imaging of mitochondrial and sarcomeric proteins in acute cardiac infarction tissue," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, 2018.

[17] X. Wang, J. Han, D. B. Hardie, J. Yang, J. Pan, and C. H. Borchers, "Metabolomic profiling of prostate cancer by matrix assisted laser desorption/ionization-Fourier transform ion cyclotron resonance mass spectrometry imaging using Matrix Coating Assisted by an Electric Field (MCAEF)," *Biochim. Biophys. Acta - Proteins Proteomics*, vol. 1865, no. 7, pp. 755–767, 2017.

[18] Y. Otsuka, S. Satoh, J. Naito, M. Kyogaku, and H. Hashimoto, "Visualization of cancer-related chemical components in mouse pancreas tissue by tapping-mode scanning probe electrospray ionization mass spectrometry," *J. Mass Spectrom.*, vol. 50, no. 10, pp. 1157–1162, 2015.

[19] J. H. Hong *et al.*, "Global changes of phospholipids identified by MALDI imaging mass spectrometry in a mouse model of Alzheimer's disease," *J. Lipid Res.*, vol. 57, no. 1, pp. 36–45, Jan. 2016.

[20] A. Cassese *et al.*, "Spatial Autocorrelation in Mass Spectrometry Imaging," *Anal. Chem.*, vol. 88, no. 11, pp. 5871–5878, 2016.

[21] I. Chernyavsky, S. Nikolenko, F. von Eggeling, T. Alexandrov, and M. Becker, "Analysis and Interpretation of Imaging Mass Spectrometry Data by Clustering Mass-to-Charge Images According to Their Spatial Similarity," *Anal. Chem.*, vol. 85, no. 23, pp. 11189–11195, 2013.

[22] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *Ann. Math. Stat.*, vol. 18, no. 1, pp. 50–60, Mar. 1947.

[23] M. Martins-Green *et al.*, "Cigarette smoke toxins deposited on surfaces: Implications for human health," *PLoS One*, vol. 9, no. 1, pp. 1–12, 2014.

[24] G. J. Patti, O. Yanes, and G. Siuzdak, "Metabolomics: the apogee of the omics trilogy," *Nat. Rev. Mol. Cell Biol.*, vol. 13, no. 4, pp. 263–269, Mar. 2012.

[25] M. W. Fagerland, "t-tests, non-parametric tests, and large studies—a paradox of statistical practice?," *BMC Med. Res. Methodol.*, vol. 12, no. 1, p. 78, Dec. 2012.

[26] N. Adhami, S. R. Starck, C. Flores, and M. M. Green, "A health threat to bystanders living in the homes of smokers: How smoke toxins deposited on surfaces can cause insulin resistance," *PLoS One*, vol. 11, no. 3, pp. 1–19, 2016.

[27] P. Ràfols *et al.*, "Assessing the potential of sputtered gold nanolayers in mass spectrometry imaging for metabolomics applications," *PLoS One*, vol. 13, no. 12, p. e0208908, Dec. 2018.

[28] P. Ràfols *et al.*, "rMSI: an R package for MS imaging data handling and visualization," *Bioinformatics*, vol. 33, no. 15, Mar. 2017.

[29] P. Ràfols, E. del Castillo, O. Yanes, J. Brezmes, and X. Correig, "Novel automated workflow for spectral alignment and mass calibration in MS imaging using a sputtered Ag nanolayer," *Anal. Chim. Acta*, vol. 1022, pp. 61–69, Aug. 2018.

[30] M. Statistics, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other Author (s): H . B . Mann and D . R . Whitney Source : The Annals of Mathematical Statistics , Vol . 18 , No . 1 (Mar ., 1947), pp . 50-60 Published by : Institute," vol. 18, no. 1, pp. 50–60, 2019.

[31] D. S. Wishart *et al.*, "HMDB 4.0: The human metabolome database for 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D608–D617, 2018.

[32] T. D. Mak, E. C. Laiakis, M. Goudarzi, and A. J. Fornace, "MetaboLyzer: A Novel Statistical Workflow for Analyzing Postprocessed LC–MS Metabolomics Data," *Anal. Chem.*, vol. 86, no. 1, pp. 506–513, Jan. 2014.

7. Supporting Information

7.1 Calculation of the similarity parameters between ROIs

To determine the ions that are expressed differently in two given ROIs, we calculate three parameters:

The null concentration parameter (Z parameter)

The Z_{ijk} parameter is calculated according to Eq. A1:

$$Z_{ijk} = \frac{\frac{Nz_{ij}}{N_j}}{\frac{Nz_{ik}}{N_k}} \quad \forall\, i \in I; \quad \forall\, j,k \in S_p \qquad \text{Eq. A1}$$

where Z_{ijk} is the parameter that accounts for the null values (i.e. the non-detected values) of the i ion when comparing the j and k ROIs; Nz_{ij} and Nz_{ik} are the number of pixels with null

values of the *i* ion in *j* and *k* ROIs, respectively; N_j and N_k are the total number of ROI pixels in *j* and *k*, respectively; *I* is the set of ions and *Sp* is the set of ROIs.

The equation calculates the ratio between the null values of a particular ion in the two ROIs. A value of $Z_{ijk} > Z_{high}$ (Z_{high} being a positive value greater than 1) means that the *i* ion is more expressed in *k* ROI than in *j* ROI, while $Z_{ijk} < Z_{low}$ (Z_{low} being a positive value much lower than 1) means that the *i* ion is less expressed in *k* ROI than in *j* ROI.

The importance of this parameter is assessed in Figure S4. For clusters 1 to 7, we plotted, the percentage of pixels that have null concentration for every ion.

The Z_{high} and Z_{low} values are calculated by following these steps:

The *Z* values of all ions, for all cluster-pairs, are calculated according to Eq. A1

An ordered rank list of all the *Z* values is created.

Z_{low} is determined considering that this value is a certain percentile P_Z of the rank list of Z values.

Z_{high} is determined considering that this value is a certain percentile 100 - P_Z of the rank list of Z values.

Non-null concentration parameters (V parameters)

Provided that the distribution of the ions concentration is non-normal, we considered the *U* Mann-Whitney test and the concentration fold change (FC) between two ROIs, as a non-null concentration parameter.

Generally speaking, if N_j and N_k are high, the random variable *U* can be regarded as normally distributed [30]. The U_{ijk} parameter is then normalized following Eq. A2:

$$V_{ijk} = \frac{U_{ijk} - m_u}{\sigma_u} \qquad \text{Eq.A2}$$

where m_u and σ_u are the average and standard deviation of U_{ijk} and V_{ijk} is a random variable with a normalized Gaussian distribution. If *V* has values close to 1 the similarity between the distributions is high, while values close to zero indicate disparate distributions. The value obtained for *V* indicates the similarity between the distributions of two ROIs for an ion.

Another parameter often used to compare sets of magnitudes is the fold change, defined as the ion median concentration quotient between two ROIs (Eq. A3):

$$FC_{ijk} = \frac{M_{ij}}{M_{ik}} \qquad \text{Eq. A3}$$

where M_{ij} is the distribution median of the *i* ion in *j* ROI and M_{ik} is the same for *k* ROI. For every *i* ion, the FC_{ijk} parameter is calculated between the *j* and *k* ROIs.

For a pair of ROIs, a Volcano plot [32] can be drawn from the *V* and *FC* parameters. In this representation, the position occupied by the ions is important: the ions located in the top corners generate very different distributions in the two ROIs. The ions at the top left are under-expressed ($V_{ijk} < V_{high}$ & $Fc_{ijk} < Fc_{low}$) and the ions at the top right are over-expressed ($V_{ijk} < V_{high}$ & $Fc_{ijk} > Fc_{high}$).

The values V_{high}, Fc_{high} and Fc_{low} are calculated following the same steps as for Z_{high} and Z_{low}, but with a difference in the percentile value. The ions located in the areas of interest must satisfy the probability of being within a range associated with two random variables; that is to say:

$P[V_{ijk} \leq V_{high}, Fc_{ijk} \leq Fc_{low}]$ for under-expressed ions and $P[V_{ijk} \leq V_{high}, Fc_{ijk} \geq Fc_{high}]$ for over-expressed ions. Assuming that these are independent random variables, we obtain $P[V_{ijk} \leq V_{high}] = P[Fc_{ijk} \leq Fc_{low}] = P[Fc_{ijk} \geq Fc_{high}] = \sqrt{P_z/100}$. That is, the percentile that must be used to determine the cutoff values in the volcano plot should be $P_V = 10 * \sqrt{P_Z}$

7.2 Determination of the discriminating figure values and generation of the discriminant ions lists

The *contrast parameter* $C_{ij|S_p}$ of the *i* ion between the *j* ROI and all the ROIs (set S_p) is calculated according to Eq. B1:

$$C_{ij|S_p} = \frac{\frac{1}{N_j}\sum_{p=1}^{N_j} m_{ip}^{j}}{\frac{1}{N}\sum_{k=0}^{N_{S_p}}\sum_{p=1}^{N_k} m_{ip}^{k}} \qquad \text{Eq. B1}$$

where *N* are the total number of pixels in S_p, N_j and N_k are the number of pixels in the *j* and *k* ROI's respectively. N_{S_p} is the total number of ROIs in set S_p, m_{ip}^{j} and m_{ip}^{k} are the magnitude of the *i* ion in pixel *p* of the *j* and *k* ROI, respectively. The list is ordered according to the $C_{ij|S_p}$, assuming that high values mean high contrast and vice-versa.

CHAPTER 4

rMSIannotation: a peak annotation tool for mass spectrometry imaging based on the analysis of isotopic intensity ratios

ABSTRACT: Mass spectrometry imaging (MSI) consist of spatially located spectra with thousands of peaks. Only a fraction of these peaks corresponds to unique monoisotopic peaks, as mass spectra include isotopes, adducts and fragments of compounds. Current peak annotation solutions depend on matching MS features to compounds libraries. We present rMSIannotation, a peak annotation algorithm to annotate carbon isotopes and adducts in metabolomics and lipidomics imaging mass spectrometry datasets without using supporting libraries. rMSIannotation measures and evaluates the intensity ratio between carbon isotopic peaks and models their distribution across the *m/z* axis of the compounds in the Human Metabolome Database. Monoisotopic peak selection is based on the isotopic likelihood score (ILS) made of three components: image morphology correlation, validation of isotopic intensity ratios, and peak centroid mass deviation. rMSIannotation proposes pairs of peaks that can be adducts based on three scores: isotopic pattern coherence, image correlation and mass error. We validated rMSIannotation with three MALDI-MSI datasets which were manually annotated by experts, and compared the annotations obtained with rMSIannotation and with the METASPACE annotation platform. rMSIannotation replicated more than 90% of the manual annotation reported in FT-ICR datasets and expanded the list of annotated compounds with additional monoisotopic peaks and neutral masses. Finally, we evaluated isotopic peak annotation as a data reduction method for MSI by comparing the results of PCA and *k-means* segmentation before and after removing non-monoisotopic peaks. The results show that monoisotopic peaks retain most of the biologic variance in the dataset.

1. Introduction

Mass spectrometry imaging (MSI) is a technique that can spatially resolve the chemical composition of a variety of bio-samples, including animal and plant tissues, to reveal their biological mechanisms.[1–3] An MSI dataset consists of a collection of mass spectra localized in the pixels of an image. Raw mass spectra need to be processed to reduce the variance introduced during acquisition (electronic noise, mass drifts, intensity fluctuations, etc.).[4] The information in a processed dataset consists of spatially resolved discrete *m/z* features, which undergo data analysis steps such as multivariate statistics and compound identification to obtain biological knowledge.[5–7]

Mass spectrometry dataset contains redundant information, since a single chemical compound generates multiple peaks, which can be attributed to isotopes, adducts, fragments, and different ionization states. Therefore, the redundant variables in the dataset tend to enlarge the data size and hinder statistical analysis.[8] Reducing this redundancy to obtain statistically relevant variables is crucial to unveiling biological knowledge.[9–11]

In this study, we define peak annotation as the process of automatically grouping all peaks related to the same molecule, and the ion species to which they correspond.[11–13] This involves labeling carbon monoisotopic (M+0) and isotopic ions (M+1, M+2, etc.), adducts of the same compounds ($[M+Na]^+$, $[M+K]^+$, etc.) and, when possible, assigning putative molecular classes with the Kendrick mass defect.[14,15] Besides, a neutral monoisotopic mass can be determined if two or more adducts can be annotated for a given compound. This allows the assignment of molecular formulas with higher confidence. Peak annotation is an essential step prior to peak identification, which consists in searching the annotated peaks in libraries of chemical compounds to assign them a putative chemical formula and name using MS data and confirming each assignment through MS^n data and orthogonal techniques.[16]

Moreover, peak annotation algorithms are reliable variable selection approaches and greatly facilitate the identification process. Annotation ideally allows unifying all peaks coming from

the same compound, reducing the number of statistical variables to only one independent variable per compound.

Peak annotation algorithms are more established in LC-MS-based experiments than in MSI. Although LC-MS and MSI have different data structure and content, some peak annotation strategies in LC-MS can be adapted to MSI datasets. Notable examples are:

(1) R package CAMERA[11] annotates carbon isotopes, adducts, and fragments in a peak list by first grouping peaks by peak shape correlation, retention time similarity and correlation across samples, and then by checking M+1/M+0 isotopic ratios, and adduct distances. Ratios between M+0 and M+1 isotopes are computationally pre-established.

(2) R package CliqueMS[17] annotates adducts using the similarity between coelution profiles and a similarity network based on the natural frequency of adduct formation observed in real samples.

(3) R package Astream[18] annotates isotopes, fragments, and adducts by using intensity correlations across samples, retention time differences, and expected *m/z* differences.

In MSI there is no chromatographic separation before ionization and ions frequently overlap, even with high resolving power spectrometers (> 20,000). Since MSI is an imaging technique, spatial correlation methods can be used to increase peak annotation confidence. To our knowledge, only two annotation tools have been developed specifically for MSI applications:

(1) R package MassPix[19] annotates M+0, M+1 and M+2 isotopes by searching for intensity ratios between peaks below user defined ratios. After deisotoping, it searches for the *m/z* of M+0 peaks in a self-developed library of lipids to tentatively annotate and identify them. MassPix does not consider spatial information or colocalization among isotopic ion images.

(2) METASPACE annotation platform[20] is an online annotation tool in which users upload their MSI datasets to be annotated. Its annotation workflow consists of generating isotopic patterns from metabolites databases and matching them with the experimental MSI data using three different metrics: spatial chaos measure, spatial isotope measure and spectral isotope measure. Matches with an overall score higher than a threshold are then given a false discovery rate score based on a target-decoy approach.[21] The results of this workflow are pairs of matching adducts and formulae, which lead to tentative *m/z* identifications. On the downside, it is important to notice that METASPACE requires to uploads datasets with a high mass accuracy (<3 ppm) and a resolving power over 70k (*m/z* 200) for reliable results. In addition, METASPACE may be impractical for large experiments since datasets must be uploaded through the internet. Finally, despite having METASPACE 's source code available, it still suffers from the black box effect where users are restricted to visualize the annotation results and are not able to finely control/adapt the annotation tool themselves.

Both MassPix and METASPACE use generated isotopic patterns from libraries of metabolites, which restrict the annotation to compounds already reported in the libraries. To overcome this limitation, we propose rMSIannotation, a new annotation tool based on library-free criteria optimized for compounds below 1200 Da, included in the MSI data processing R package rMSIproc.[22] rMSIannotation takes advantage of the high number of pixels in an MSI dataset to annotate carbon-based isotopes with single and multiple charges using three scores: (1) image morphology, which considers the colocalization among related *m/z* ion images, (2) isotopic pattern profile, which asserts the plausibility of isotopic ratios given an *m/z* ratio and (3) centroid mass deviation, which evaluates the theoretical distance of carbon isotopic patterns. Additionally, monoisotopic ions found by the algorithm are compared with theoretical mass distances of adducts to generate tentative neutral masses. The algorithm has been tested and validated using *in silico* datasets, experimental datasets with manual identifications and by comparing the annotations produced by rMSIannotation with the results provided by METASPACE. Users can freely access and/or contribute to rMSIproc at <https://github.com/prafols/rMSIproc>.

2. Materials and methods

2.1 Imaging datasets

Three published datasets were used to test the algorithm: (1) a MALDI-TOF dataset consisting of bovine ovarian follicles[23], (2) a MALDI-FT-ICR dataset consisting of a bloom-forming alga during infection[24] and (3) a MALDI-FT-ICR dataset consisting of coronal 12 µm-thick brain sections of adult wild-type C57 mice.[20]

2.1.1 MALDI-TOF dataset

The MALDI-TOF dataset consists of a collection of bovine ovarian follicles.[23] The dataset was kindly provided by the authors. Details of sample preparation and data acquisition can be found in the original paper. The authors identified 43 metabolites in the MSI dataset by first, analyzing lipid extracts from the follicular cells with high-resolution LC-MS and direct infusion MS/MS structural analyses and second, searching the identifications in the MSI dataset. The raw data was exported to imzML format using Bruker FlexImaging software and the dataset was then processed using the rMSIproc processing workflow.[22] The processing pipeline consisted of: (1) smoothing by Savitzky-Golay using a kernel size of 7, (2) spectra alignment with two iterations, a 400 ppm max shift, an oversampling of 2 and references for low, mid and high of 0, 0.5 and 0.8, (3) mass calibration using previously identified peaks (*m/z* 524.372, *m/z* 760.586 and *m/z* 824.557) and (4) peak-picking with an SNR threshold set to 5, a peak detector window of 12, a peak oversampling of 10, a binning tolerance of 5 scans and a binning filter of 0.05. The result was a peak matrix with a total of 235 peaks and 15293 pixels within the *m/z* range between 100 and 1200.

2.1.2 MALDI-FT-ICR dataset 1

The MALDI-FT-ICR dataset 1 consists of a bloom-forming alga (Emiliana huxleyi) during infection with a virus.[24] The dataset was available from MetaboLights[25] stored in the study with reference MTBLS769. Details of sample preparation and data acquisition can be found in the original paper, in which the authors identified 37 metabolites using LC-MS and LC-MS/MS in lipidomic experiments performed in liquid cultures. The raw data was exported to imzML format using Bruker FlexImaging and the dataset was then processed using rMSIproc. The processing pipeline consisted of: (1) smoothing by Savitzky-Golay using a kernel size of 7, (2) a spectra alignment with two iterations, a 300 ppms max shift, an oversampling of 2 and references for low, mid and high of 0, 0.5 and 1, (3) mass calibration using four previously identified peaks (*m/z* 689.5024, *m/z* 749.5153, *m/z* 802.5469, *m/z* 826.6199 and *m/z* 902.5782) to facilitate the comparison of the results and (4) peak-picking with an SNR threshold set to 20, a detector window of 10, an oversampling of 10, a binning tolerance of 6 scans and a binning filter of 0.05. The result was a peak matrix with a total of 4047 peaks and 10517 pixels within the *m/z* range of 100 to 1200.

2.1.3 MALDI-FT-ICR dataset 2

The MALDI-FT-ICR dataset 2 consists of four coronal 12 µm-thick brain sections of an adult wild-type C57 mice.[20] The dataset was available from MetaboLights[25] stored in the study with reference number MTBLS313. Details of sample preparation and data acquisition can be found in the original paper. In the original work, the dataset consisted of ten sections of two different animals. In this work, we used four sections out of five from the first animal as the data for one

section was missing. The authors annotated 35 molecules for the first animal using the METASPACE platform and validated 16 representative annotations with LC-MS/MS.

The data was obtained from individual imzML files in processed mode containing the peaks list of each section, which was transformed to rMSIproc's peak matrix format using a mass binning of 10 ppms and a bin filter of 1%. After that, the four peak matrices were combined in a single dataset using rMSIproc's processing pipeline. The resulting peak matrix contained 1011 peaks and 53241 pixels within a mass range from *m/z* 100 to *m/z* 1180.

2.2 Description of the algorithm

The algorithm consists of two modules: isotope annotation and adduct annotation. The isotope annotation module detects pairs of isotope candidates and computes the isotopic detection metrics for all the peaks in the dataset. The adduct annotation module use the information generated by the isotope annotation module and proposes pairs of peaks that could be adducts of the same compound. Lastly, all the annotations generated are organized in three groups: two groups for the adduct module, differentiated by the amount of information gathered during the annotation; and one for the isotope module containing information on the monoisotopic ions. Supplementary Figure S9 shows a flow diagram of the algorithm.

2.2.1 Input data format

Raw spectra undergo rMSIproc's processing workflow,[22] which consists of spectral smoothing, spectral alignment, mass re-calibration, peak picking, and peak binning of all the pixels in the image. The result of this workflow is a peak matrix, in which pixels of the image are arranged in rows, *m/z* features are arranged in columns and the *m/z* axis is shared between all pixels.

The annotation algorithm uses the rMSIproc peak matrix format as input. Alternatively, rMSIproc can create a peak matrix from an imzML file already centroided by third-party software. However, it is recommended to use raw data in profile mode to take full advantage of the complete rMSIproc processing workflow.

2.2.2 Isotope annotation

First, all *m/z* features in the peak matrix are assumed to be M+0 ions and, for all of them, a list of possible M+1 candidates is generated looking for peaks at a mass distance of 1.00336 Da within a user-defined windows (depending on the spectral resolution of the MS analyzer), expressed in number of raw spectra data points. We prefer to specify this mass distance in data points instead of ppm since it provides a more constant metric thought all the mass range. Alternatively, if spectral data is not available in profile mode, the mass tolerance can be specified in ppm. The mass distance is divided by the charge number, if isotopes of ions with multiple charges are being searched for.

Next, the *m/z* features with one M+1 candidate or more are evaluated pairwise with the isotopic likelihood score (ILS) which was developed in-house and consists of the combination of three different scores: 1) the image morphology score, 2) the isotopic pattern profile score and 3) the centroid mass deviation score. Before computation, the pixels with zero value are removed pairwise from both *m/z* features to increase the discriminant power of the score.

1. The image morphology score considers that *m/z* features belonging to the same isotopic pattern are colocalized. We estimate colocalization by least squares regression between the intensities of M+0 and the M+1 candidate across all the pixels using the coefficient of determination (R^2). Ions are colocalized if the coefficient is close to 1.

2. The isotopic pattern profile score examines the relationship between the experimental and the theoretical M+1/M+0 intensity ratios. The experimental intensity ratio is defined as the slope of a linear model produced by least squares regression between the M+0 and the M+1 candidate intensities. The theoretical intensity ratio is calculated inputting the *m/z* of the monoisotopic candidate to a self-developed carbon isotopic ratio model (CIR model). The carbon isotopic ratio model contains the distribution of carbon isotopes intensities ratios across the *m/z* axis up to *m/z* 1200 and delivers the most probable intensity ratio for a given peak mass (see section 1 of supplementary information). Lastly, the experimental and theoretical intensity ratios are subtracted and fitted in a Gaussian score function which preserves the expected variability of the carbon isotopic ratio model. The score gets close to one as the measured intensity ratio of a pair of peaks is more likely to result from an actual isotopic profile.

3. The centroid mass deviation score compares the experimental mass distance between M+0 and its M+1 with the theoretical mass distance between carbon isotopes (considering the charge). The user defines the error tolerance for the mass deviation, which can be introduced in ppms or number of data points. The score gets close to one as the error tolerance reduces.

The three scores are multiplied to calculate the ILS. The pairs of *m/z* features with an ILS greater than the user-defined threshold constitute a monoisotopic/isotopic peak pair. Once all the true M+0 *m/z* features have been found, the full procedure is repeated to evaluate the M+N candidates for all the M+0 *m/z* features until no more candidates are found or N has reached the maximum number of iterations. The number of isotopes (N) to search for is a user-defined parameter.

2.2.3 Adduct annotation

The algorithm searches for pairs of ions (discarding the features annotated as isotopes) whose mass difference fits with a candidate adduct ($[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, user-defined adducts, and neutral losses) within a mass tolerance in ppms to generate putative neutral masses. For each pair of adduct ions, the algorithm calculates three scores to guide the user to select the more probable adduct pairs. The scores are:

1. Isotopic pattern coherence. When two monoisotopic ions are adducts of the same compound, their M+1/M+0 intensity ratio should be the same (unless the ion forming the adduct contains carbon, which would slightly modify the isotopic pattern). This is calculated as the standard error of the mean M+1/M+0 intensity ratios of both monoisotopic ions. Small standard error of the mean indicates good isotopic pattern coherence.

2. Correlation between the two ions intensities using Pearson's R. We assume that adducts of the same compound exhibit some degree of colocalization. It is expected to obtain less degree of colocalization between adducts peaks than between isotopes peaks due to salts concentrations variations related to tissue morphology. Nevertheless, the ion images between adducts of the same compound should be still similar and very rarely show complementary spatial distributions.

3. Mass error between the M+0 peaks and their putative neutral mass. The neutral mass is calculated by subtracting the molecular mass of each adduct ion and averaging the resulting neutral masses. Small mass errors indicate a precise putative neutral mass assignation.

The algorithm allows each *m/z* feature to be part of different adduct pairs (e.g., an $[M+Na]^+$ ion can be paired with an $[M+H]^+$ ion and with an $[M+K]^+$ ion) and even labeled as different adducts in different pairs (e.g., an ion can be labeled as $[M+Na]^+$ in one pair and as $[M+K]^+$ in a different pair). The calculated scores of each annotation are stored along with each adduct pair, which enables the user post-evaluation of all possible adducts pairs to select the most feasible annotations. The user is the responsible to choose/validate the more feasible annotations provided by the algorithm.

Finally, the adduct annotation module generates a list with the neutral masses and its annotation scores, facilitating the search in compound libraries for tentative identification.

2.2.4 Feature annotation groups and output information

The annotations are divided into three groups (A, B and C) depending on the information available to reliably annotate each *m/z* feature.

Group 'A' contains neutral masses from pairs of M+0 ions cataloged as adducts, where at least one isotope is identified for every M+0 ion. The three scores described for adducts can be computed for all these pairs.

Group 'B' contains neutral masses from pairs of ions in which, one ion is an M+0, but not the other. The isotopic information is not available for the second ion since the algorithm failed to assign the corresponding M+1 peak. Therefore, isotopic pattern coherence cannot be computed in this annotation group.

Group 'C' contains the *m/z* ratios of all M+0 annotated ions. Ions only reported in group C are, therefore, annotated as M+0, but their adduct identity is unknown. This group consists of a summary of the isotope annotation module, in which ILS is the key quality parameter.

The output of rMSIannotation consists of the annotations in groups A, B and C (which can be exported as CSV files); the computations of the ILS for all candidates during isotope annotation, and two vectors of the monoisotopic and isotopic ions. The vectors of monoisotopic and isotopic ions can be used to filter the peak matrix to remove the isotopic peaks, or to work with only the monoisotopic ions found.

3. Results

First, we tested the performance of rMSIannotation using two *in silico* MSI datasets. The datasets were developed simulating TOF and FT-ICR mass analyzers experiment in which we know a priori the identity of all the *m/z* features. Section 2 of supplementary information contains the detailed procedure. Then, we used different ILS thresholds with the *in silico* datasets to test the performance of rMSIannotation's criteria and to obtain optimal ILS thresholds. The optimal ratios found were 0.55 to 0.7 for TOF datasets and 0.7 to 0.8 for FT-ICR datasets. Next, we compared the number of coinciding annotations produced by sweeping the ILS threshold in a range of 0.2 to 0.9 for the MALDI-TOF dataset and MALDI-FT-ICR dataset 1. This allowed us to determine whether the optimal ILS thresholds obtained with the *in silico* dataset were applicable to experimental data. The results show that the number of annotations provided by rMSIannotation coinciding with the manual annotations decreases slowly as we increase the ILS threshold until it reaches the optimal thresholds (Figure S5). After this point, the number of coinciding annotations drastically decreases. This suggests that the optimal ILS thresholds obtained *in silico* are applicable experimental data and can be setup as default parameter values. Refer to Section 3 of supplementary information for the complete study.

Second, we annotated using rMSIannotation three experimental datasets acquired with TOF and FT-ICR mass analyzers from papers that reported manually identified compounds. We compared the reported annotations with the ones generated by rMSIannotation. Later, we compared the annotations of rMSIannotation with the results obtained using METASPACE annotation platform on the FT-ICR datasets.

Finally, we evaluated the effects of retaining only M+0 ions during the post-processing of MSI datasets, using principal component analysis (PCA) and *k-means* clustering.

3.1 MALDI-TOF annotation results

The MALDI-TOF dataset consists of a collection of bovine ovarian follicle tissues in which the authors identified 43 metabolites (see Figure 1). After the raw data had been processed, the peak matrix was fed to the annotation algorithm. The parameters used were: isotope search up to M+3, isotope mass tolerance in data points mode and up to 4 data points (~100 ppms at *m/z* 800 for this dataset), ILS threshold set to 0.6 and default $[M+K]^+$, $[M+H]^+$ and $[M+Na]^+$ adducts searched for within a window of 30 ppm.

With these parameters, rMSIannotation generated 16 putative neutral masses in group A and 22 in group B, and found a total of 42 monoisotopic ions in group C. All the annotations of each group are presented in Supplementary Table S1, S2 and S3.

First, we compared the adduct ions found in groups A and B with the adduct ions from the original publication. This was done by searching in groups A and B for the exact masses of the compounds identified. Then, we searched for monoisotopic ions without adduct annotation in group C. Table 1 shows the monoisotopic ions found by rMSIannotation that coincide with those identified in the original work. The ions in group C that also appear in groups A or B (annotated as monoisotopic ions) display its ILS value

We annotated as monoisotopic 23 of the ions in the list of 43 provided by the authors in the original study (see Figure 1). There are three causes explaining why the other 20 ions provided by the authors were not annotated as monoisotopic ions by rMSIannotation: (1) the peak picking algorithm detected only the M+0 ion due to low intensity of the subsequent isotopes; (2) all the ions of the compound have an intensity group below the S/N ratio, and (3) overlapping isotopic patterns of isobaric species which could not be properly resolved by the mass analyzer. Causes 1 and 2 are related to the presence of only one peak per compound in the MSI dataset as the provided identifications were obtained using LC-MS and direct infusion MS/MS. In addition, we further analyzed the case of overlapping with *in silico* overlapping isotopic patterns with different resolving power to determine how it affects rMSIannotation (section 4 of supplementary information). The results show that, rMSIannotation is tolerant to some extend of peak overlapping and the resulting annotation depend on the two overlapped compounds abundance ratios and on the spectral resolving power. As expected, a higher resolving power increases the annotation performance, but even when lowering the resolving power the algorithm still provides reliable results by annotating peak in the isotopic pattern (M+1, M+2…) only when isotopic ratio criteria is met. Therefore, monoisotopic peaks (M+0) highly overlapped with the M+1 peak of another molecule will not be annotated as part of an isotopic pattern of the former molecule. Supplementary Table S4 shows which category applies to the non-annotated ions and Supplementary Figures S10, S11 and S12 show examples of each group defined above, respectively. It is worth mentioning that some of the non-annotated ions could have been annotated by reducing the SNR in the preprocessing steps of the datasets although uninformative noisy peaks may be introduced hampering the subsequent data analysis.

Lastly, we used the Human metabolome database[26] and Lipid maps[27] to putatively identify the ions annotated by rMSIannotation that had not been identified in the original paper. We identified 1 neutral mass with a mass error below 30 ppms that belonged to the CHCA molecule used as matrix (we found 9 common adduct ions by hand in group C), and 4 more monoisotopic masses, resulting in 13 new monoisotopic ions identified. Supplementary Table S5 shows the putative name and molecular formula for the 4 monoisotopic masses in group C (CHCA related annotations are excluded).

3.2 MALDI-FT-ICR annotation results

The MALDI-FT-ICR dataset 1 consists of a bloom-forming alga (*Emiliana huxleyi*) analyzed during a viral infection.[24] The authors of the original paper identified 37 metabolites. The algorithm parameters used were: isotope search up to M+3, isotope mass tolerance in ppm mode up to 10 ppms, ILS threshold set to 0.7 and $[M+K]^+$, $[M+H]^+$ and $[M+Na]^+$ adducts searched up to a maximum of 5 ppm mass tolerance.

With these parameters, rMSIannotation generated 31 putative neutral masses in group A and 95 putative neutral masses in group B, and found a total of 187 monoisotopic ions in group C. All the annotations of each group are presented in Supplementary Table S6, S7 and S8.

Considering all the matching annotations, we found 28 ions on the list of 37 provided by the authors of the original study (Figure 1) and, we obtained 2 additional adducts for two of the compounds in the original work annotation list. Table 2 shows all the coinciding annotations. We observed that in this dataset several M+1 peaks (and subsequent isotopes) have some pixels with null value due to the data reduction mode for FT-ICR raw data which automatically discarded low intensity signals. This produces a bias in the isotopic pattern profile score which can increase or decrease the real ILS score. To solve this problem, the algorithm is designed to discard pairwise pixels with null values to ensure proper linear modelization. Supplementary Figure S13 shows the example of ion *m/z* 826.620 corresponding to compound DGCC 40:7, in which the ILS is 0.877 if null pixels are included and 0.984 if null pixels are discarded.

rMSIannotation was not able to annotate 9 of the manually identified compounds because of their low intensity. This means that the M+1 and subsequent isotopes were not present in the peak matrix or there were too many null pixels to be properly corrected by the algorithm. Supplementary Table S9 shows these compounds and Supplementary Figure S14 shows the case of ions *m/z* 826.640 and *m/z* 812.622.

Various compound libraries were used to tentatively assign the new annotations generated by rMSIannotation not reported in the original work. Supplementary Table S10 shows the putative names and molecular formulae assigned to 19 monoisotopic masses, according to METLIN,[28] Lipid Maps[27] and Dictionary of Natural Products.[29] It is worth mentioning that rMSIannotation helped to find different adducts of common alkenones produced by *Emiliana huxleyi*.[30,31]

Additionally, we submitted the MALDI-FT-ICR dataset 1 to METASPACE to compare its performance against rMSIannotation. Table 3 lists all the manually identified compounds by the authors of the datasets and shows which ions were annotated by rMSIannotation and/or METASPACE. In case of METASPACE, we show the results for FDR 10%, which are showcased as the default results in the online platform, and the results for FDR 20%. The libraries selected in METASPACE were the Human Metabolome Database, Lipid Maps and Chemical Entities of Biological Interest. For FDR 10%, taking as a reference the manually identified compounds, METASPACE found 12 coinciding monoisotopic ions, and for FDR 20% found 20, which is less than the 28 found by rMSIannotation.

To further compare the performance of rMSIannotation with the METASPACE annotation platform, we tested rMSIannotation with the MALDI-FT-ICR dataset 2, consisting of four coronal brain sections of two adult wild-type C57 mice, which were previously annotated by the authors of METASPACE, reporting 31 compounds. The parameters used with rMSIannotation for the MALDI-FT-ICR dataset 2 were: ILS threshold set to 0.7, isotope mass tolerance in ppm mode up to 5 ppms and $[M+K]^+$, $[M+H]^+$ and $[M+Na]^+$ adducts searched up to a maximum of 5 ppm mass tolerance. rMSIannotation was able to putatively identify all the 31 compounds annotated using METASPACE. Moreover, rMSIannotation found 202 monoisotopic ions in group C and a total of 263 neutral masses combining groups A and B. Table 4 show the lists of the 31 annotated ions, together with its ILS values. We obtained ILS values over 0.9 for every annotated compound indicating high confidence in the annotation and confirming the original METASPCAE results.

3.3 Effect of reducing variables to monoisotopic ions in multivariate analysis

In section 3.2 we have shown the ability of rMSIanotation to identify monoisotopic ions and, thereby, to annotate the isotopes which carries redundant information. There are also many peaks that are not annotated that could correspond to overlapped peaks, matrix derived peaks, etc. In this section, we used Principal Component Analysis (PCA) and image segmentation to verify whether discarding the redundant peaks corresponding to the molecules isotopic pattern would represent a significant loss of biological information.

We standardized the data and then we compared the PCA scores of the complete dataset with the PCA scores of a reduced version of the dataset containing only monoisotopic ions. This involves selecting 17.87% (42 out of 235) of the variables for the TOF dataset and 4.62% (187 out of 4047) of the variables for the FT-ICR dataset 1. We compared the images produced by the three principal components on the tissue in each case. We compared the spatial structures displayed in the principal component images of the complete and the reduced datasets by computing its similarity using Pearson's R correlation. For the TOF dataset the correlations were: R = 0.99 (PC1), R = 0.96 (PC2), and R = 0.90 (PC3); for the FT-ICR dataset 1 the correlations were: R = 0.91 (PC1), R = -0.87 (PC2), R = 0.90 (PC3). In both cases, the principal components exhibit a very similar distribution. Figure 2 shows the images of the first three principal components encoded in RGB color space for each studied case. As it can be seen, the tissue morphology is preserved in the reduced dataset.

Next, we analyzed the importance of monoisotopic peaks in the loadings of the first two principal components. Figure 3 shows the loadings of PC1 and PC2 of all *m/z* features on both peak matrices and distinguishes between monoisotopes, isotopes and non-annotated ions. The monoisotopic ions tend to have larger loadings on the PCA, indicating that the variance is mainly led by monoisotopic peaks.

Finally, we also analyzed the extent to which monoisotopic peaks influence a segmentation process. To this end, we applied the k-means algorithm to the datasets with all the peaks, and with only the monoisotopic ions. The number of clusters was selected to suit the morphology of the tissues. Figure 4 shows the results of this procedure. The clusters have the same pixel distribution for both datasets, which indicates that the monoisotopic peaks have a predominant role in establishing the centers of the clustering.

4. Discussion

The annotation of low molecular weight compounds (below 1200 Da) in MSI datasets still has some limitations. As shown in MALDI-TOF annotation results, datasets acquired with TOF mass spectrometers with a resolving power less than 30,000 tend to suffer from overlapping peaks (i.e. isobaric species with very similar mass do not resolve completely). This problem can still arise, although to a much lesser extent, with high resolving power MS analyzers. For instance, if the M+0 peak of a compound A overlaps the M+1 peak of another compound B, the ILS of the M+0 peak of compound B decreases, making it difficult to annotate (supplementary information, section 4). Moreover, when both compounds are co-localized in the same regions of the tissue, the overlapping is harder to detect as peak picking cannot find pixels where both peaks are well resolved. These cases could be addressed with peak deconvolution algorithms, which would split all the isotopic ions from overlapping peaks increasing the scores of peak picking algorithms and generating more annotations. At the same time, peak deconvolution algorithms could benefit from previous peak annotation results by searching for overlapped peaks for which the peak annotation algorithm has previously failed. This would reduce the load of the overall process. As far as we know, no deconvolution

algorithms have been reported with this exclusive purpose in the context of MSI, which could be a line of further work. We presume that overlapping is one of the reasons why METASPACE encourages users to submit ultra-high-resolution datasets.

Adduct annotation is a problem that is harder to address than isotope annotation. First, there are no general rules applicable to the intensity ratios between M+0 adduct ions, since adduct generation depends on experimental conditions.[17] Some compounds tend to ionize better with one specific adduct[32], but this still depends heavily on the sample preparation and the matrix applied. Second, the mass distances between adducts may be like mass distances between different compounds or neutral loses. For example, the mass of the ammonium cation is 18.034 Da, which is very close to the mass of a neutral loss of water (18.011 Da). And third, the colocalization of adducts of the same compound can be affected by the natural abundance of the adduct elements, for instance, some structures in the brain tissue have a high intrinsic concentration of potassium which can affect the distribution of potassium adducts across the tissue and their intensity in comparison to other adducts.[33] Therefore, we rely on correlations between adduct ions to assess to likelihood of a set of peaks to originate from the same chemical compound. These limitations result in adduct annotations being less reliable than isotope annotations. To address this, the rMSIannotation strategy consists in presenting to the user all the possible annotations with its scores to facilitate a manually guided confirmation of the results.

In the presented FT-ICR dataset 1, rMSIannotation found more coinciding annotations with the original paper than METASPACE. For the FT-ICR dataset 2, we were able to replicate the previous METASPACE annotations. These results could be attributed to the differences in the isotope annotation criteria. METASPACE annotations are based on isotopic patterns generated using libraries, reducing the possible annotations to the compounds available in those libraries. This limits the annotation of MSI experiment from understudied organisms like microalgae and precludes compound discovery. On the other hand, rMSIannotation measures and validates isotope peaks intensity using intrinsic chemical information, common for all organic compounds, without relying on compound libraries. Moreover, the output of rMSIannotation can be easily integrated in custom R scripts to filter ions and select the non-redundant features to approach the bio-statistical analysis more reliably.

We also investigated the isotope annotation module as a variable selection method by retaining only monoisotopic peaks. The results show that monoisotopic peaks play a predominant role (i.e. a considerable weight in the loadings) in determining the result of a PCA (Figure 2 and 3), and in establishing the centers of a common clustering procedure like k-means (Figure 4). This is probably because monoisotopic peaks have more intensity than their isotopes (this only applies to molecules with fewer than 93 carbon atoms) and that annotated monoisotopic peaks tend to have larger intensities than non-annotated monoisotopic peaks. This suggests that the PCA analysis is mainly driven by monoisotopic peaks, whereas the rest of MS signals are just introducing redundancy to the data.[32]

5. Conclusion

We presented rMSIannotation, a software tool that annotates carbon isotopes and adducts for MSI dataset in the low mass range. rMSIannotation is useful for putative identification of compounds and variable reduction strategies; and can be integrated in any low-weight compounds MSI data analysis workflows. The results show that rMSIannotation automatically extracts valuable information from both high (TOF) and ultra-high (FT-ICR) resolution spectrometers. The presented algorithm demonstrated a high performance and annotation confidence when compared to the established metabolomics MSI annotation platform: METASPACE and to the manual annotation approaches.

The tool is integrated into the MSI processing R package rMSIproc <https://github.com/prafols/rMSIproc>, which processes and annotates data within the same software environment. This expands the possibilities of MSI data analysis for biological research by reducing data processing and manual inspection time.

6.References

(1) McDonnell, L. A. ;; Heeren, R. M. A. ; Imaging Mass Spectrometry. Mass Spectrometry Reviews 2007, 26, 606–643. https://doi.org/10.1002/mas.

(2) Rohner, T. C.; Staab, D.; Stoeckli, M. MALDI Mass Spectrometric Imaging of Biological Tissue Sections. Mechanisms of Ageing and Development 2005, 126 (1), 177–185. https://doi.org/https://doi.org/10.1016/j.mad.2004.09.032.

(3) Chughtai, K.; Heeren, R. M. A. Mass Spectrometric Imaging for Biomedical Tissue Analysis. Chemical Reviews 2010, 110 (5), 3237–3277. https://doi.org/10.1021/cr100012c.

(4) Norris, J. L.; Cornett, D. S.; Mobley, J. A.; Andersson, M.; Seeley, E. H.; Chaurand, P.; Caprioli, R. M. Processing MALDI Mass Spectra to Improve Mass Spectral Direct Tissue Analysis. International Journal of Mass Spectrometry 2007, 260 (2), 212–221. https://doi.org/https://doi.org/10.1016/j.ijms.2006.10.005.

(5) McDonnell, L. A.; van Remoortere, A.; van Zeijl, R. J. M.; Deelder, A. M. Mass Spectrometry Image Correlation: Quantifying Colocalization. Journal of Proteome Research 2008, 7 (8), 3619–3627. https://doi.org/10.1021/pr800214d.

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

(7) Alexandrov, T. MALDI Imaging Mass Spectrometry: Statistical Data Analysis and Current Computational Challenges. BMC Bioinformatics 2012, 13 (16), S11. https://doi.org/10.1186/1471-2105-13-S16-S11.

(8) del Castillo, E.; Sementé, L.; Torres, S.; Ràfols, P.; Ramírez, N.; Martins-Green, M.; Santafe, M.; Correig, X. RMsikeyion: An Ion Filtering r Package for Untargeted Analysis of Metabolomic LDI-MS Images. Metabolites 2019, 9 (8). https://doi.org/10.3390/metabo9080162.

(9) Thomas, S. A.; Race, A. M.; Steven, R. T.; Gilmore, I. S.; Bunch, J. Dimensionality Reduction of Mass Spectrometry Imaging Data Using Autoencoders. In 2016 IEEE Symposium Series on Computational Intelligence (SSCI); 2016; pp 1–7.

(10) McDonnell, L. A.; van Remoortere, A.; de Velde, N.; van Zeijl, R. J. M.; Deeldera, A. M. Imaging Mass Spectrometry Data Reduction: Automated Feature Identification and Extraction. Journal of the American Society for Mass Spectrometry 2010, 21 (12), 1969–1978. https://doi.org/10.1021/jasms.8b03661.

(11) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. Analytical Chemistry 2012, 84 (1), 283–289. https://doi.org/10.1021/ac202450g.

(12) Wang, L.; Xing, X.; Chen, L.; Yang, L.; Su, X.; Rabitz, H.; Lu, W.; Rabinowitz, J. D. Peak Annotation and Verification Engine for Untargeted LC–MS Metabolomics. Analytical Chemistry 2019, 91 (3), 1838–1846. https://doi.org/10.1021/acs.analchem.8b03132.

(13) Domingo-Almenara, X.; Montenegro-Burke, J. R.; Benton, H. P.; Siuzdak, G. Annotation: A Computational Solution for Streamlining Metabolomics Analysis. Analytical Chemistry 2018, 90 (1), 480–489. https://doi.org/10.1021/acs.analchem.7b03929.

(14) Hughey, C. A.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G.; Qian, K. Kendrick Mass Defect Spectrum: A Compact Visual Analysis for Ultrahigh-Resolution Broadband Mass Spectra. Analytical Chemistry 2001, 73 (19), 4676–4681. https://doi.org/10.1021/ac010560w.

(15) Lerno, L. A.; German, J. B.; Lebrilla, C. B. Method for the Identification of Lipid Classes Based on Referenced Kendrick Mass Analysis. Analytical Chemistry 2010, 82 (10), 4236–4245. https://doi.org/10.1021/ac100556g.

(16) Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W.-M.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; Hankemeier, T.; Hardy, N.; Harnly, J.; Higashi, R.; Kopka, J.; Lane, A. N.; Lindon, J. C.; Marriott, P.; Nicholls, A. W.; Reily, M. D.; Thaden, J. J.; Viant, M. R. Proposed Minimum Reporting Standards for Chemical Analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics : Official journal of the Metabolomic Society 2007, 3 (3), 211–221. https://doi.org/10.1007/s11306-007-0082-2.

(17) Senan, O.; Aguilar-Mogas, A.; Navarro, M.; Capellades, J.; Noon, L.; Burks, D.; Yanes, O.; Guimerà, R.; Sales-Pardo, M. CliqueMS: A Computational Tool for Annotating in-Source Metabolite Ions from LC-MS Untargeted Metabolomics Data Based on a Coelution Similarity Network. Bioinformatics 2019, 35 (20), 4089–4097. https://doi.org/10.1093/bioinformatics/btz207.

(18) Alonso, A.; Julià, A.; Beltran, A.; Vinaixa, M.; Díaz, M.; Ibañez, L.; Correig, X.; Marsal, S. AStream: An R Package for Annotating LC/MS Metabolomic Data. Bioinformatics (Oxford, England) 2011, 27 (9), 1339–1340. https://doi.org/10.1093/bioinformatics/btr138.

(19) Bond, N. J.; Koulman, A.; Griffin, J. L.; Hall, Z. MassPix: An R Package for Annotation and Interpretation of Mass Spectrometry Imaging Data for Lipidomics. Metabolomics 2017. https://doi.org/10.1007/s11306-017-1252-5.

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

(21) Reiter, L.; Claassen, M.; Schrimpf, S. P.; Jovanovic, M.; Schmidt, A.; Buhmann, J. M.; Hengartner, M. O.; Aebersold, R. Protein Identification False Discovery Rates for Very Large Proteomics Data Sets Generated by Tandem Mass Spectrometry. Molecular & Cellular Proteomics 2009, 8 (11), 2405 LP – 2417. https://doi.org/10.1074/mcp.M900317-MCP200.

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

(23) Bertevello, P. S.; Teixeira-Gomes, A. P.; Seyer, A.; Carvalho, A. V.; Labas, V.; Blache, M. C.; Banliat, C.; Cordeiro, L. A. V.; Duranthon, V.; Papillier, P.; Maillard, V.; Elis, S.; Uzbekova, S. Lipid Identification and Transcriptional Analysis of Controlling Enzymes in Bovine Ovarian Follicle. International Journal of Molecular Sciences 2018, 19 (10). https://doi.org/10.3390/ijms19103261.

(24) Schleyer, G.; Shahaf, N.; Ziv, C.; Dong, Y.; Meoded, R. A.; Helfrich, E. J. N.; Schatz, D.; Rosenwasser, S.; Rogachev, I.; Aharoni, A.; Piel, J.; Vardi, A. In Plaque-Mass Spectrometry Imaging of a Bloom-Forming Alga during Viral Infection Reveals a Metabolic Shift towards Odd-Chain Fatty Acid Lipids. Nature Microbiology 2019, 4 (3), 527–538. https://doi.org/10.1038/s41564-018-0336-y.

(25) Haug, K.; Cochrane, K.; Nainala, V. C.; Williams, M.; Chang, J.; Jayaseelan, K. V.; O'Donovan, C. MetaboLights: A Resource Evolving in Response to the Needs of Its

Scientific Community. Nucleic Acids Research 2019, 48 (D1), D440–D444. https://doi.org/10.1093/nar/gkz1019.

(26) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; Sayeeda, Z.; Lo, E.; Assempour, N.; Berjanskii, M.; Singhal, S.; Arndt, D.; Liang, Y.; Badran, H.; Grant, J.; Serra-Cayuela, A.; Liu, Y.; Mandal, R.; Neveu, V.; Pon, A.; Knox, C.; Wilson, M.; Manach, C.; Scalbert, A. HMDB 4.0: The Human Metabolome Database for 2018. Nucleic Acids Research 2017, 46 (D1), D608–D617. https://doi.org/10.1093/nar/gkx1089.

(27) Fahy, E.; Sud, M.; Cotter, D.; Subramaniam, S. LIPID MAPS Online Tools for Lipid Research. Nucleic Acids Research 2007, 35 (suppl_2), W606–W612. https://doi.org/10.1093/nar/gkm324.

(28) Guijas, C.; Montenegro-Burke, J. R.; Domingo-Almenara, X.; Palermo, A.; Warth, B.; Hermann, G.; Koellensperger, G.; Huan, T.; Uritboonthai, W.; Aisporna, A. E.; Wolan, D. W.; Spilker, M. E.; Benton, H. P.; Siuzdak, G. METLIN: A Technology Platform for Identifying Knowns and Unknowns. Analytical chemistry 2018, 90 (5), 3156–3164. https://doi.org/10.1021/acs.analchem.7b04424.

(29) CRC Press, Taylor & Francis Group, an I. G. company 2020 (P2). Dictionary of Natural Products 29.1 Chemical Search http://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml (accessed Jul 22, 2020).

(30) Fulton, J. M.; Kendrick, B. J.; DiTullio, G. R.; van Mooy, B. A. S. Alkenone Unsaturation during Virus Infection of Emiliania Huxleyi. Organic Geochemistry 2017, 111, 82–85. https://doi.org/https://doi.org/10.1016/j.orggeochem.2017.06.001.

(31) Llewellyn, C. A.; Evans, C.; Airs, R. L.; Cook, I.; Bale, N.; Wilson, W. H. The Response of Carotenoids and Chlorophylls during Virus Infection of Emiliania Huxleyi (Prymnesiophyceae). Journal of Experimental Marine Biology and Ecology 2007, 344 (1), 101–112. https://doi.org/https://doi.org/10.1016/j.jembe.2006.12.013.

(32) Garate, J.; Lage, S.; Martín-Saiz, L.; Perez-Valle, A.; Ochoa, B.; Boyano, M. D.; Fernández, R.; Fernández, J. A. Influence of Lipid Fragmentation in the Data Analysis of Imaging Mass Spectrometry Experiments. Journal of the American Society for Mass Spectrometry 2020, 31 (3), 517–526. https://doi.org/10.1021/jasms.9b00090.

(33) Hankin, J. A.; Farias, S. E.; Barkley, R. M.; Heidenreich, K.; Frey, L. C.; Hamazaki, K.; Kim, H.-Y.; Murphy, R. C. MALDI Mass Spectrometric Imaging of Lipids in Rat Brain Injury Models. Journal of the American Society for Mass Spectrometry 2011, 22 (6). https://doi.org/10.1021/jasms.8b04045.

7. Tables and Figures

Table 1. Coinciding annotations of MALDI-TOF dataset between rMSIannotation and author's manual identifications.

Name	Formula	Adduct	*m/z*	Mass error (ppm)	Annotation group	ILS
Phosphocholine	$C_5H_{15}NO_4P$	$[M]^+$	184.141	364.560	C	0.842
LPC 16a:0	$C_{24}H_{50}NO_7P$	$[M+H]^+$	496.367	54.866	A	0.848
LPC 16a:0	$C_{24}H_{50}NO_7P$	$[M+Na]^+$	518.348	50.717	B	---
LPC 18a:1	$C_{26}H_{52}NO_7P$	$[M+H]^+$	522.387	60.460	B	0.812
LPC 18a:0	$C_{26}H_{54}NO_7P$	$[M+H]^+$	524.372	1.781	C	0.746
LPC 16a:0	$C_{24}H_{50}NO_7P$	$[M+K]^+$	534.318	41.833	A	0.662
LPC 18a:1	$C_{26}H_{52}NO_7P$	$[M+Na]^+$	544.363	47.099	B	---
LPC 18a:1	$C_{26}H_{52}NO_7P$	$[M+K]^+$	560.338	47.653	B	---
SM(d18:1/C16:0)	$C_{39}H_{79}N_2O_6P$	$[M+H]^+$	703.576	1.632	B	---
SM(d18:1/C16:0)	$C_{39}H_{79}N_2O_6P$	$[M+Na]^+$	725.543	19.014	A	0.853
PC 32a:0	$C_{40}H_{80}NO_8P$	$[M+H]^+$	734.562	10.116	A	0.866
SM(d18:1/C16:0)	$C_{39}H_{79}N_2O_6P$	$[M+K]^+$	741.510	27.960	A	0.916
PC 32a:0	$C_{40}H_{80}NO_8P$	$[M+Na]^+$	756.526	33.542	A	0.672
PC 34a:2	$C_{42}H_{80}NO_8P$	$[M+H]^+$	758.548	28.253	B	0.705
PC 34a:1	$C_{42}H_{82}NO_8P$	$[M+H]^+$	760.574	14.569	A	0.829
PC 32a:0	$C_{40}H_{80}NO_8P$	$[M+K]^+$	772.508	22.411	A	0.641
PC 34a:2	$C_{42}H_{80}NO_8P$	$[M+Na]^+$	780.537	18.418	B	---
PC 34a:1	$C_{42}H_{82}NO_8P$	$[M+Na]^+$	782.539	35.814	A	0.877
PC 36a:2	$C_{44}H_{84}NO_8P$	$[M+H]^+$	786.585	19.999	A	0.692
PC 34a:2	$C_{42}H_{80}NO_8P$	$[M+K]^+$	796.522	4.159	B	---
PC 34a:1	$C_{42}H_{82}NO_8P$	$[M+K]^+$	798.517	30.009	A	0.866
PC 36a:2	$C_{44}H_{84}NO_8P$	$[M+Na]^+$	808.555	34.229	B	---
PC 36a:2	$C_{44}H_{84}NO_8P$	$[M+K]^+$	824.530	32.277	A	0.642

*A missing ILS value correspond to ions where the isotopic pattern could not be annotated and are exclusively in group B.

Table 2. Coinciding annotations of MALDI-FT-ICR dataset 1 between rMSIannotation and author's manual identifications.

Name	Formula	Adduct	*m/z*	Mass error (ppm)	Annotation group	ILS
Sulfonioglycerolipid 28:0	$C_{38}H_{72}O_8S$	$[M+H]^+$	689.502	0.710	C	0.959
Sulfonioglycerolipid 30:0	$C_{40}H_{76}O_8S$	$[M+H]^+$	717.534	0.382	C	0.935
DGCC 36:6	$C_{46}H_{77}NO_8$	$[M+H]^+$	772.573	0.653	C	0.996
PC 36:6	$C_{44}H_{76}NO_8P$	$[M+H]^+$	778.538	1.732	C	0.955
DGCC 37:6	$C_{47}H_{79}NO_8$	$[M+H]^+$	786.588	1.376	C	0.992
Sulfonioglycerolipid 36:6	$C_{46}H_{76}O_8S$	$[M+H]^+$	789.533	1.125	C	0.984
PDPT 36:6	$C_{44}H_{75}O_8PS$	$[M+H]^+$	795.500	1.167	C	0.972
TG 46:1	$C_{49}H_{92}O_6$	$[M+Na]^+$	799.679	1.309	C	0.943
PDPT 37:6	$C_{45}H_{77}O_8PS$	$[M+H]^+$	809.516	1.472	C	0.836
Sulfonioglycerolipid 38:6	$C_{48}H_{80}O_8S$	$[M+H]^+$	817.565	0.784	C	0.975
PDPT 38:6	$C_{46}H_{79}O_8PS$	$[M+H]^+$	823.530	1.524	C	0.941
DGCC 40:7	$C_{50}H_{83}NO_8$	$[M+H]^+$	826.620	0.632	C	0.984
TG 48:1	$C_{51}H_{96}O_6$	$[M+Na]^+$	827.710	1.148	C	0.919
PC 40:7	$C_{48}H_{82}NO_8P$	$[M+H]^+$	832.585	1.183	C	0.767
Sulfonioglycerolipid 40:7	$C_{50}H_{82}O_8S$	$[M+H]^+$	843.579	3.861	C	0.879
TG 50:6	$C_{53}H_{90}O_6$	$[M+Na]^+$	845.664	0.376	C	0.735
PDPT 40:7	$C_{48}H_{81}O_8PS$	$[M+H]^+$	849.547	0.663	C	0.971
TG 50:2	$C_{53}H_{98}O_6$	$[M+Na]^+$	853.726	1.619	C	0.805
PC 44:12	$C_{52}H_{80}NO_8P$	$[M+H]^+$	878.570	0.524	C	0.968
PDPT 42:9	$C_{50}H_{81}O_8PS$	$[M+H]^+$	895.530	1.899	C	0.856
TG 54:7	$C_{57}H_{96}O_6$	$[M+Na]^+$	899.710	1.531	B	0.966
BLL 44:12	$C_{54}H_{79}NO_{10}$	$[M+H]^+$	902.578	0.264	C	0.796
TG 56:7	$C_{59}H_{100}O_6$	$[M+H]^+$	905.759	0.295	B	---
TG 54:7	$C_{57}H_{96}O_6$	$[M+K]^+$	915.685	1.256	B	---
TG 56:7	$C_{59}H_{100}O_6$	$[M+Na]^+$	927.742	1.311	B	0.899
TG 58:16	$C_{61}H_{86}O_6$	$[M+Na]^+$	937.634	0.080	C	0.910
TG 58:12	$C_{61}H_{94}O_6$	$[M+Na]^+$	945.695	1.332	C	0.883
TG 58:11	$C_{61}H_{96}O_6$	$[M+Na]^+$	947.711	0.458	C	0.907
TG 58:10	$C_{61}H_{98}O_6$	$[M+Na]^+$	949.727	1.332	C	0.914
TG 58:9	$C_{61}H_{100}O_6$	$[M+Na]^+$	951.743	1.310	C	0.888

*A missing ILS value correspond to ions where the isotopic pattern could not be annotated and are exclusively in group B.

Table 3. Coinciding annotations of MALDI-FT-ICR dataset 1 between rMSIannotation and METASPACE.

Formula	Adduct	*m/z*	METASPACE (FDR 10%)	METASPACE (FDR 20%)	rMSIannotation
$C_{38}H_{72}O_8S$	$[M+H]^+$	689.502	-	-	x
$C_{40}H_{76}O_8S$	$[M+H]^+$	717.534	-	-	x
$C_{40}H_{77}O_8PS$	$[M+H]^+$	749.515	-	x	-
$C_{46}H_{77}NO_8$	$[M+H]^+$	772.573	-	-	x
$C_{44}H_{76}NO_8P$	$[M+H]^+$	778.538	x	x	x
$C_{47}H_{79}NO_8$	$[M+H]^+$	786.588	-	-	x
$C_{46}H_{76}O_8S$	$[M+H]^+$	789.533	-	x	x
$C_{44}H_{75}O_8PS$	$[M+H]^+$	795.501	x	x	x
$C_{49}H_{92}O_6$	$[M+Na]^+$	799.679	x	x	x
$C_{46}H_{75}NO_{10}$	$[M+H]^+$	802.547	-	-	-
$C_{45}H_{77}O_8PS$	$[M+H]^+$	809.516	x	x	x
$C_{44}H_{87}NO_{10}$	$[M+Na]^+$	812.622	-	-	-
$C_{50}H_{94}O_6$	$[M+Na]^+$	813.695	x	x	-
$C_{48}H_{80}O_8S$	$[M+H]^+$	817.565	-	-	x
$C_{46}H_{79}O_8PS$	$[M+H]^+$	823.533	x	x	x
$C_{45}H_{87}NO_{10}$	$[M+Na]^+$	824.622	-	-	-
$C_{50}H_{83}NO_8$	$[M+H]^+$	826.620	-	-	x
$C_{45}H_{89}NO_{10}$	$[M+Na]^+$	826.640	-	-	-
$C_{51}H_{96}O_6$	$[M+Na]^+$	827.712	x	x	x
$C_{48}H_{82}NO_8P$	$[M+H]^+$	832.585	-	-	x
$C_{50}H_{82}O_8S$	$[M+H]^+$	843.579	-	-	x
$C_{53}H_{90}O_6$	$[M+Na]^+$	845.664	-	x	x
$C_{48}H_{81}O_8PS$	$[M+H]^+$	849.547	x	x	x
$C_{53}H_{98}O_6$	$[M+Na]^+$	853.726	-	x	x
$C_{53}H_{100}O_6$	$[M+Na]^+$	855.746	-	-	-
$C_{52}H_{80}NO_8P$	$[M+H]^+$	878.569	x	x	x
$C_{49}H_{91}NO_{11}$	$[M+Na]^+$	892.649	-	-	-
$C_{50}H_{81}O_8PS$	$[M+H]^+$	895.531	-	x	x
$C_{57}H_{96}O_6$	$[M+Na]^+$	899.712	x	x	x
$C_{54}H_{79}NO_{10}$	$[M+H]^+$	902.578	-	-	x
$C_{59}H_{100}O_6$	$[M+H]^+$	905.759	-	x	x
$C_{57}H_{108}O_6$	$[M+Na]^+$	911.804	-	-	-
$C_{57}H_{96}O_6$	$[M+K]^+$	915.685	-	-	x
$C_{59}H_{100}O_6$	$[M+Na]^+$	927.742	x	x	x
$C_{61}H_{86}O_6$	$[M+Na]^+$	937.634	-	-	x
$C_{61}H_{94}O_6$	$[M+Na]^+$	945.695	-	x	x
$C_{61}H_{96}O_6$	$[M+Na]^+$	947.711	-	x	x
$C_{61}H_{98}O_6$	$[M+Na]^+$	949.727	x	x	x
$C_{61}H_{100}O_6$	$[M+Na]^+$	951.743	-	x	x

Table 4. Coinciding annotations of MALDI-FT-ICR dataset 2 between rMSIannotation and METASPACE.

Formula	Adduct	*m/z*	rMSIannotation	ILS
$C_{35}H_{66}O_4$	$[M+H]^+$	551.503	x	0.977
$C_{37}H68O_4$	$[M+H]^+$	577.519	x	0.975
$C_{28}H_{33}O_{14}$	$[M+Na]^+$	616.176	x	0.988
$C_{37}H_{71}O_8P$	$[M+Na]^+$	697.478	x	0.938
$C_{37}H_{71}O_8P$	$[M+K]^+$	713.452	x	0.965
$C_{39}H_{73}O_8P$	$[M+Na]^+$	723.494	x	0.957
$C_{41}H_{83}N_2O_6P$	$[M+H]^+$	731.606	x	0.909
$C_{40}H_{80}NO_8P$	$[M+H]^+$	734.569	x	0.988
$C_{39}H_{73}O_8P$	$[M+K]^+$	739.468	x	0.951
$C_{39}H_{79}N_2O_6P$	$[M+K]^+$	741.531	x	0.963
$C_{41}H_{82}NO_8P$	$[M+H]^+$	748.585	x	0.979
$C_{41}H_{83}N_2O_6P$	$[M+Na]^+$	753.588	x	0.976
$C_{40}H_{80}NO_8P$	$[M+Na]^+$	756.551	x	0.977
$C_{42}H_{84}NO_8P$	$[M+H]^+$	762.601	X	0.962
$C_{41}H_{83}N_2O_6P$	$[M+K]^+$	769.562	x	0.982
$C_{43}H_{74}NO_7P$	$[M+Na]^+$	770.510	x	0.912
$C_{40}H_{78}NO_8P$	$[M+K]^+$	770.510	x (isobaric)	0.912
$C_{43}H_{76}NO_7P$	$[M+Na]^+$	772.525	x	0.967
$C_{40}H_{80}NO_8P$	$[M+K]^+$	772.525	x (isobaric)	0.967
$C_{42}H_{84}NO_8P$	$[M+Na]^+$	784.583	x	0.924
$C_{45}H_{76}NO_7P$	$[M+Na]^+$	796.525	x	0.896
$C_{42}H_{80}NO_8P$	$[M+K]^+$	796.525	x (isobaric)	0.896
$C_{45}H_{80}NO_7P$	$[M+Na]^+$	800.557	x	0.797
$C_{42}H_{84}NO_8P$	$[M+K]^+$	800.557	x (isobaric)	0.797
$C_{43}H_{78}NO_8P$	$[M+K]^+$	806.510	x	0.921
$C_{44}H_{80}NO_8P$	$[M+K]^+$	820.525	x	0.968
$C_{44}H_{84}NO_8P$	$[M+K]^+$	824.557	x	0.950
$C_{44}H_{86}NO_8P$	$[M+K]^+$	826.572	x	0.970
$C_{45}H_{78}NO_8P$	$[M+K]^+$	830.510	x	0.942
$C_{46}H_{84}NO_8P$	$[M+Na]^+$	832.583	x	0.937
$C_{46}H_{80}NO_8P$	$[M+K]^+$	844.525	x	0.968
$C_{46}H_{82}NO_8P$	$[M+K]^+$	846.541	x	0.919
$C_{46}H_{84}NO_8P$	$[M+K]^+$	848.557	x	0.958
$C_{48}H_{91}NO_8$	$[M+K]^+$	848.638	x	0.959
$C_{48}H_{84}NO_8P$	$[M+K]^+$	872.557	x	0.933

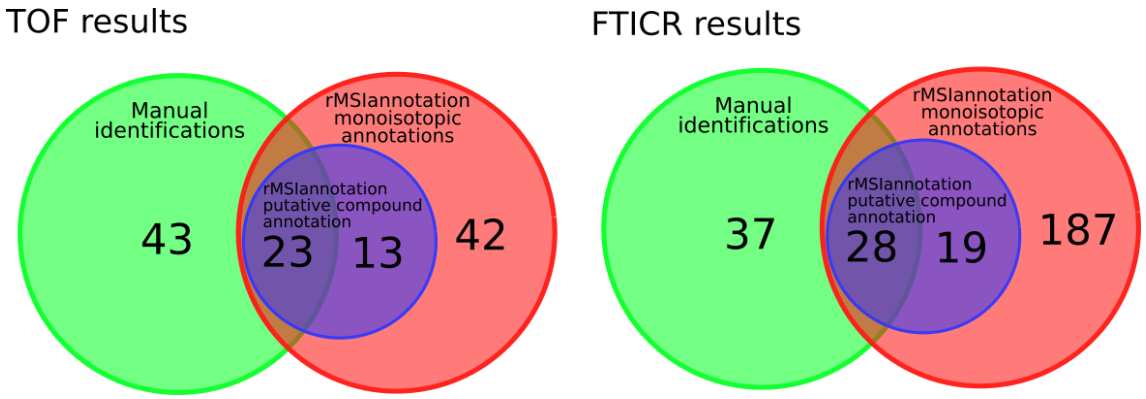


Figure 1. Diagrams representing the number of identifications reported by the authors of the datasets, the number of M+0 annotations produced by rMSIannotation in group C and the number of coinciding and new putative compound annotations found using rMSIannotation.

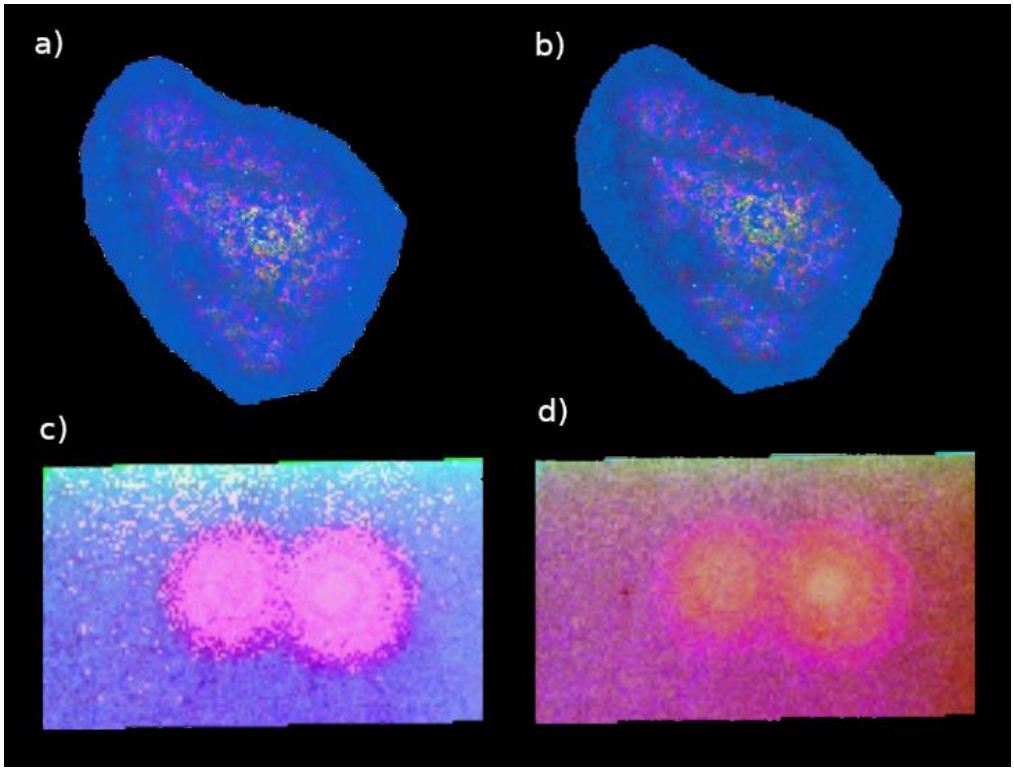


Figure 2. Representation of the first three principal components on the tissue in RGB. Red channel for PC1, green channel for PC2 and blue channel for PC3. a) TOF dataset with all the peaks. b) TOF dataset with only annotated monoisotopic peaks. c) FT-ICR dataset with all the peaks. d) FT-ICR with only annotated monoisotopic peaks.

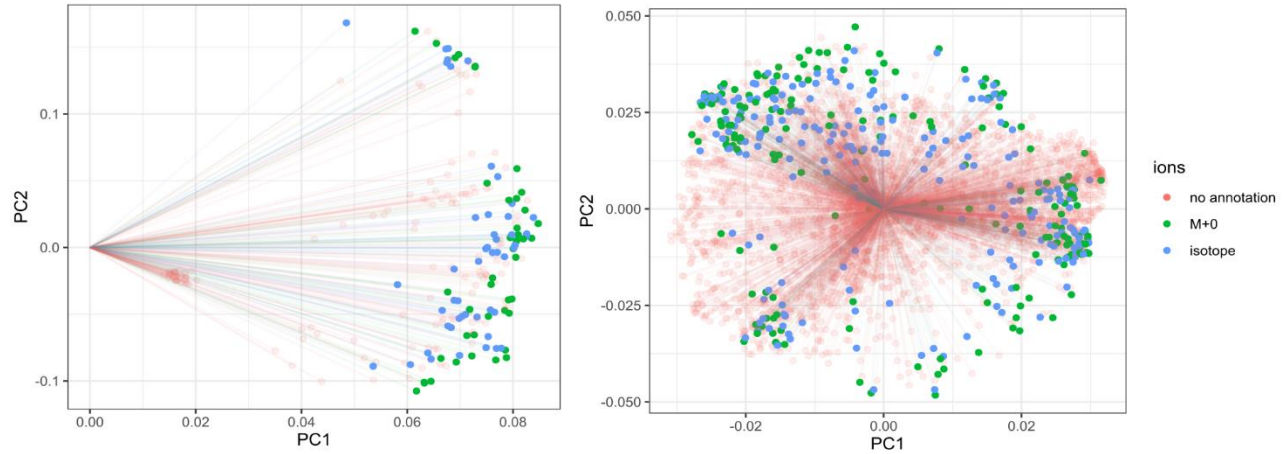


Figure 3. a) Loadings of PC1 and PC2 of the TOF dataset b) Loadings of PC1 and PC2 of the FT-ICR datasets. Every point in the graphs represents an m/z feature in the datasets. Green points represent the peaks annotated as monoisotopic, blue points are peaks annotated as isotopes (M+1, M+2, etc.) and red points are peaks that have not been annotated by rMSIannotation.

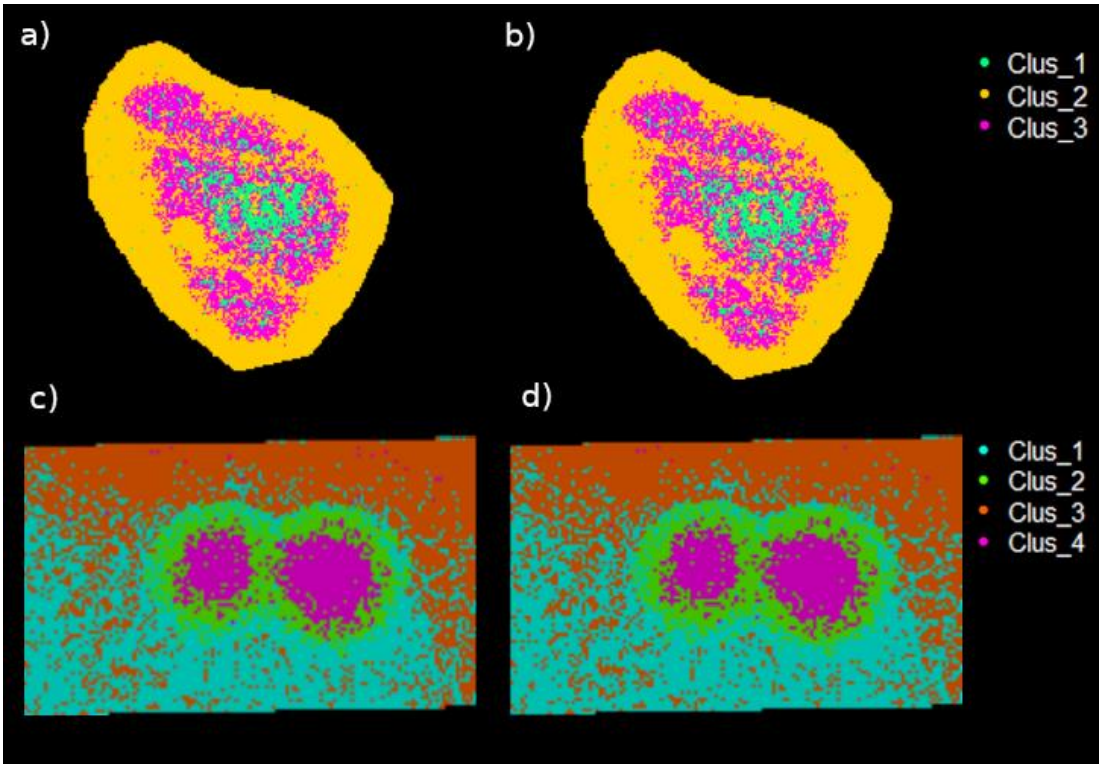


Figure 4. a) k-means clustering of the TOF dataset with all peaks with k = 3. b) k-means clustering of the TOF dataset with only the monoisotopic ions with k = 3. c) k-means clustering of the FT-ICR dataset with all peaks with k = 4. d) k-means clustering of the FT-ICR dataset with only the monoisotopic ions with k = 4.

7. Supporting information

7.1. Carbon isotopic ratio (CIR) model

The natural abundance of carbon isotopes is 98.98% for ^{12}C and 1.11% for ^{13}C . This produces unique isotopic patterns related with the number of carbon atoms and, hence, the molecular weight. In order to calculate the isotopic pattern score (see Section 3.2), a reference for the M+1/M+0 intensity ratio needs to be generated to evaluate the experimental ratio between isotopic and monoisotopic candidate peaks. To do so, a carbon isotopic ratio (CIR) model is derived, which explains the overall tendency in intensity ratios between isotopes and their monoisotopic peaks in terms of their molecular weight (in Da). If the ions detected are single charged, the CIR model produces a theoretical reference of M+0/M+1 ratio for a given molecular weight.

The CIR model was derived using all the molecules in the Human Metabolome Database (HMDB) with a molecular weight below 1200 Da. Figure S1 shows the M+1/M+0 isotope intensity ratios versus the monoisotopic neutral mass of all molecules in the HMDB, where each point represents a single molecular formula. There is a clear positive linear relationship between both variables, as carbon atoms are the main building block in organic compounds.

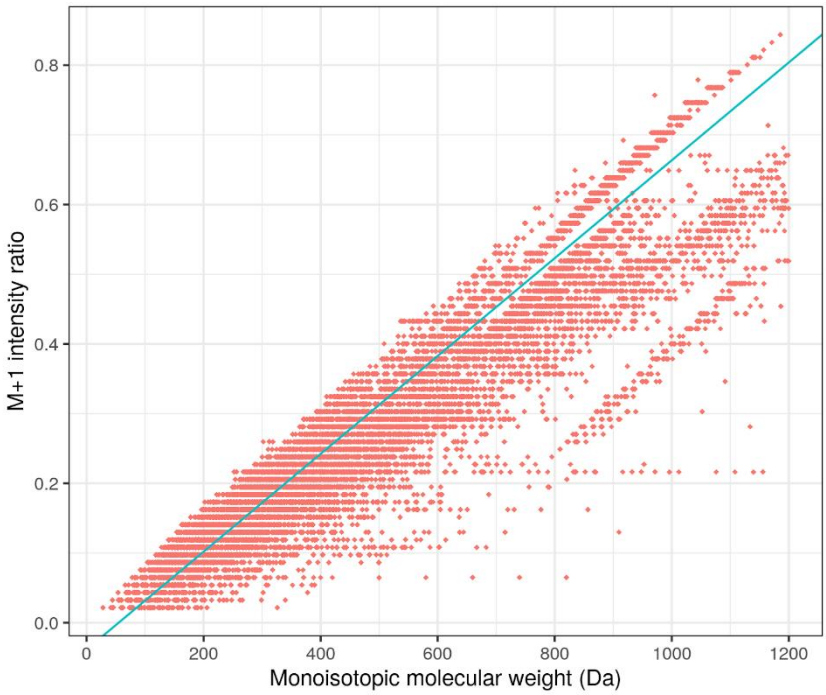


Figure S1. Scatter plot representing the monoisotopic molecular weight of metabolites included in HMDB (below m/z 1200) against the M+1/M+0 intensity ratio. Each point in the image represents a unique molecular formula. The blue line indicates the best fit of the model.

Using this information, a linear model was inferred to calculate an expected M+N/M+0 intensity ratio given an *m/z* of a mono-charged ion. For N = 1, the relationship between M+1/M+0 intensity ratio and molecular weight is linear. The model follows the equation:

$$\frac{M+1\ intensity}{M+0\ intensity} = 7.02 \cdot 10^{-4} \cdot m/z - 0.03851$$

7.2. *In silico* dataset

An *in silico* dataset was developed with the aim to provide a ground truth for testing the algorithm developed. The dataset was designed to display similar characteristics of a real MSI dataset:

1) several m/z signals, some of which are correlated as they simulate to be produced by the same compounds; 2) a defined morphology caused by signals at different intensities; 3) spatial noise; and 4) variations in the m/z axis.1) The theoretical isotopic pattern of 7 organic molecules was generated using the enviPat software[1]. The chemical formulas of the seven molecules were: $C_{38}H_{72}O_8S$, $C_{40}H_{77}O_8PS$, $C_{45}H_{120}$, $C_{49}H_{92}O_6$, $C_{50}H_{83}NO_8$, $C_{52}H_{80}NO_8P$ and $C_{61}H_{96}O_6$. For each molecule, four carbon isotopes (M+0 to M+3) of a single adduct ($[M+H]^+$, $[M+Na]^+$ or $[M+K]^{+)}$ were generated. To challenge the algorithm, a false peak was added 1 Da before all M+0 peaks. This allows the algorithm to compute a score for all the real peaks of the patterns. Also, two out of seven molecules were given extra adducts, one consisting of $[M+Na]^+$ and $[M+K]^+$, and the other $[M+H]^+$, $[M+Na]^+$ and $[M+K]^+$. So there was a total of 49 peaks.

2) A brain image of 258717 pixels was constructed by taking the morphology of a mouse brain image from the Brain Atlas[2] as a reference. The brain was segmented into 11 regions. All the peaks from the 7 molecules were placed in the regions at different concentrations to preserve the theoretical isotopic patterns and produce clear differences between regions. Figure

S2 shows the mean spectrum of the 11 regions in the brain along with the molecular formula of the simulated compounds.

3) Seven (one per molecule) simplex noise patterns (using the R package ambient[3]) were added to all the peaks from the same molecule to generate spatial structures similar to biological tissue morphologies. Also, two different levels of Gaussian noise, with standard deviations of 0.05 and 0.01, were applied to each individual peak to create two slightly different matrices: one simulating experiments using a TOF mass analyzer (noisier) and the other a FT-ICR (less noisy). The Gaussian noise applied produces differences in the isotopic patterns between pixels and results in more heterogeneous regions.

4) Different levels of mass shift were added to the mass axis of both peak matrices using normal noise to simulate differences in mass accuracy between mass analyzers. For the TOF dataset the mass shift had a standard deviation of 20 ppms while for the FT-ICR the standard deviation was 5 ppms.

The resulting peak matrices of this process had 49 *m/z* features as columns, and 258717 pixels as rows. Although the number of m/z features is smaller than in a real dataset, the purpose of this matrix was not to accurately reproduce a real dataset but to provide a ground truth with which to test the criteria developed. Figure S3 illustrates several steps of the process.

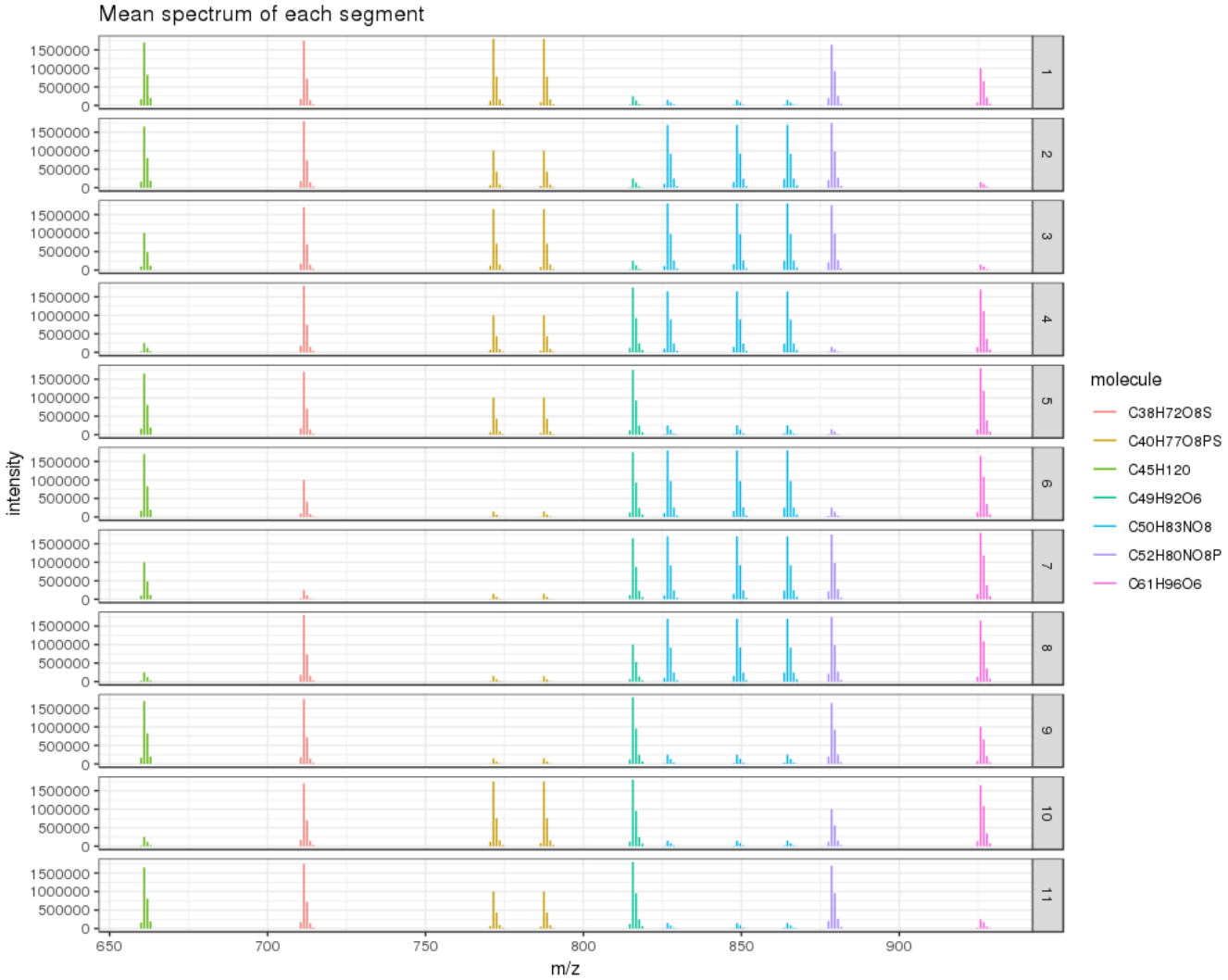


Figure S2. Representation of the *in silico* isotopic patterns of the seven molecules for each of the eleven regions

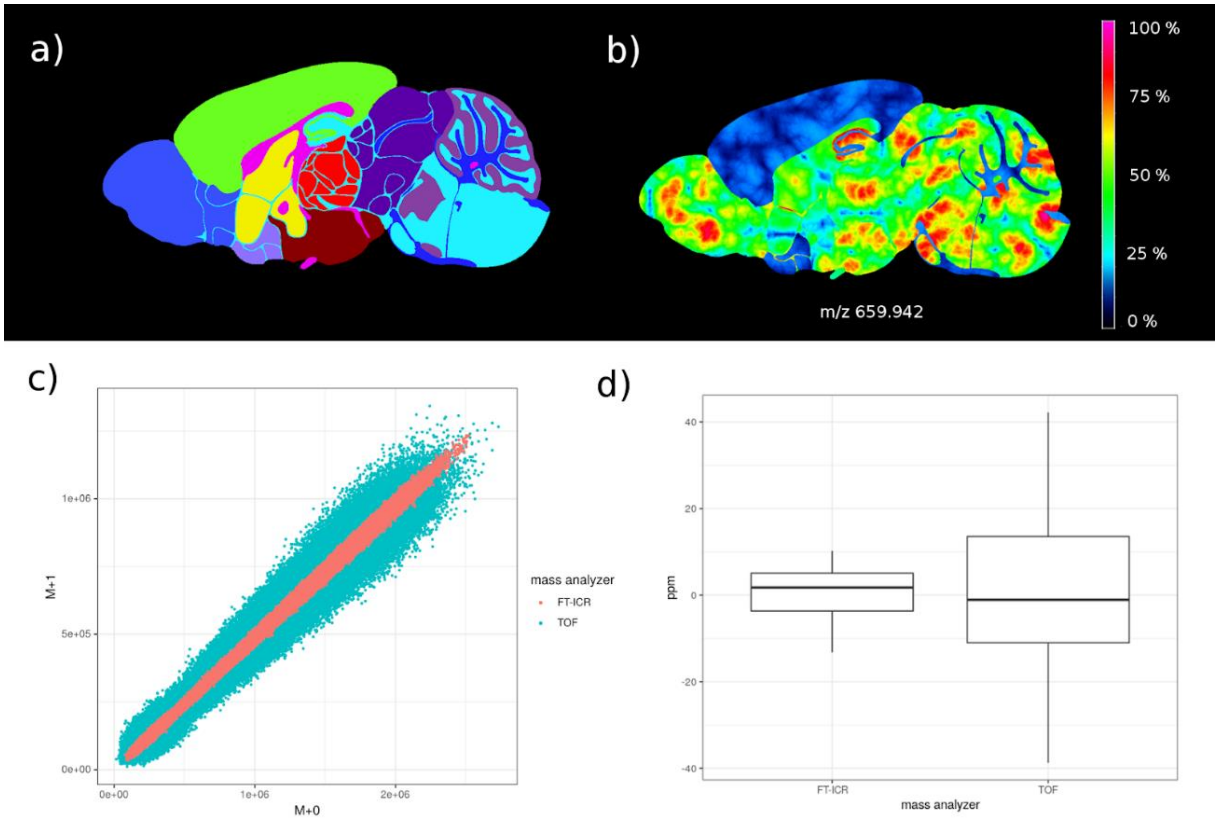


Figure S3. a) Brain segmented into 11 regions used as support for the *in silico* data **b)** Intensity representation of the *in silico* ion *m/z* 659.942. The up-regulated and down-regulated regions are distinguishable, and the simplex noise pattern generates globular structures all over the image. **c)** Scatter plot of the intensity of one the M+0 ions against its M+1 intensity from both peak matrices. Each point represents a pixel of the image. Different levels of noise were consistently applied through all the pixels in an attempt to simulate the performance of different mass analyzers. **d)** Box plot of the mass deviation introduced into the mass axis of the peak matrices. This mass indetermination represents the difference in the performance of peak picking algorithms with TOF and FT-ICR mass analyzers.

7.3. ILS threshold optimization

In order to determine the ILS optimal threshold to detect the maximum number of real monoisotopic peaks without introducing artifacts and to evaluate the robustness of rMSIannotation, we worked on *in silico* datasets plus experimental MSI data. In this section we will outline the procedure that we implemented for ILS threshold determination. The development of the in silico datasets is shown in section 2 of the supplementary materials and the experimental datasets are introduced in section 2 of the main article and explored in section 4 of the main article.The *in silico* datasets were analyzed using ROC curve analysis and the Matthews correlation coefficient (MCC) as a measure of the quality of binary classifiers.[4] To create the ROC curve, we ranged the ILS threshold from 0 to 1 in 200 steps. For each score value, we calculated its confusion matrix: a matrix that summarizes the number of correct and incorrect predictions that the algorithm can make on a set of test data for which the true values are known (i.e. the *in silico* dataset). The confusion matrix has four types of observations: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), which are necessary for defining precision (TP/(TP+FP)) and MCC.Figure S4 summarizes the performance tests of the annotation algorithm. Solid lines represent the FT-ICR *in silico* peak

matrix and dashed lines the TOF *in silico* peak matrix. The recommended ILS threshold is reached when MCC (red line) and precision (blue line) coincide at 1, which means that there are no misclassified peaks. The recommended ILS threshold for the *in silico* TOF database is in the range 0.55~0.7, and for *in silico* FT-ICR it is in the range 0.7~0.8. A higher threshold will ensure more robust annotations (precision keeps scoring 1) at the expense of reducing the number of annotated peaks (represented by decreasing MCC). These results show that, in real datasets, the optimal threshold will depend on the resolving power of the instrument. We recommend thresholds of 0.55~0.7 for TOF datasets and 0.7~0.8 for FT-ICR datasets as starting points.

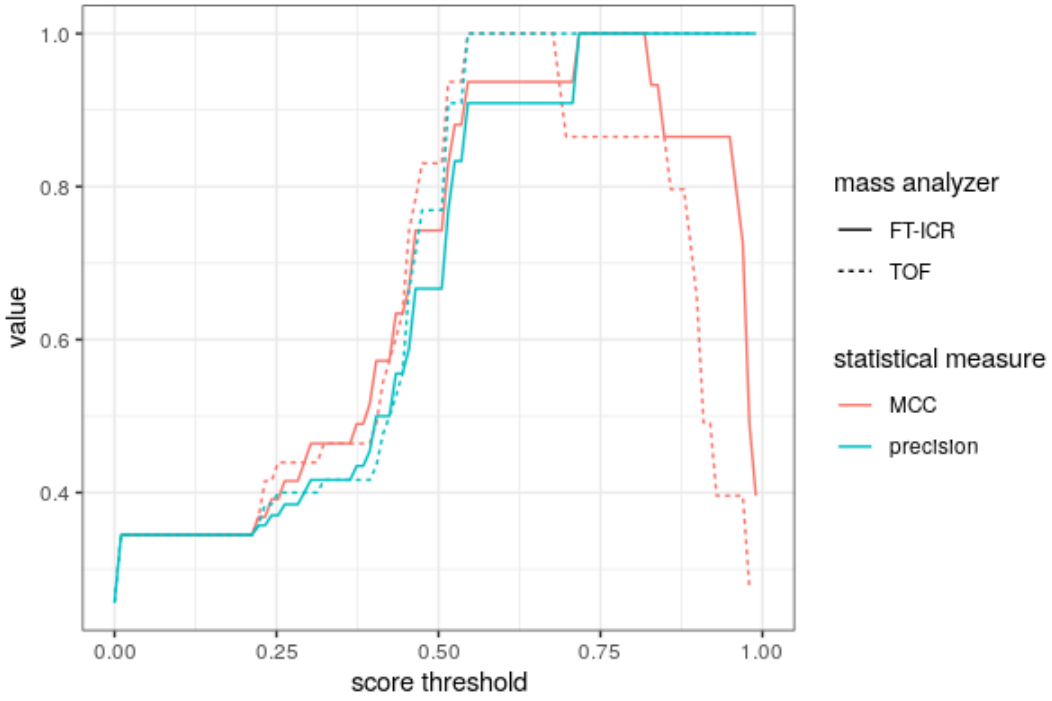


Figure S4. Analysis of recommended Isotopic likelihood score thresholds for the *in silico* FT-ICR (solid line) and TOF (dashed line) peak matrix using precision (TruePositives/(TruePositives+FalsePositives)) and Matthews correlation coefficient (MCC). Optimal thresholds are in the range 0.55~0.7 for TOF datasets and in the range 0.7~0.8 for FT-ICR datasets.

After evaluating this result, we tested thresholds between 0.2 and 1 in steps of 0.05 for the experimental datasets. For each threshold, we compared the ions considered to be monoisotopes by the algorithm with those identified by the authors of the datasets. Figure S5 shows the number of coinciding annotations for each dataset at each threshold. The results show that, in both cases, the number of coinciding annotations remains constant close to the recommended thresholds and rapidly decreases after them. Low thresholds values produce more annotations with less confidence, whereas higher thresholds give us more reliable results but less annotated peaks.

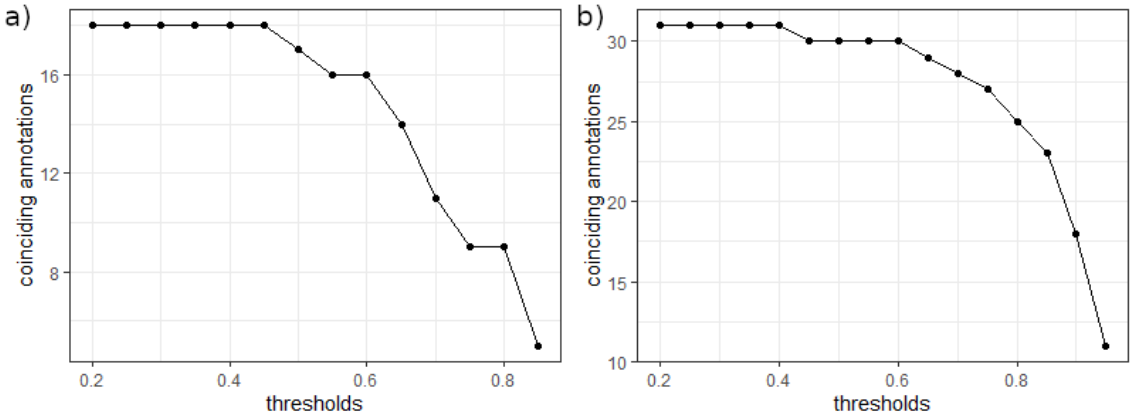


Figure S5. Analysis of Isotopic likelihood score threshold for the experimental datasets. a) TOF dataset. b) FT-ICR dataset 1. In both cases, the number of coinciding annotations starts to decrease around the recommended thresholds found with the *in silico* datasets.

7.4. Overlapping isotopic patterns

To investigate how rMSIannotation responds to overlapping isotopic patterns, we have studied *in silico* the case of two molecules detected in the TOF dataset with failed annotations due to overlap: phosphatidylcholine(32a:1) with molecular formula $[C_{40}H_{78}NO_8P+Na]^+$ and *m/z* 754.5357, and sphingomyelin(d18:1/C17:0) with molecular formula $[C_{40}H_{81}N_2O_6P + K]+$ and *m/z* 755.5464 (from now on, molecule A and B). The monoisotopic peak (M+0) of molecule B and the M+1 peak of molecule A overlaps, with a difference in the *m/z* ratio of 0.0073 Da. Peaks M+1 peak of molecule B and M+2 peak of molecule A also overlaps. To generate the spectra we used enviPat [1] in three different resolving powers: 40k, 120k and 240k, with abundance rations between molecules from 0 to 2; this is the abundance of molecule B over molecule A. Later, we peak picked the *in silico* spectra and obtained a peak list in which we computed the Isotope likelihood score (ILS) for the pairs of peaks with the *m/z* ratio closest to the theoretic isotopic pattern. In a real experiment, if there are big differences between the spatial distribution of molecules A and B, the peak picking algorithm could probably detect the overlapped peaks as individuals, since one molecule could be detected in a region and the other in a different one, as the spectra in those regions are not overlapped. Nevertheless, in this *in silico* analysis we considered the worst-case-scenario with a base morphology score of 0.95, in which both molecules exhibit the same morphology.

Figures S6 to S8 show the results of this procedure for different resolving powers. The top panel of the figures display the isotopic patterns profiles for both molecules and the resulting overlap. Rows represent different peaks and columns represent discrete values of the abundance ratio between molecules. First row represents the M+0 peak of molecule A, second row the M+0 peak of molecule B overlapped with the M+1 peak of molecule A and third row, the overlap between the M+1 peak of molecule B and the M+2 peaks of molecule A. The bottom panel shows how ILS is affected by changes in the abundance ratio between molecules for both monoisotopic peaks and how the deformation of the profile affects the peak picking algorithm, reducing or increasing the *m/z* ratio difference in ppms from the theoretic value for each peak. For resolving powers below 120k the peaks are completely overlapped, allowing the simultaneous annotation of both compounds only in a narrow range of abundance ratios. Figure S6 shows the results for a resolving power of 40k in which, rMSIannotation annotates both compounds in the range of abundance ratios of 0.2 to 0.4. For resolving powers near 120k, the range of abundances for simultaneous annotation increases. Figure S7 shows the results for a resolving power of 120k in which, rMSIannotation annotates both compounds in the range

of abundance ratios of 0.2 to 1. For resolving power higher than 120k, the simultaneous annotation of both molecules is possible in a bigger abundance ratio range. Figure S8 shows the results for a resolving power of 240k in which, rMSIannotation annotates both compounds for the complete range of abundance ratios tested. To summarize, rMSIannotation can annotate overlapping compounds in a range of abundance ratio between the overlapping compounds which depends on the resolving power of the dataset. As the resolving power increases, the range of abundance ratio for simultaneous annotation also increases and, in case of being outside of this ranges, the compound with a higher relative intensity will be the one annotated.

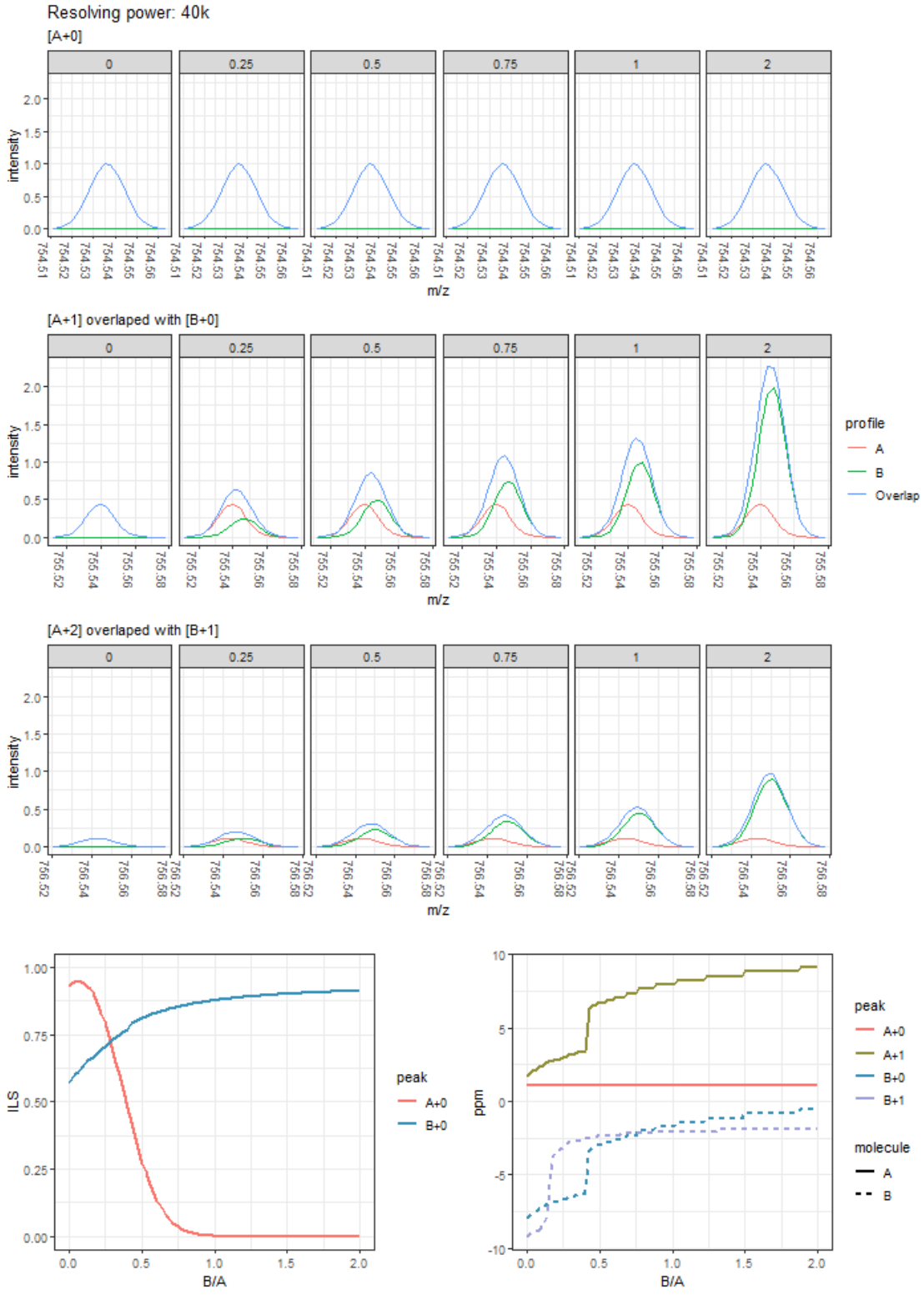


Figure S6. Analysis of *in silico* overlapping isotopic patterns with a resolving power of 40k. Top panel displays the profile of the peaks for each molecule and for the overlap. Bottom panel displays isotopic likelihood score and *m/z* ratio ppm error to the theoretic value for each peak. In the graphics, *A* refers to ion $[C_{40}H_{78}NO_8P+Na]^+$ and *B* refers to ion $[C_{40}H_{81}N_2O_6P + K]^+$. In this case, the overlapping allows the annotation of both compounds (ILS > 0.6, for low resolving power datasets) only in the range of abundance ration from 0.2 to 0.4 approximately.

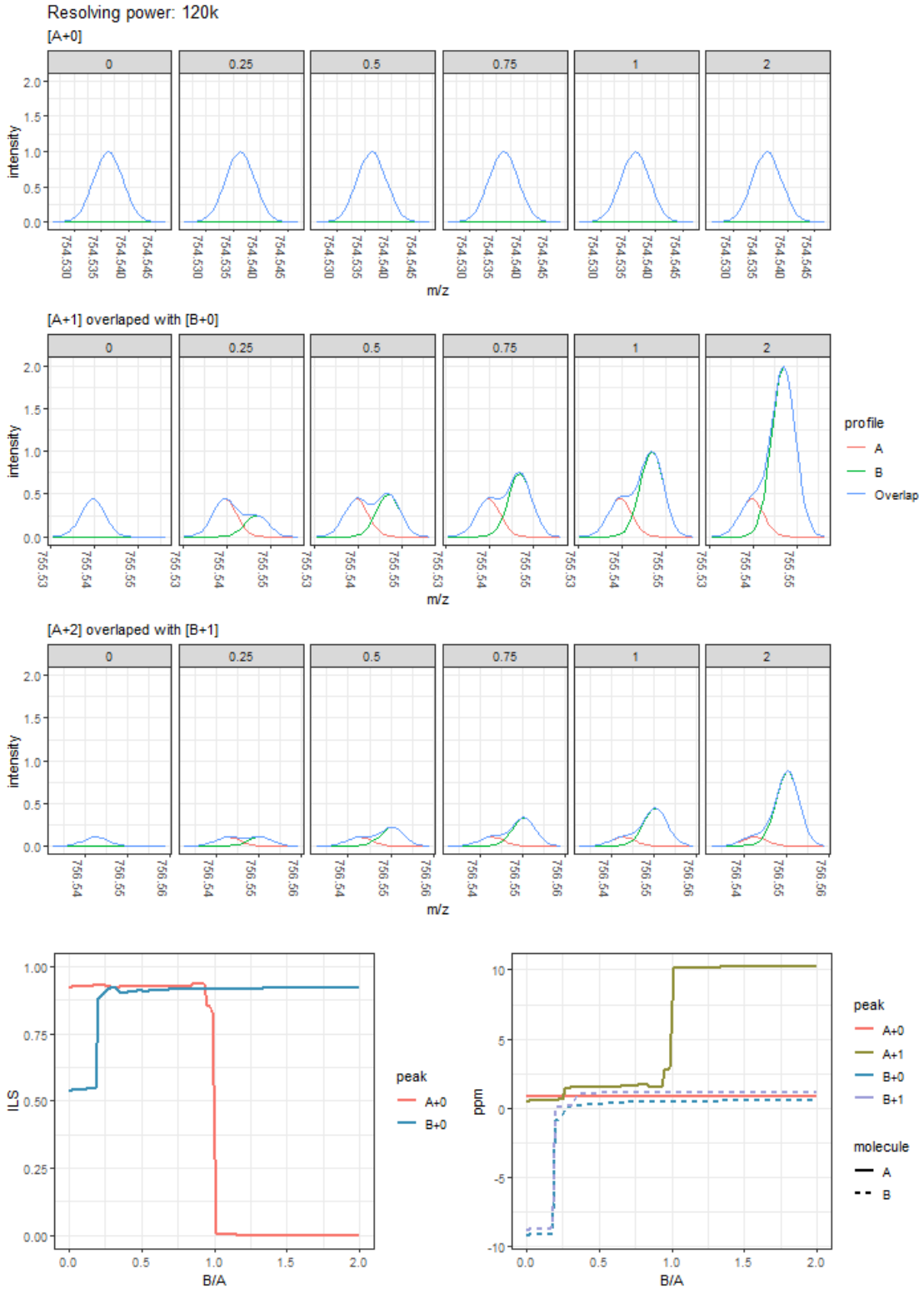


Figure S7. Analysis of *in silico* overlapping isotopic patterns with a resolving power of 120k. Top panel displays the profile of the peaks for each molecule and for the overlap. Bottom panel displays isotopic likelihood score and *m/z* ratio ppm error to the theoretic value for each peak. In the graphics *A* refers to ion $[C40H78NO8P+Na]^+$ and *B* refers to ion $[C40H81N2O6P + K]^+$. In this case, the resolving power allows for the annotation of both compounds for the range of abundance ratio from 0.2 to 1. After that, the deformation on the peak shape A+1 due to the increase of abundance of peak B+0 hides the peak A+1.

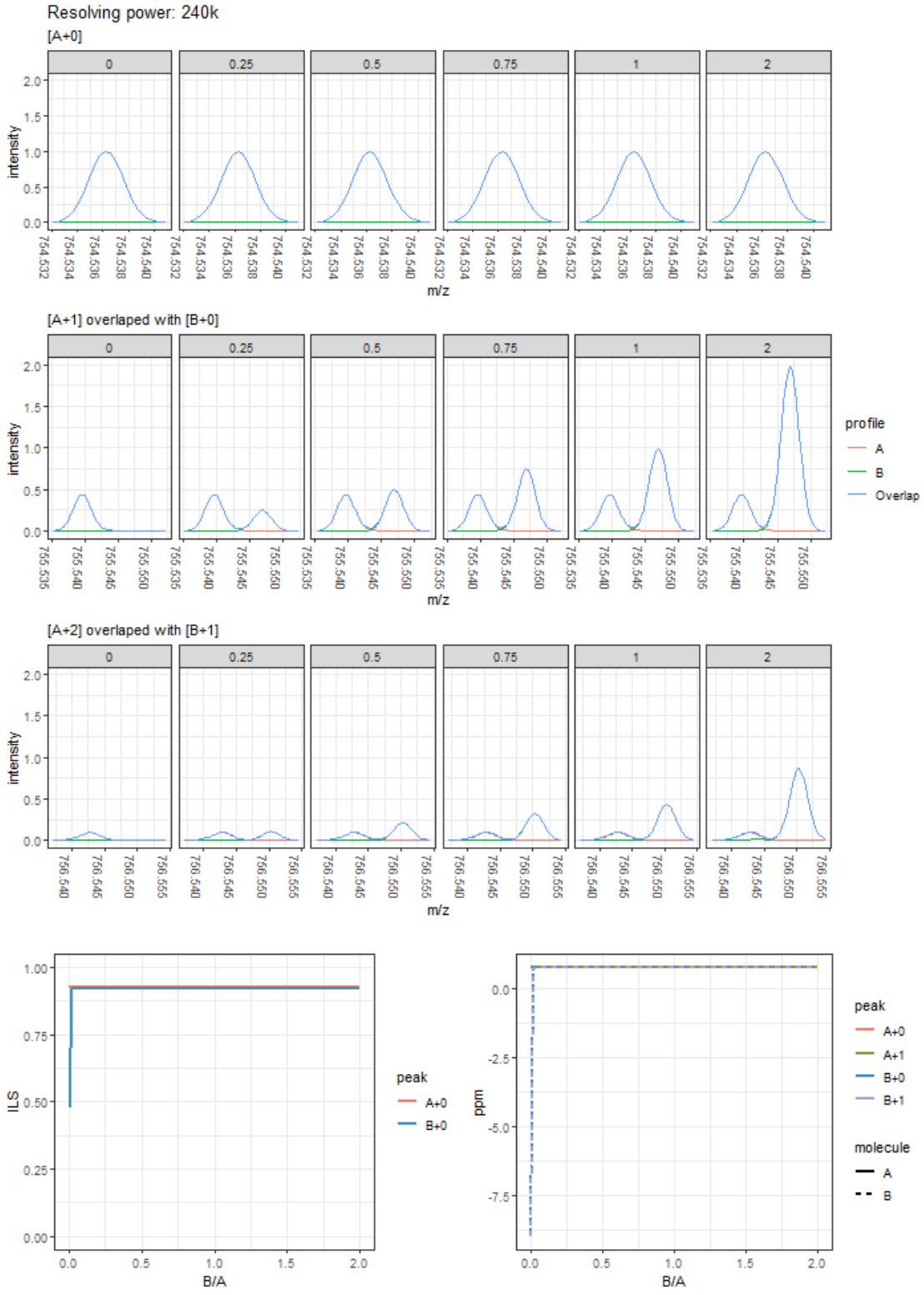


Figure S8. Analysis of *in silico* overlapping isotopic patterns with a resolving power of 240k. Top panel displays the profile of the peaks for each molecule and for the overlap. Bottom panel displays isotopic likelihood score and *m/z* ratio ppm error to the theoretic value for each peak. In the graphics *A* refers to ion $[C40H78NO8P+Na]^+$ and *B* refers to ion $[C40H81N2O6P + K]^+$. In this case, the resolving power allows the annotation of both compounds in the complete abundance ratio range.

7.5. Supplementary Figures

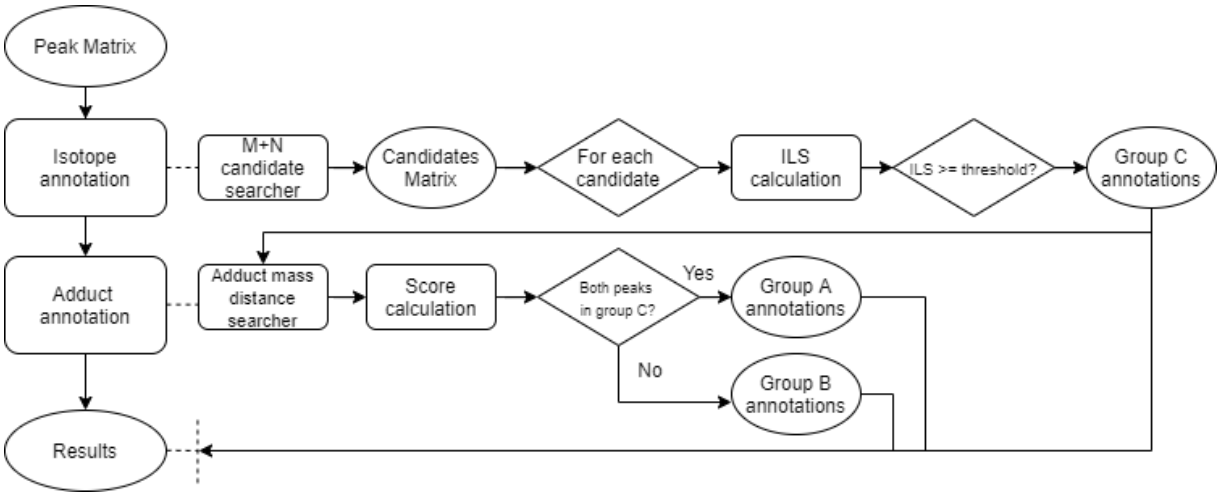


Figure S9. Flow diagram of the peak annotation algorithm rMSIannotation. Rounded objects refer to data structures and rectangles to algorithmic processes.

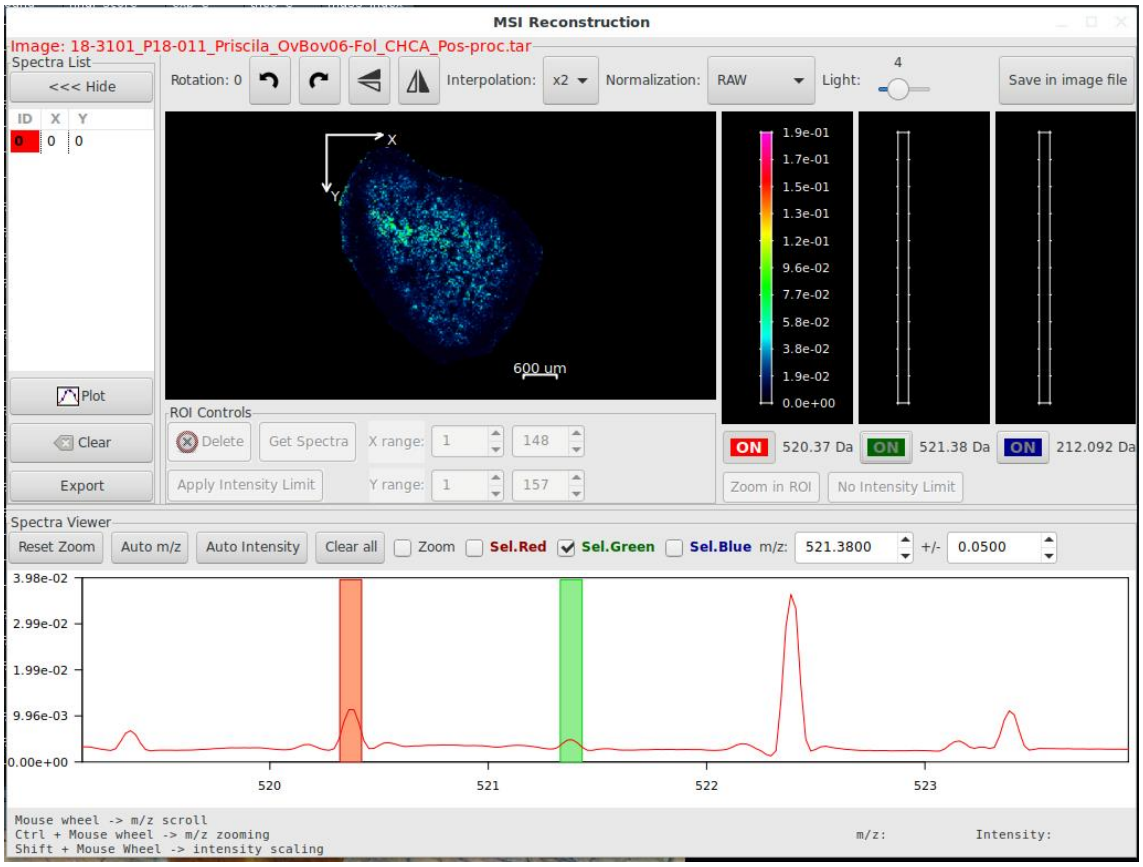


Figure S10. TOF dataset peak *m/z* 520.34 (red) cannot be labeled as M+0 by the algorithm because the processing parameters discarded peak *m/z* 521.34 (green) because it has SNR < 5

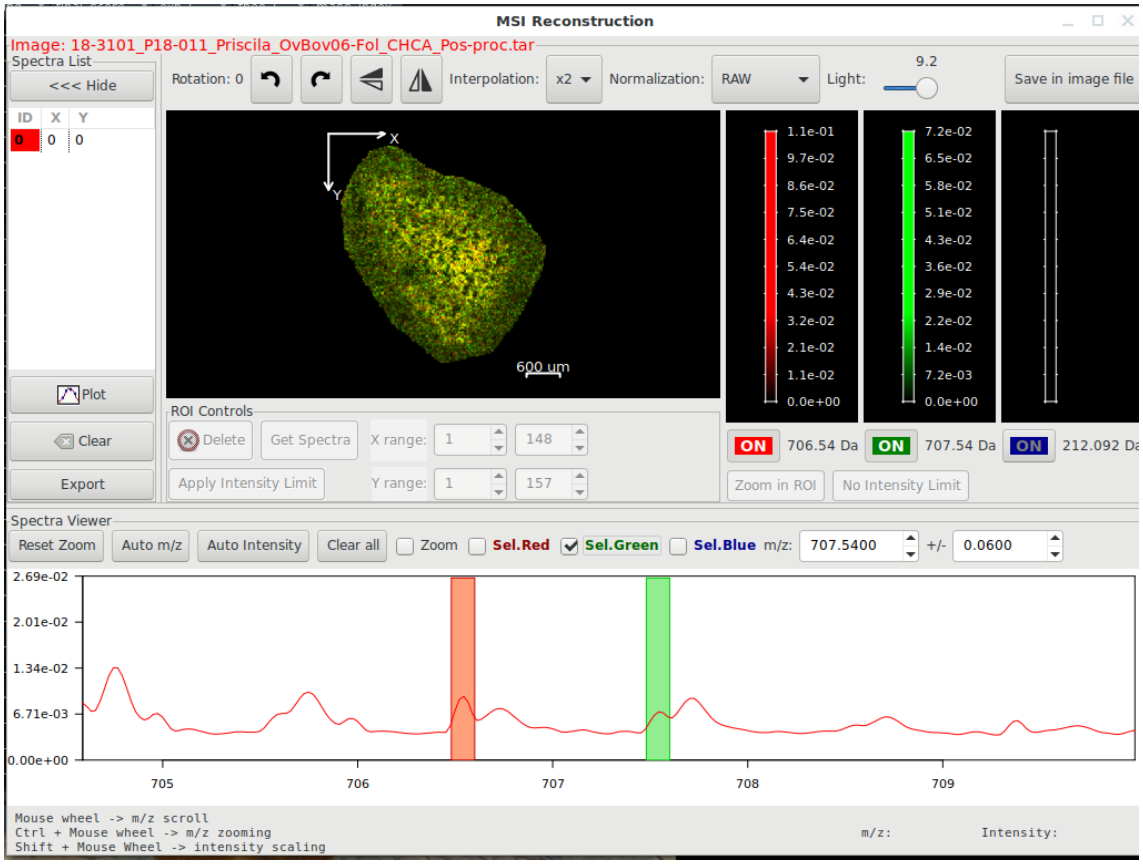


Figure S11. TOF dataset peaks *m/z* 706.54 (red) and *m/z* 707.54 (green) cannot be evaluated because neither of the ions has SNR > 5.

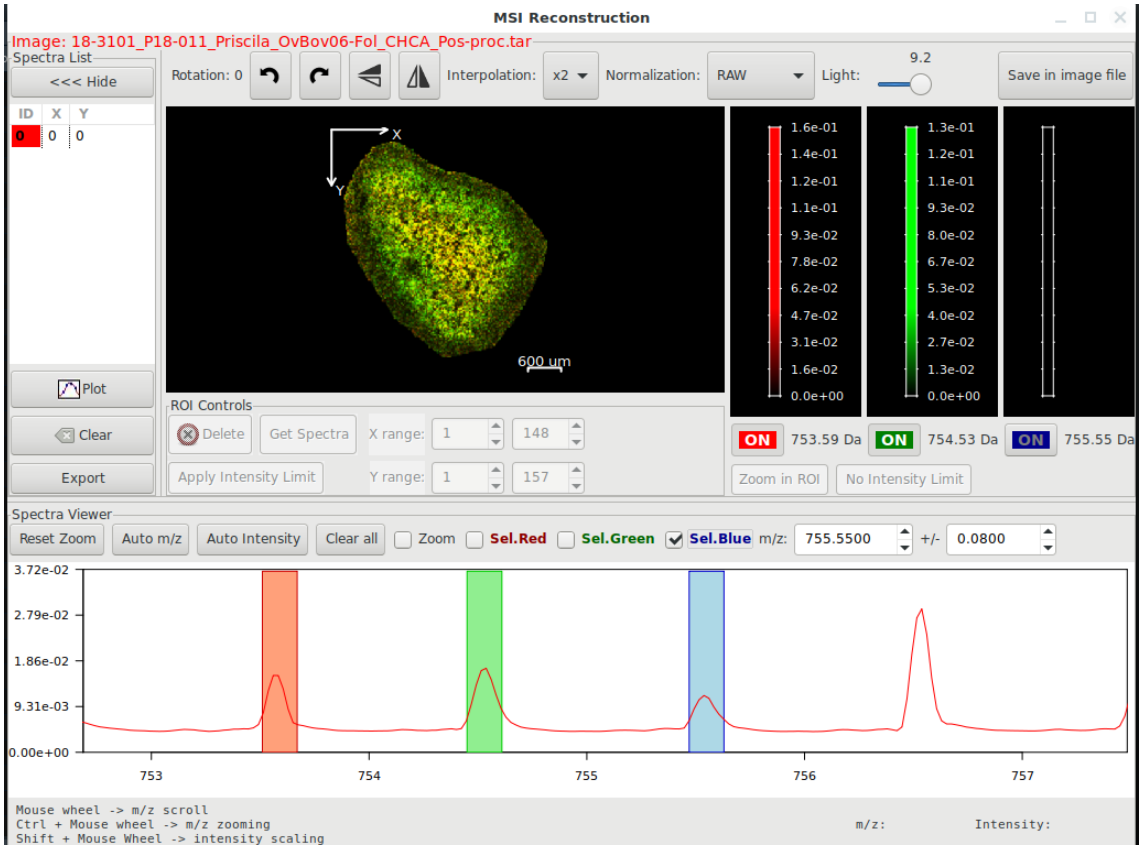


Figure S12. Overlapping of compounds SM(d18:1/C18:0) (exact neutral mass 730.5988 Da) and PC 32a:1 (exact neutral mass 731.5465 Da) of the TOF dataset. Peak *m/z* 753.588 (red) is the $[M+Na]^+$ monoisotopic ion of SM(d18:1/C18:0) which cannot be labeled as a monoisotopic ion because the M+1 peak in *m/z* 754.53 m/z (green) overlaps with the $[M+Na]^+$ monoisotopic ion of PC 32a:1 which, at the same time, cannot be labeled as monoisotopic because peak *m/z* 755.54 (blue) has not enough SNR.

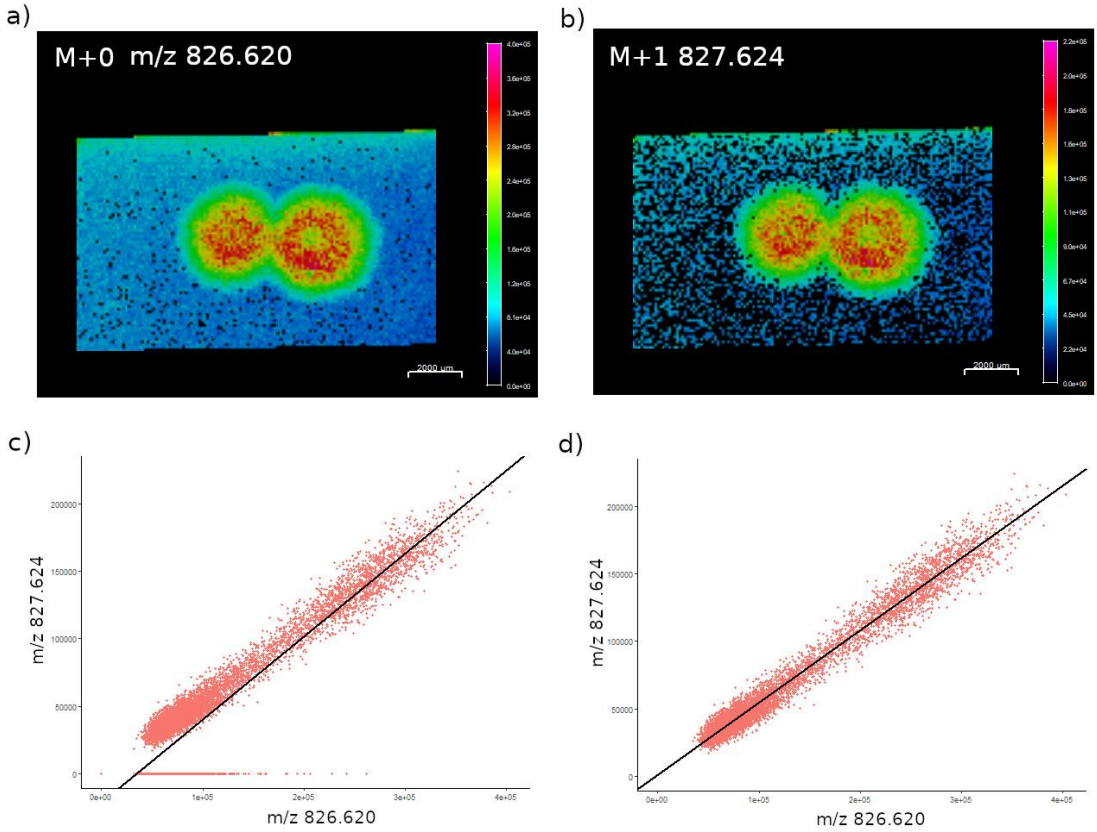


Figure S13. a) Intensity map of M+0 *m/z* 826.620 of the FT-ICR dataset. b) Intensity map of M+1 *m/z* 827.624 of the FT-ICR dataset. c) Scatter plot of M+0 and M+1 ions without null pixel correction. The scatter plot shows that there are some pixels in the M+1 ion which have zero intensity due to the low intensity of this compound in some pixels of the sample. This produces bad linear modeling of the slope and correlation, which worsens the results of the isotopic likelihood test. d) Scatter plot of M+0 and M+1 ions after removing the pixels with zero intensity. Removing them leads to a better modeling of the data: the linearity increases, and the isotopic likelihood test results are better. The large number of observations in an MSI experiment means that some observations from the original data can be discarded without losing predictive power.

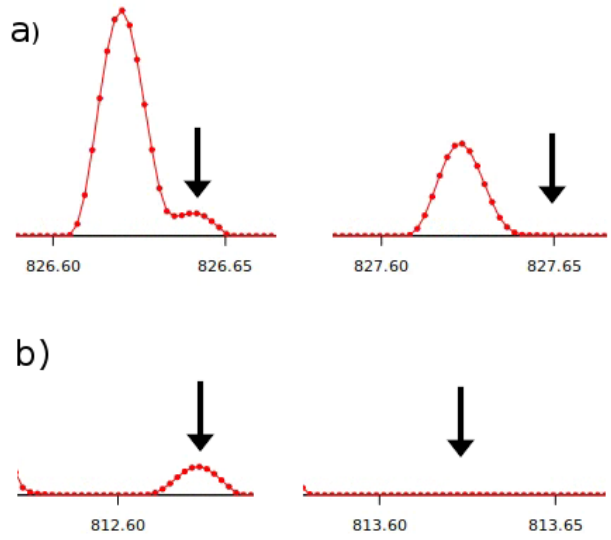


Fig S14. a) M+0 peak *m/z* 826.640 (the small one) and M+1 *m/z* 827.643 mean spectrum regions. b) M+0 peak *m/z* 812.622 and M+1 ion *m/z* 813.625 mean spectrum regions. In both cases, the M+1 peak does not appear in the peak matrix so the algorithm cannot evaluate them.

7.6. Supplementary Tables

Table. S1 Annotations produced by rMSIannotation in group A for the TOF dataset

Neutral mass (Da)	Adducts	Adduct 1 mass (*m/z*)	Adduct 2 mass (*m/z*)	Intensity ratio	Intensity ratio Std. Error	Correlation	Mass error (ppms)
189.114	$[M+Na]^+$ & $[M+H]^+$	212.106	190.120	0.137	0.067	0.838	25.269
211.101	$[M+K]^+$ & $[M+H]^+$	250.067	212.106	0.066	0.004	0.933	22.176
227.074	$[M+K]^+$ & $[M+Na]^+$	266.035	250.067	0.064	0.002	0.983	27.230
227.079	$[M+Na]^+$ & $[M+H]^+$	250.067	228.088	0.062	0.001	0.944	13.801
378.149	$[M+K]^+$ & $[M+H]^+$	417.113	379.158	0.189	0.053	0.731	0.981
438.071	$[M+K]^+$ & $[M+Na]^+$	477.037	461.058	0.169	0.001	0.945	12.397
495.357	$[M+K]^+$ & $[M+H]^+$	534.318	496.367	0.203	0.006	0.863	9.558
649.040	$[M+K]^+$ & $[M+Na]^+$	688.002	672.031	0.386	0.009	0.909	4.558
702.550	$[M+K]^+$ & $[M+Na]^+$	741.510	725.543	0.407	0.036	0.945	9.243
733.540	$[M+K]^+$ & $[M+Na]^+$	772.508	756.526	0.368	0.001	0.849	10.747
733.545	$[M+Na]^+$ & $[M+H]^+$	756.526	734.562	0.403	0.033	0.819	24.233
733.549	$[M+K]^+$ & $[M+H]^+$	772.508	734.562	0.402	0.034	0.804	13.485
759.552	$[M+K]^+$ & $[M+Na]^+$	798.517	782.539	0.437	0.001	0.939	5.764
759.558	$[M+Na]^+$ & $[M+H]^+$	782.539	760.574	0.402	0.033	0.855	21.943
759.560	$[M+K]^+$ & $[M+H]^+$	798.517	760.574	0.404	0.035	0.835	16.179
785.572	$[M+K]^+$ & $[M+H]^+$	824.530	786.585	0.434	0.030	0.745	13.158

Table. S2 Annotations produced by rMSIannotation in group B for the TOF dataset

Neutral mass (Da)	Adducts	Adduct 1 mass (*m/z*)	Adduct 2 mass (*m/z*)	Correlation	Mass error (ppms)
183.132	$[M+H]^+$ & $[M+Na]^+$	184.141	206.121	0.864	10.223
183.134	$[M+H]^+$ & $[M+K]^+$	184.141	222.099	0.824	16.182
211.099	$[M+H]^+$ & $[M+Na]^+$	212.106	234.089	0.940	5.311
211.101	$[M+K]^+$ & $[M+Na]^+$	250.067	234.089	0.978	16.866
334.163	$[M+H]^+$ & $[M+Na]^+$	335.167	357.156	0.652	19.583
400.117	$[M+H]^+$ & $[M+Na]^+$	401.125	423.107	0.851	0.297
400.122	$[M+H]^+$ & $[M+K]^+$	401.125	439.090	0.889	24.868
416.103	$[M+H]^+$ & $[M+Na]^+$	417.113	439.090	0.905	10.700
422.095	$[M+K]^+$ & $[M+Na]^+$	461.058	445.086	0.913	4.682
422.097	$[M+K]^+$ & $[M+H]^+$	461.058	423.107	0.809	10.768
438.078	$[M+K]^+$ & $[M+H]^+$	477.037	439.090	0.853	20.421
454.048	$[M+Na]^+$ & $[M+H]^+$	477.037	455.056	0.826	1.246
495.357	$[M+K]^+$ & $[M+Na]^+$	534.318	518.349	0.879	9.114
495.359	$[M+H]^+$ & $[M+Na]^+$	496.367	518.349	0.839	0.445
521.376	$[M+H]^+$ & $[M+Na]^+$	522.387	544.364	0.775	10.22
521.377	$[M+H]^+$ & $[M+K]^+$	522.387	560.338	0.799	9.130
633.071	$[M+K]^+$ & $[M+Na]^+$	672.031	656.063	0.814	8.881
702.561	$[M+Na]^+$ & $[M+H]^+$	725.543	703.576	0.830	21.654
757.544	$[M+H]^+$ & $[M+Na]^+$	758.548	780.537	0.718	9.593
757.550	$[M+H]^+$ & $[M+K]^+$	758.548	796.523	0.755	24.617
785.566	$[M+K]^+$ & $[M+Na]^+$	824.530	808.555	0.767	1.385
785.571	$[M+H]^+$ & $[M+Na]^+$	786.585	808.555	0.750	14.543

Table. S3 Annotations produced by rMSIannotation in group C for the TOF dataset

Monoisotopic mass (*m/z*)	**ILS**	**Isotopic intensity ratio**
172.113	0.957	0.134
184.140	0.842	0.073
184.265	0.603	0.306
190.119	0.865	0.204
212.106	0.967	0.070
228.088	0.950	0.061
250.066	0.923	0.062
266.034	0.893	0.065
294.145	0.958	0.128
335.167	0.933	0.135
379.157	0.977	0.241
379.914	0.888	0.139
401.124	0.876	0.166
417.113	0.781	0.136
428.420	0.719	0.219
461.058	0.774	0.170
477.037	0.756	0.167
496.366	0.848	0.197
522.386	0.812	0.239
524.391	0.746	0.230
534.318	0.662	0.208
578.431	0.799	0.428
635.018	0.686	0.427
635.031	0.706	0.433
635.117	0.605	0.367
672.031	0.843	0.376
676.779	0.933	0.437
688.002	0.829	0.395
692.720	0.879	0.430
725.542	0.853	0.371
733.711	0.739	0.458
734.562	0.866	0.435
741.510	0.916	0.443
749.617	0.772	0.394
756.526	0.672	0.369
758.548	0.705	0.397
760.573	0.829	0.369
772.508	0.641	0.367
782.539	0.877	0.435
786.584	0.692	0.404
798.517	0.866	0.438
824.530	0.642	0.464

Table. S4 Ions manually annotated in the original publication but not found in group C during the annotation procedure due to: (1) the M+1 peak is not in the peak matrix, (2) M+0 and M+1 are not in the peak matrix and (3) peak overlap. (*) Although the algorithm found the monoisotopic peak of choline and evaluated it with a good morphology score and isotopic pattern profile score (both above 0.9), the mass error score is very low (0.55) resulting in a final score of 0.53. This kind of problem is common with TOF mass analyzers in the low m/z ratio range, where mass accuracy decreases exponentially.

Name	Molecular formula	Adduct	Exact mass (*m/z*)	Misclassification group
Choline (*)	$C_5H_{14}NO$	$[M+H]^+$	104.173	*
LPC 16a:0	$C_{24}H_{50}NO_7P$	$[M+Na]^+$	518.348	1
LPC 18a:2	$C_{26}H_{50}NO_7P$	$[M+H]^+$	520.340	1
LPC 20a:4	$C_{28}H_{50}NO_7P$	$[M+H]^+$	544.340	1
LPC 18a:1	$C_{26}H_{52}NO_7P$	$[M+K]^+$	560.311	1
SM(d18:1/C16:0)	$C_{39}H_{80}N_2O_6P$	$[M+H]^+$	703.575	1
PC 30a:0	$C_{38}H_{76}NO_8P$	$[M+H]^+$	706.539	2
PC 32a:1	$C_{40}H_{78}NO_8P$	$[M+H]^+$	732.554	1
PC 33a:1	$C_{41}H_{80}NO_8P$	$[M+H]^+$	746.570	2
PC 33a:0	$C_{41}H_{82}NO_8P$	$[M+H]^+$	748.585	2
SM(d18:1/C18:0)	$C_{41}H_{83}N_2O_6P$	$[M+Na]^+$	753.588	3
PC 32a:1	$C_{40}H_{78}NO_8P$	$[M+Na]^+$	754.535	1-3
SM(d18:1/C17:0)	$C_{40}H_{81}N_2O_6P$	$[M+K]^+$	755.546	2
PC 33a:1	$C_{41}H_{80}NO_8P$	$[M+Na]^+$	768.551	2-3
SM(d18:1/C18:0)	$C_{41}H_{83}N_2O_6P$	$[M+K]^+$	769.562	3
PC 33a:0	$C_{41}H_{82}NO_8P$	$[M+Na]^+$	770.567	1
PE 38:1	$C_{43}H_{84}NO_8P$	$[M+H]^+$	774.601	2
PC 34a:2	$C_{42}H_{80}NO_8P$	$[M+Na]^+$	780.551	1
PC 34a:0	$C_{42}H_{84}NO_8P$	$[M+Na]^+$	784.582	1
PC 36a:1	$C_{44}H_{86}NO_8P$	$[M+H]^+$	788.616	1
PC 35a:2	$C_{43}H_{82}NO_8P$	$[M+Na]^+$	794.567	2
PC 35a:1	$C_{43}H_{84}NO_8P$	$[M+Na]^+$	796.582	1
PC 36a:3	$C_{44}H_{82}NO_8P$	$[M+Na]^+$	806.567	2
PC 36a:2	$C_{44}H_{84}NO_8P$	$[M+Na]^+$	808.582	1
PC 36a:1	$C_{44}H_{86}NO_8P$	$[M+Na]^+$	810.598	1
PC 36:0	$C_{44}H_{88}NO_8P$	$[M+Na]^+$	812.614	2
PC 36a:3	$C_{44}H_{82}NO_8P$	$[M+K]^+$	822.584	2
PC 36a:1	$C_{44}H_{86}NO_8P$	$[M+K]^+$	826.572	1

Table S5. Putative name and molecular formula for some compounds annotated by rMSIannotation in group C.

m/z	Adduct	Molecular formula	Mass error (ppm)	Name
428.420	$[M+H]^+$	$C_{26}H_{53}NO_3$	25	Cer(26:0)
578.431	$[M+K]^+$	$C_{30}H_{60}NO_7P$	27	LPC(22:1)
733.711	$[M+H-H2O]^+$	$C_{48}H_{94}O_5$	5	DG(45:0)
749.617	$[M+Na]^+$	$C_{47}H_{82}O_5$	15	DG(44:5)

Table. S6 Annotations produced by rMSIannotation in group A for the FT-ICR dataset 1

Neutral mass (Da)	Adducts	Adduct 1 mass (*m/z*)	Adduct 2 mass (*m/z*)	Intensity ratio mean	Intensity ratio Std. Error	Correlation	Mass error (ppms)
312.182	$[M+Na]^+$ & $[M+H]^+$	335.171	313.189	0.190	0.0088	0.159	0.576
312.182	$[M+K]^+$ & $[M+Na]^+$	351.145	335.171	0.185	0.0035	0.644	0.558
312.182	$[M+K]^+$ & $[M+H]^+$	351.145	313.189	0.194	0.0052	-0.170	1.134
528.526	$[M+K]^+$ & $[M+Na]^+$	567.490	551.516	0.355	0.0120	0.184	0.057
530.542	$[M+K]^+$ & $[M+Na]^+$	569.506	553.531	0.368	0.0146	0.394	0.797
544.558	$[M+K]^+$ & $[M+Na]^+$	583.521	567.547	0.379	0.0129	0.482	0.082
602.226	$[M+Na]^+$ & $[M+H]^+$	625.216	603.234	0.314	0.0251	0.599	0.220
624.367	$[M+Na]^+$ & $[M+H]^+$	647.357	625.375	0.413	0.0015	0.758	0.372
624.367	$[M+K]^+$ & $[M+Na]^+$	663.331	647.357	0.398	0.0135	0.869	0.700
624.367	$[M+K]^+$ & $[M+H]^+$	663.331	625.375	0.399	0.0150	0.486	0.328
624.378	$[M+Na]^+$ & $[M+H]^+$	647.368	625.385	0.435	0.0242	0.491	1.189
642.206	$[M+Na]^+$ & $[M+H]^+$	665.196	643.214	0.358	0.0027	0.562	0.788
709.586	$[M+Na]^+$ & $[M+H]^+$	732.576	710.593	0.407	0.0180	0.829	0.307
716.525	$[M+K]^+$ & $[M+H]^+$	755.489	717.534	0.451	0.0322	0.898	1.199
840.711	$[M+Na]^+$ & $[M+H]^+$	863.701	841.719	0.584	0.0033	0.568	0.207
842.727	$[M+Na]^+$ & $[M+H]^+$	865.717	843.735	0.605	0.0005	0.620	0.037
852.720	$[M+Na]^+$ & $[M+H]^+$	875.712	853.727	0.511	0.0134	0.858	3.437
854.725	$[M+Na]^+$ & $[M+H]^+$	877.715	855.733	0.558	0.0104	0.526	0.322
856.742	$[M+Na]^+$ & $[M+H]^+$	879.732	857.750	0.610	0.0053	0.651	0.289
870.565	$[M+K]^+$ & $[M+H]^+$	909.530	871.573	0.704	0.0364	0.242	1.416
870.565	$[M+Na]^+$ & $[M+H]^+$	893.556	871.573	0.751	0.0101	0.829	1.277
870.566	$[M+K]^+$ & $[M+Na]^+$	909.530	893.556	0.714	0.0466	0.849	0.139
876.721	$[M+K]^+$ & $[M+Na]^+$	915.685	899.711	0.584	0.0270	0.842	0.368
880.752	$[M+Na]^+$ & $[M+H]^+$	903.744	881.758	0.580	0.0170	0.614	3.758
904.752	$[M+Na]^+$ & $[M+H]^+$	927.743	905.759	0.597	0.0087	0.497	1.755
926.736	$[M+Na]^+$ & $[M+H]^+$	949.727	927.743	0.615	0.0273	0.191	2.885
954.390	$[M+Na]^+$ & $[M+H]^+$	977.379	955.398	0.576	0.0581	0.327	0.491
1052.645	$[M+K]^+$ & $[M+Na]^+$	1091.613	1075.639	0.569	0.0076	0.863	0.571
1088.878	$[M+Na]^+$ & $[M+H]^+$	1111.863	1089.880	0.628	0.0197	0.147	1.121
1102.666	$[M+K]^+$ & $[M+Na]^+$	1141.630	1125.656	0.592	0.0052	0.882	0.755
1116.904	$[M+Na]^+$ & $[M+H]^+$	1139.894	1117.912	0.657	0.0510	0.235	0.808

Table. S7 Annotations produced by rMSIannotation in group B for the FT-ICR dataset 1

Neutral mass (Da)	Adducts	Adduct 1 mass (*m/z*)	Adduct 2 mass (*m/z*)	Correlation	Mass error (ppms)
154.022	$[M+Na]^+$ & $[M+H]^+$	177.012	155.030	-0.030	0.010
154.023	$[M+Na]^+$ & $[M+H]^+$	177.012	155.031	0.354	4.658
176.005	$[M+H]^+$ & $[M+Na]^+$	177.012	198.995	0.346	4.337
274.205	$[M+Na]^+$ & $[M+H]^+$	297.194	275.213	0.305	3.365
289.191	$[M+Na]^+$ & $[M+H]^+$	312.181	290.198	-0.152	4.682
289.191	$[M+Na]^+$ & $[M+H]^+$	312.181	290.198	-0.112	4.638
289.191	$[M+Na]^+$ & $[M+H]^+$	312.181	290.198	-0.106	3.625
289.192	$[M+Na]^+$ & $[M+H]^+$	312.181	290.200	-0.167	1.756
290.199	$[M+Na]^+$ & $[M+H]^+$	313.189	291.206	-0.079	3.517
290.200	$[M+Na]^+$ & $[M+H]^+$	313.189	291.209	-0.393	4.321
401.266	$[M+H]^+$ & $[M+Na]^+$	402.274	424.256	0.791	0.271
424.197	$[M+H]^+$ & $[M+Na]^+$	425.206	447.186	0.938	3.728
446.223	$[M+H]^+$ & $[M+K]^+$	447.230	485.186	0.935	0.621
446.223	$[M+H]^+$ & $[M+Na]^+$	447.230	469.212	0.823	0.481
448.198	$[M+H]^+$ & $[M+Na]^+$	449.206	471.188	0.302	0.612
450.214	$[M+H]^+$ & $[M+Na]^+$	451.222	473.204	0.394	0.017
464.193	$[M+H]^+$ & $[M+Na]^+$	465.201	487.183	0.185	0.015
464.193	$[M+H]^+$ & $[M+K]^+$	465.201	503.157	0.942	0.309
465.201	$[M+Na]^+$ & $[M+K]^+$	488.191	504.165	0.311	0.149
480.189	$[M+H]^+$ & $[M+Na]^+$	481.196	503.178	-0.034	0.583
487.183	$[M+H]^+$ & $[M+Na]^+$	488.191	510.173	0.483	0.297
487.183	$[M+H]^+$ & $[M+K]^+$	488.191	526.147	-0.177	0.813
512.278	$[M+H]^+$ & $[M+Na]^+$	513.286	535.268	0.798	0.641
514.294	$[M+H]^+$ & $[M+Na]^+$	515.302	537.283	0.560	1.055
528.526	$[M+K]^+$ & $[M+H]^+$	567.490	529.533	0.802	1.215
528.526	$[M+Na]^+$ & $[M+H]^+$	551.516	529.533	0.547	1.158
530.542	$[M+Na]^+$ & $[M+H]^+$	553.531	531.550	0.553	0.069
530.542	$[M+K]^+$ & $[M+H]^+$	569.506	531.550	0.796	0.728
537.283	$[M+H]^+$ & $[M+K]^+$	538.291	576.247	0.074	0.024
542.542	$[M+Na]^+$ & $[M+K]^+$	565.532	581.506	0.183	0.040
542.542	$[M+Na]^+$ & $[M+H]^+$	565.532	543.550	0.627	0.077
544.557	$[M+K]^+$ & $[M+H]^+$	583.521	545.565	0.775	0.641
544.557	$[M+Na]^+$ & $[M+H]^+$	567.547	545.565	0.579	0.723
558.417	$[M+K]^+$ & $[M+Na]^+$	597.380	581.408	0.065	3.086
572.481	$[M+Na]^+$ & $[M+H]^+$	595.472	573.487	0.501	4.632
587.405	$[M+K]^+$ & $[M+H]^+$	626.368	588.413	0.344	0.480
592.268	$[M+H]^+$ & $[M+Na]^+$	593.276	615.258	0.178	0.876
600.368	$[M+H]^+$ & $[M+Na]^+$	601.375	623.358	0.286	3.065
600.512	$[M+Na]^+$ & $[M+H]^+$	623.503	601.519	0.353	4.154
602.226	$[M+Na]^+$ & $[M+K]^+$	625.216	641.190	0.480	0.488
602.226	$[M+H]^+$ & $[M+K]^+$	603.234	641.190	0.226	0.268
612.475	$[M+Na]^+$ & $[M+K]^+$	635.465	651.439	0.180	0.114
622.497	$[M+H]^+$ & $[M+Na]^+$	623.503	645.488	0.093	3.928
624.208	$[M+H]^+$ & $[M+Na]^+$	625.216	647.198	0.463	0.308
624.208	$[M+H]^+$ & $[M+K]^+$	625.216	663.172	-0.060	0.504
624.378	$[M+H]^+$ & $[M+K]^+$	625.385	663.343	0.156	3.137

624.379	$[M+Na]^+$ & $[M+K]^+$	647.368	663.343	0.448	1.948
625.360	$[M+H]^+$ & $[M+Na]^+$	626.368	648.349	0.735	1.408
640.362	$[M+H]^+$ & $[M+Na]^+$	641.370	663.351	0.663	0.857
654.342	$[M+H]^+$ & $[M+Na]^+$	655.350	677.331	0.243	1.483
709.586	$[M+H]^+$ & $[M+K]^+$	710.593	748.550	0.755	1.243
709.586	$[M+Na]^+$ & $[M+K]^+$	732.576	748.550	0.971	0.936
712.507	$[M+Na]^+$ & $[M+K]^+$	735.496	751.470	0.974	0.283
732.473	$[M+Na]^+$ & $[M+K]^+$	755.463	771.437	0.983	0.028
738.243	$[M+H]^+$ & $[M+Na]^+$	739.251	761.232	0.103	0.318
766.466	$[M+Na]^+$ & $[M+K]^+$	789.455	805.430	0.981	0.488
776.379	$[M+H]^+$ & $[M+Na]^+$	777.387	799.368	0.754	0.409
776.689	$[M+Na]^+$ & $[M+K]^+$	799.679	815.653	0.982	0.211
777.531	$[M+H]^+$ & $[M+Na]^+$	778.539	800.520	0.762	1.574
778.223	$[M+H]^+$ & $[M+Na]^+$	779.231	801.213	0.494	0.070
794.492	$[M+H]^+$ & $[M+Na]^+$	795.500	817.482	0.466	0.424
804.721	$[M+Na]^+$ & $[M+K]^+$	827.711	843.685	0.975	0.073
816.558	$[M+H]^+$ & $[M+Na]^+$	817.565	839.548	0.122	0.534
826.705	$[M+H]^+$ & $[M+Na]^+$	827.711	849.696	0.201	4.368
830.737	$[M+Na]^+$ & $[M+K]^+$	853.727	869.701	0.980	0.417
848.540	$[M+H]^+$ & $[M+Na]^+$	849.547	871.531	-0.044	3.293
848.604	$[M+K]^+$ & $[M+Na]^+$	887.570	871.593	0.864	3.653
850.704	$[M+Na]^+$ & $[M+H]^+$	873.695	851.712	0.894	1.079
865.609	$[M+K]^+$ & $[M+Na]^+$	904.572	888.599	0.964	1.384
874.705	$[M+H]^+$ & $[M+Na]^+$	875.712	897.696	0.928	2.456
876.722	$[M+K]^+$ & $[M+H]^+$	915.685	877.732	0.007	3.005
876.723	$[M+Na]^+$ & $[M+H]^+$	899.711	877.732	-0.344	2.637
877.562	$[M+H]^+$ & $[M+Na]^+$	878.570	900.551	0.634	1.480
878.723	$[M+H]^+$ & $[M+Na]^+$	879.732	901.712	0.313	1.967
878.740	$[M+Na]^+$ & $[M+H]^+$	901.729	879.749	0.420	2.408
892.550	$[M+H]^+$ & $[M+K]^+$	893.556	931.516	0.049	4.799
893.807	$[M+H]^+$ & $[M+Na]^+$	894.813	916.799	0.904	4.294
903.564	$[M+H]^+$ & $[M+Na]^+$	904.572	926.554	0.696	0.010
904.751	$[M+H]^+$ & $[M+K]^+$	905.759	943.715	0.834	0.224
904.752	$[M+Na]^+$ & $[M+K]^+$	927.743	943.715	0.225	1.531
908.524	$[M+H]^+$ & $[M+Na]^+$	909.530	931.516	0.713	4.582
914.643	$[M+Na]^+$ & $[M+K]^+$	937.634	953.605	0.288	2.582
922.705	$[M+Na]^+$ & $[M+K]^+$	945.696	961.668	0.394	2.001
924.722	$[M+Na]^+$ & $[M+K]^+$	947.711	963.687	0.345	1.603
926.736	$[M+H]^+$ & $[M+K]^+$	927.743	965.702	-0.375	3.178
926.738	$[M+Na]^+$ & $[M+K]^+$	949.727	965.702	0.271	0.293
928.519	$[M+Na]^+$ & $[M+K]^+$	951.509	967.482	-0.220	1.070
950.736	$[M+H]^+$ & $[M+Na]^+$	951.743	973.726	0.599	1.083
972.229	$[M+H]^+$ & $[M+Na]^+$	973.236	995.218	0.358	0.030
1100.657	$[M+Na]^+$ & $[M+K]^+$	1123.649	1139.614	-0.128	0.712
1142.927	$[M+H]^+$ & $[M+Na]^+$	1143.923	1165.910	0.266	0.434
1182.753	$[M+Na]^+$ & $[M+H]^+$	1205.744	1183.757	-0.331	0.548
1188.903	$[M+Na]^+$ & $[M+H]^+$	1211.893	1189.912	0.188	0.388
1216.935	$[M+Na]^+$ & $[M+H]^+$	1239.926	1217.942	0.243	1.603
1238.919	$[M+H]^+$ & $[M+Na]^+$	1239.926	1261.910	0.533	1.425

Table. S8 Annotations produced by rMSIannotation in group C for the FT-ICR dataset 1

Monoisotopic mass (*m/z*)	**ILS**	**Isotopic intensity ratio**
177.012	0.726	0.062
208.459	0.753	0.008
273.037	0.734	0.116
297.194	0.746	0.185
312.181	0.936	0.226
312.686	0.873	0.082
313.189	0.955	0.198
332.329	0.862	0.232
335.171	0.790	0.181
351.145	0.804	0.188
360.361	0.845	0.230
402.274	0.915	0.244
425.206	0.941	0.255
435.785	0.833	0.137
447.230	0.717	0.217
449.206	0.880	0.269
451.222	0.731	0.237
465.201	0.777	0.247
481.196	0.936	0.271
488.191	0.804	0.250
513.286	0.778	0.279
514.293	0.937	0.311
515.302	0.805	0.302
530.493	0.706	0.277
535.429	0.796	0.302
538.291	0.747	0.274
549.488	0.756	0.323
551.516	0.898	0.367
553.531	0.920	0.382
558.524	0.937	0.389
565.532	0.854	0.359
567.490	0.827	0.343
567.547	0.923	0.391
569.506	0.862	0.353
581.399	0.951	0.405
583.521	0.865	0.366
584.540	0.919	0.392
585.223	0.845	0.314
586.556	0.926	0.402
591.878	0.749	0.299
593.276	0.913	0.326
595.472	0.826	0.382
597.380	0.716	0.314
601.375	0.724	0.298
603.234	0.783	0.289
612.571	0.923	0.408
623.503	0.731	0.364
625.216	0.881	0.339
625.375	0.995	0.414
625.385	0.887	0.459

626.368	0.723	0.248
635.465	0.815	0.365
641.370	0.847	0.364
643.214	0.885	0.360
647.357	0.989	0.411
647.368	0.832	0.411
649.519	0.784	0.397
655.350	0.776	0.319
661.053	0.790	0.362
663.331	0.938	0.384
665.196	0.816	0.354
667.313	0.854	0.346
672.541	0.853	0.384
689.502	0.959	0.401
695.468	0.791	0.344
695.504	0.844	0.430
700.573	0.798	0.370
703.518	0.769	0.360
710.593	0.906	0.425
717.415	0.803	0.400
717.534	0.935	0.418
723.499	0.703	0.343
726.588	0.861	0.387
732.576	0.829	0.389
735.496	0.714	0.405
737.392	0.718	0.367
739.251	0.825	0.370
743.550	0.827	0.405
755.463	0.938	0.431
755.489	0.940	0.483
761.392	0.835	0.419
772.573	0.996	0.501
777.387	0.865	0.422
778.539	0.955	0.468
779.231	0.758	0.422
786.588	0.992	0.505
789.455	0.802	0.429
789.534	0.984	0.505
789.550	0.766	0.406
793.381	0.733	0.394
795.500	0.972	0.474
797.069	0.873	0.482
797.241	0.902	0.477
799.679	0.943	0.498
800.376	0.885	0.452
800.604	0.961	0.505
803.550	0.883	0.532
809.516	0.836	0.435
811.553	0.833	0.613
817.565	0.976	0.516
821.594	0.700	0.428
823.366	0.806	0.424
823.531	0.941	0.469
826.620	0.984	0.535

827.711	0.919	0.511
830.579	0.928	0.518
832.585	0.767	0.491
837.062	0.949	0.535
841.205	0.921	0.473
841.719	0.927	0.587
843.581	0.879	0.509
843.735	0.924	0.605
845.664	0.735	0.450
849.547	0.971	0.519
851.554	0.826	0.465
853.727	0.805	0.497
855.733	0.900	0.568
857.750	0.933	0.615
859.540	0.785	0.430
863.701	0.934	0.580
865.717	0.943	0.604
866.782	0.973	0.552
871.573	0.730	0.740
872.603	0.953	0.595
873.695	0.845	0.510
875.712	0.891	0.524
877.715	0.884	0.547
878.570	0.968	0.567
879.732	0.950	0.604
880.798	0.861	0.525
881.758	0.929	0.597
887.570	0.875	0.583
893.556	0.704	0.761
894.813	0.964	0.582
895.531	0.858	0.509
899.711	0.966	0.610
901.729	0.829	0.608
902.578	0.796	0.511
903.744	0.870	0.563
904.572	0.846	0.543
905.461	0.899	0.573
905.759	0.912	0.605
908.589	0.720	0.477
909.530	0.864	0.667
913.587	0.736	0.453
915.685	0.871	0.556
920.829	0.860	0.549
922.845	0.701	0.480
927.743	0.899	0.587
937.400	0.751	0.495
937.634	0.910	0.603
945.696	0.883	0.609
947.711	0.907	0.633
949.727	0.914	0.642
951.509	0.882	0.537
951.743	0.888	0.624
953.759	0.927	0.646
955.398	0.969	0.634

959.542	0.921	0.592
966.812	0.884	0.613
973.236	0.921	0.638
977.379	0.750	0.518
978.907	0.736	0.530
993.574	0.737	0.541
1008.595	0.718	0.761
1025.600	0.889	0.668
1035.624	0.853	0.608
1045.726	0.949	0.695
1075.639	0.831	0.577
1089.880	0.834	0.607
1091.413	0.872	0.691
1091.613	0.729	0.561
1111.863	0.841	0.647
1113.555	0.784	0.657
1117.912	0.799	0.606
1123.640	0.718	0.582
1125.656	0.776	0.596
1131.393	0.732	0.665
1139.894	0.898	0.708
1141.630	0.726	0.586
1143.928	0.875	0.811
1161.616	0.815	0.828
1205.740	0.702	0.916
1211.895	0.741	0.691
1213.712	0.800	0.665
1237.909	0.769	0.814
1239.926	0.717	0.708

Supplementary Table S9. Ions not found by the annotation algorithm that were manually annotated in the original publication for the FT-ICR dataset 1. The M+1 ion of most of them does not appear in the peak matrix or there are too many null pixels to be corrected.

Name	Molecular formula	Adduct	Exact mass (Da)
PDPT 32:1	$C_{40}H_{77}O_8PS$	$[M+H]^+$	749.515
BLL 36:6 (*)	$C_{46}H_{75}NO_{10}$	$[M+H]^+$	802.547
vGSL-like (t16:0/h22:0)	$C_{44}H_{87}NO_{10}$	$[M+Na]^+$	812.622
TAG 47:1	$C_{50}H_{94}O_6$	$[M+Na]^+$	813.695
vGSL (t17:0/h22:1)	$C_{45}H_{87}NO_{10}$	$[M+Na]^+$	824.622
vGSL (t17:0/h22:0)	$C_{45}H_{89}NO_{10}$	$[M+Na]^+$	826.640
TAG 50:1	$C_{53}H_{100}O_6$	$[M+Na]^+$	855.746
sGSL d18:2	$C_{49}H_{91}NO_{11}$	$[M+Na]^+$	892.649
TAG 54:1	$C_{57}H_{108}O_6$	$[M+Na]^+$	911.804

Supplementary Table. S10 Putative name and molecular formula for some compounds annotated by rMSIannotation. (*) Common alkenones specific of *Emiliana huxleyi*. The annotation level indicates the table from which the annotated adduct was found.

m/z	Adduct	ILS	Annotation group	Molecular formula	Mass error (ppm)	Putative annotation
297.194	$[M+Na]^+$	0.746	B	$C_{17}H_{26}N_2O$	3	Alkaloid
313.189	$[M+H]^+$	0.955	A	$C_{19}H_{24}N_2O_2$	4	Alkaloid
335.171	$[M+Na]^+$	0.790	A	$C_{19}H_{24}N_2O_2$	5	Alkaloid
351.145	$[M+K]^+$	0.804	A	$C_{19}H_{24}N_2O_2$	3	Alkaloid
513.286	$[M+H]^+$	0.778	B	$C_{30}H_{40}O_7$	3	Terpenoid
515.302	$[M+H]^+$	0.805	B	$C_{30}H_{42}O_7$	4	Terpenoid
535.268	$[M+Na]^+$	--	B	$C_{30}H_{40}O_7$	3	Terpenoid
(*) 553.531	$[M+Na]^+$	0.920	A	$C_{37}H_{70}O$	2	Heptatriacontadien
(*) 565.532	$[M+Na]^+$	0.854	B	$C_{38}H_{70}O$	2	Octratriacontadienone
(*) 567.547	$[M+Na]^+$	0.923	B	$C_{38}H_{72}O$	1	Octratriacontadienone
(*) 569.506	$[M+K]^+$	0.862	A	$C_{37}H_{70}O$	1	Heptatriacontadien
581.399	$[M+H]^+$	0.951	C	$C_{40}H_{52}O_3$	1	Carotenoid
(*) 583.521	$[M+K]^+$	0.865	A	$C_{38}H_{72}O$	1	Octratriacontadienone
585.223	$[M+K]^+$	0.845	C	$C_{33}H_{38}O_7$	2	Xanthone
635.465	$[M+Na]^+$	0.815	B	$C_{39}H_{64}O_5$	1	DG(36:6)
735.496	$[M+Na]^+$	0.714	B	$C_{47}H_{68}O_5$	1	DG(44:12)
751.471	$[M+K]^+$	--	B	$C_{47}H_{68}O_5$	1	DG(44:12)
865.717	$[M+Na]^+$	0.943	A	$C_{50}H_{82}O_8S$	4	Sulfoglycerolipid(40:7)
843.735	$[M+H]^+$	0.924	A	$C_{50}H_{82}O_8S$	4	Sulfoglycerolipid(40:7)

7.7. Supplementary References

(1) Loos, M.; Gerber, C.; Corona, F.; Hollender, J.; Singer, H. Accelerated Isotope Fine Structure Calculation Using Pruned Transition Trees. *Analytical Chemistry* 2015, *87* (11), 5738–5744. https://doi.org/10.1021/acs.analchem.5b00941.

(2) Lein, E. S. et al. Genome-Wide Atlas of Gene Expression in the Adult Mouse Brain. *Nature* 2007, *445* (7124), 168–176. https://doi.org/10.1038/nature05453.

(3) Pedersen, T. L.; Peck, J. Ambient: A Generator of Multidimensional Noise. 2018.

(4) Chicco, D.; Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics* 2020, *21* (1), 6. https://doi.org/10.1186/s12864-019-6413-7.

CHAPTER 5

Fuzzy c-means improves the evaluation of segmentation processes in mass spectrometry imaging

Abstract: Objective. Mass spectrometry imaging (MSI) produces molecular images of biological tissues by localizing mass spectra over their surface. The images are used to study the distribution of molecules in different regions of the sample to extract information on disease dysregulations, cell epigenetics or drug localization. Multiple hard clustering algorithms have been used to define regions of interest over the sample. But it is common to observe molecular images with fluctuations in intensity inside and between regions, pointing to soft transitions between regions. Therefore, the use of soft clustering methods capable of representing these transitions is required. Methods. We combine the soft clustering algorithm fuzzy c-means and MSI data to represent the transitions between clusters and present a new score to evaluate different clustering results. Results. We study the effects of the algorithm's parameters on MSI data. We develop a workflow to analyze multiple human cancer samples using the new score. Which allows us to see the transition between different tissue types and clusters associated with cancer and healthy tissue. Significance. Soft clustering algorithms allow to unveil and represent the transitions between different tissue types in a spatial segmentation analysis of MSI data, which result in clusters more closely related to how biologic tissue types are interconnected.

1. Introduction

The organs that constitute animal bodies are composed of a variety of biological tissues that by themselves are very heterogeneous, as they are formed by diverse cell types. Thus, various families of compounds like small molecules, lipids, and proteins are distributed differently across the several morphological structures that form the tissues. In recent years, mass spectrometry imaging (MSI) has become an invaluable tool [1], [2] to unveil the spatial distribution of molecules directly on biological tissues by producing ion images with spatial resolutions down to 1 μm[3]–[6]. For each sampling point (pixel), a spectrum contains information on the abundance of typically hundreds or thousands of ions desorbed from the tissue. An ion image is generated by visualizing the intensity values of the ion at each pixel. One of the most usual approaches to combining the information from all ion images is spatial segmentation, which categorizes the spectral data into a certain number of regions or segments localized on the tissue. Studying the differences in molecular signatures between segments can lead to the discovery of specific biomarkers for pathologies or intrinsic differences between tissue regions and even cell types, which is a common scenario in studies related to cancer, like tumor margins of tumor classification[7]. In many cases, regions of interest (ROI) are defined by correspondence with histological annotations of consecutive samples. However, the unsupervised definition of segments based on spectral similarity is an alternative method in case of lacking histologically defined ROIs. To do so, the most common approach is applying clustering methods.

Clustering methods are unsupervised machine learning procedures that find groups of data with common characteristics using similarity metrics. The most widespread segmentation methods in MSI are k-means[8], [9] and hierarchical clustering[10]–[13], due to their simplicity and coverage in different programming languages and software packages. Both algorithms are based on assigning each pixel to a unique cluster by comparing them using various metrics, like euclidean distance, the cosine similarity, or Pearson correlation. K-means groups the data in a user-defined number of clusters maximizing the distance between cluster centroids, while hierarchical clustering develops a complete dendrogram of different numbers of clusters by grouping or splitting previous steps in the dendrogram. For instance, these algorithms have been used widely in the study of cancer[14]–[17], and plant metabolism[18]–[20].

Ideally, all the segments in the tissue constitute groups of strongly related pixels. But, as the sampling beam in MSI typically cannot be focused down to the level of single cells, different types of cells are averaged at each pixel during the acquisition. This issue might lead to pixels where various types of cells coexist in different proportions and therefore, assigning each pixel to only one segment might lead to a wrong interpretation of the data. This is a severe limitation for clustering methods, and specific algorithms need to be used to handle the problem. The clustering methods that are able to assign multiple clusters to one pixel are known as soft (or fuzzy) clustering methods. Although most clustering methods used in MSI data are based on hard clustering methods, research on soft clustering for MSI led to some innovative developments. The most notable example is the clustering framework implemented in the R package Cardinal[21] combining spatially aware clustering[9], statistical regularization[22], [23], and probabilistic clustering[24]. Also, a two-step workflow based on probabilistic clustering and smoothing using Latent Dirichlet Allocation and Markov random field was proposed by *Chernyavsk et al.*[25]. More recently, spatial-DGMM (Dirichlet Gaussian Mixture Model) [26] was presented, combining GMM and spatially aware clustering. All these methods use probabilistic clustering, which assesses the probability of a pixel belonging to a cluster. These probabilities could be used to study the soft borders between clusters, but the principal results of these methods end up assigning a unique cluster to each pixel.

Still, there are more soft clustering algorithms used in different fields of image processing suitable for MSI: The simplest soft clustering method, both theoretically and in hardware resource consumption, is Fuzzy c-means (FCM)[27], [28], which originated from the extension of the classical k-means algorithm to fuzzy sets. Fuzzy sets allow pixels to belong to all clusters in different degrees of membership. The membership displays the distance of a pixel to each of the cluster centroids. Pixels with high membership to a cluster will have a spectrum like that of the centroid of the cluster, and low membership values to the other clusters. Like k-means, the number of clusters to determine and the distance metric used to evaluate similarity fundamentally affect the results. FCM also introduces the “fuzzifier”, usually labeled as m. The fuzzifier is a number higher than 1 that shapes the membership curve. Values close to 1 produce clusters with high membership pixels and almost no overlap, while bigger values reduce the overall membership of the clusters and produce more fuzziness between clusters. Controlling this parameter is a key aspect of FCM, but no definitive consensus has been reached on choosing an optimal value, and traditionally, 2 is considered to be the standard value for some authors, as it tends to group centroids close to the geometrical center of the data[29]–[34]. Supplementary figure 1 shows examples of membership curves using different values of the fuzzifier.

In MSI there are only a few published works involving the use of FCM. *Jones et al.* combined multiple statistical analyses, including FCM, to study the heterogeneity of myxofibrosarcoma[35]. In their work, FCM was one further piece of a combined workflow and the only stated information is the use of the euclidean distance and the selection of a fuzzifier of 1.25 (which was the default value in the employed Matlab method). The authors concluded that combining five independent multivariate methods provided an accurate summary of the spatio-chemical heterogeneity of myxofibrosarcoma. *Sakari et al.* did an exploratory study of the performance of different MSI workflows using FCM and K-means[36]. In their article, they evaluated the effects of different distance metrics on k-means: euclidean, city-block, cosine, and correlation; and the use of principal component analysis (PCA) before clustering to transform the *m/z* features into principal components. To do so, they used two MALDI-MSI datasets of mouse brain tissue from different spatial perspectives (sagittal and coronal) clusterized into 2 to 10 clusters to compute the Calinski-Harabasz (CH) clustering quality index and the correlation of the clusters with the manual annotation of 21 ion images representing predominant spatial patterns. For FCM with euclidean distance (the only tested setting), they

conclude that using PCA to reduce variables had limited effects on the cluster localization and shape. Regarding the CH index, they observed that 2 clusters produce the highest values in most workflows and associate it with the differentiation between tissue and non-tissue. Moreover, they observed that the CH index is biased towards the euclidean distance, as the euclidean distance minimizes the within-cluster sum of squares. The authors recommended exploring more cluster validity indexes, but multiple opinions in the MSI community consider that these indexes are not of much interest for MSI data as they estimate mathematically optimal solutions that do not need to go according to the biology of a tissue[37], [38]. In the end, the authors conclude that limited results were achieved for FCM and more investigation is required.

Following these investigations, we here explore the benefits of FCM in MSI using the dimension of membership to a cluster for every pixel. We apply FCM to three MSI datasets to observe the effects over the membership that different distance metrics, fuzzifiers, and spatial resolutions have in different MSI data analysis procedures. Moreover, we propose a new score to compare the distribution of the membership between clustering results and to select the most representative pixels of a cluster. Finally, we propose a semi-supervised workflow combining FCM and histologically defined ROIs to study human cancer samples.

2. Materials & Methods

Indium tin oxide (ITO)-coated glass slides were obtained from Bruker Daltonics (Bremen, Germany). The gold-target used for sputtering coating was obtained from Kurt J. Lesker Company (Hastings, England) with a purity grade higher than 99.995%. HPLC grade xylene was supplied by Sigma–Aldrich (Steinheim, Germany), and ethanol (96% purity) was supplied by Scharlau (Sentmenat, Spain).

2.1. MSI datasets

To explore the use of soft clustering, we have selected three different datasets. The first consists of a MALDI-TOF dataset of sagittal cut of a complete mouse brain. With this dataset, we study how the membership of the clusters is affected using different distance metrics, fuzzifiers, and the number of clusters. The second consists of a MALDI-Orbitrap dataset of two samples of the mouse cerebellum sampled with different pixel sizes. With this dataset, we study the effects that different spatial resolutions have on the membership of soft clusters. The third consists of a MALDI-TOF dataset of six human head and neck cancer samples. With this dataset, we implement a workflow for the analysis of MSI datasets using soft clustering and study the heterogeneity in tumoral tissue and the role of transition tissue between tumoral and healthy tissue.

2.1.1. MALDI-TOF sagittal mouse brain

The dataset consists of one sagittal slide of a complete mouse brain. Complete details on sample preparation and mice model handling can be found in the article by *del Castillo et a*l.[39]. The sample was sectioned from fresh-frozen material in slices 10 µm thick using a Leica CM-1950 cryostat (Leica Biosystems, Nussloch, Germany) and mounted on ITO slides. Gold nanolayers were deposited on the 10 µm tissue sections using an ATC Orion 8-HV sputtering system (AJA International, N. Scituate, MA, USA) following the procedures described by *Ràfols et al.*[40] The slide was measured using a MALDI TOF/TOF UltrafleXtreme instrument with SmartBeam II Nd:YAG/355 nm laser from Bruker Daltonics with a spatial resolution of 80 µm. Acquisitions were carried out using the medium and large laser spot size settings, operated at 2 kHz at an attenuated power of 60%, collecting a total of

500 shots per pixel. The TOF spectrometer operated in reflectron positive mode with the digitizer set at a sample rate of 1.25 GHz in a mass range between *m/z* 70 and 1200.

2.1.2. MALDI-Orbitrap mouse cerebellum

The sample was sectioned from OCT-embedded material in slices 10 µm thick using a Leica CM-1950 cryostat (Leica Biosystems, Nussloch, Germany) and mounted on ITO slides and coated with Au as described by *Ràfols et al.*[40]. The samples were measured with MALDI-MSI using a dual-ion funnel/dual MALDI/ESI Injector (Spectroglyph, Kennewick, WA) coupled to a Q Exactive Plus Orbitrap (Thermo Fisher Scientific, Bremen, Germany), as described by Bednařík el al.[41] The in-source pressure was set to 7 mbar of N2. A frequency-tripled q-switched Nd:YLF laser (Explorer OEM, Spectra-Physics, Mountain View, CA; emission wavelength: 349 nm; repetition rate: 300 Hz; resulting pulse width: ~10 ns) was operated with a pulse energy of about 50% above the ablation threshold. The laser beam was focused to an effective spot size of ~10 µm in diameter. The dataset consists of two samples of the mouse cerebellum, one of them measured with a pixel-to-pixel step size of 50 µm (MC50) and the other with 10 µm (MC10). The orbitrap mass analyzer was operated with a mass resolving power (FWHM) of 140,000 (*@m/z* 200) and a fixed "injection time" of 500 ms, resulting in data acquisition rates of 1.9 pixels per second. The "AGC target" was disabled. Experiments were controlled by XCalibur (2.8.SP1 Build 2806, Thermo Fisher Scientific) and MALDI Injector (ver. 1.3.1.0, Spectroglyph) software.

2.1.3. MALDI-TOF human head and neck cancer

The dataset consists of six human tissue samples with head and neck cancer extracted from the same patient with the approval of the Institutional Review Board (IRB) of Hospital Clínic de Barcelona (FIS PI18/0844). Two of them contain tumoral cells and are categorized as tumoral tissue (S1 and S2, analytical replicates), two are close biopsies of the tumoral region and are categorized as a transition between tumoral and healthy tissue (S3 and S4), and two are healthy tissue (S5 and S6). The samples were sliced from fresh-frozen material, coated with Gold nanolayers, and measured using a MALDI-TOF UltrafleXtreme with SmartBeam II Nd:YAG/355 nm laser from Bruker Daltonics in the positive mode in the mass range of *m/z* 400 to 1200 as described by *Ràfols et al.*[40].

2.2. MSI data processing

For each dataset, the raw data were converted to imzML and processed using the default parameters of rMSIproc[42]. The processing workflow consisted of smoothing, alignment, mass calibration using known gold clusters peaks, peak picking, and peak binning, resulting in peak matrices. Peak matrices represent pixels as rows and *m/z* features as columns, so that each row is the peak-picked spectrum of a pixel, and each column is the intensity image of an *m/z* feature. Later, the off-sample and hotspot pixels were removed using TIC and RMS filters. Only the peaks corresponding to monoisotopic ions found by rMSIannotation[43] were retained for the two mouse datasets to work with fewer variables. For the human dataset we discarded only the isotopes to keep as many relevant ions for the classification as possible. Next, we removed all the ions with a median value of zero in the remaining pixels to discard columns representing off-sample ions and the pixels were normalized by TIC in the case of TOF and by RMS in the case of Orbitrap, as using RMS normalization on TOF amplifies the noise greater than the TIC normalization [44]. Finally, all data were standardized to allow all ions in different intensity and variance levels to contribute equally to the spatial segmentation of the sample using FCM.

2.3. Fuzzy c-means and pixel fidelity score

FCM is a soft clustering method that works in iterative steps in which the cluster centroids are calculated by weighing all the pixels (n) by their membership to the clusters. Internally, FCM contains two main structures: the centroid matrix, which contains the spectral centroid of the clusters; and the membership matrix, which contains the membership of pixels to clusters and can be understood as the spatial centroid of the clusters in MSI. Given a specific number of clusters (c), the membership (u_{ik}) of a pixel (k) to a cluster (i) is computed using the distance (d) between the pixel spectrum and the centroid of the cluster divided by the sum of all the distances between the pixel and all the centroids (j). This ratio is raised by a factor that includes the fuzzifier (m), which increases the degree of fuzziness of the clusters and is usually set to 2 by default [33], [34]. This produces membership values between 0 and 1 and the sum of all the memberships of a pixel adds up to 1. Equation 1 represents the membership definition of FCM.

$$u_{ik} = \left(\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}\right)^{-1} \qquad (1)$$

For this work, we have implemented our version of the Fuzzy c-means algorithm in an R package using C++ and Rcpp[45], as we were not able to process MSI datasets with other tested packages in R, like ppclust[46] and fclust[47], due to the maximum number of observation allowed. The code is available through github (github/LlucSF/fcmR). The clustering algorithm can be initialized by selecting the number of clusters, or using ROIs labels for each pixel, which assign a cluster to each ROI and compute the first centroids considering the maximum membership of the pixel to the ROI. Using ROIs generally reduces the number of iterations the algorithms need to converge, resulting in faster executions. Other parameters that can be modified are: 1) the number of clusters, 2) the maximum number of iterations, 3) the minimum number of iterations before checking stop conditions, 4) the fuzzifier (value of m), to control how abrupt the transition between clusters can be (supplementary figure 1); and 5) the epsilon, the difference in the objective function between iterations accepted as the convergence of the algorithm. In order to evaluate the influence of different distance metrics, we implemented the Euclidean and the cosine distance following the formulas of *Smets et al.*[48].

Additionally, to compare the membership values of results with different numbers of clusters, we have developed the pixel fidelity score (PFS), which is automatically computed in our implementation of FCM. The PFS is computed at pixel level and quantifies the degree of dependency of a pixel to a unique cluster in a range between zero and one. The PFS reaches one when a pixel is a centroid pixel, a theoretical situation in which a pixel and a cluster centroid share the same spectrum and therefore, the pixel only has membership for that cluster (Eq. 2). On the other hand, the PFS reaches zero when a pixel is an unclassified pixel, a situation of a pixel with the same value of membership for all clusters (3). Therefore, clusters with high PFS pixels have limited spatial overlap with other clusters, as the pixels belong strongly to that cluster; while clusters with pixels with low PFS pixels have pixels shared between clusters, and hence they overlap in the space.

$$\mu_{cp}(c) = (1, 0, \ldots, 0) \qquad (2)$$
$$\mu_{\emptyset}(c) = (1/c, 1/c, \ldots, 1/c) \qquad (3)$$

The PFS is defined as the Euclidean distance (d) between the memberships of a pixel (2) and the membership of the unclassified pixel (3) divided by a normalization coefficient, which

corresponds to the distance between the unclassified pixel and the centroid pixel (4). This takes into account the maximum distance between any pixel and the unclassified pixel for any given number of clusters and ensures that the PFS always ranges from zero to one.

$$PFS = \frac{d(\mu_k(c),\ \mu_\emptyset(c))}{d(\mu_{cp}(c),\ \mu_\emptyset(c))} = \frac{d(\mu_k(c),\ \mu_\emptyset(c))}{\sqrt{\frac{c-1}{c}}} \qquad (4)$$

3.Results

3.1. Colocalization of m/z features and clusters

Estimating which *m/z* features are more closely related to each ROI is one of the main interests of the spatial segmentation of MSI data, as it allows comparing the molecular signature of the ROIs. In the case of clustering methods, it can be done by comparing the centroids of the cluster, but if the data has received many transformations before clustering, like normalization or scaling operations, it can be complicated to extrapolate the comparisons to the original data due to, for example, negative values and differences in the intensity scale.

An alternative approach is computing Pearson's correlation of the image produced by an *m/z* feature of interest and the binary image of a cluster, where pixels assigned to a cluster have an intensity value of one and all the others of zero. This approach can be reconsidered in terms of soft clustering by comparing *m/z* features to the distribution of the membership to a cluster which we refer to as membership maps. Hard clustering membership maps produce images where only the region assigned to the cluster has intensity values (0 and 1), whereas, in soft clustering membership maps, the information on the transitions between clusters is added in a complete scale of values between 0 and 1. Figure 1 shows the intensity map of ion *m/z* 850.658, the clustering results with four clusters (fuzzifier set to 2 and using the Euclidean distance), and the hard and soft membership maps. We observe that the ion is colocalized with cluster 3, as is the cluster with highest numerical correlation. Using the hard membership map we obtain a correlation of 0.792, and using the soft membership map the value increases up to 0.870.

To see the differences in correlation depending on the FCM parameters, we performed Pearson's correlation between all the images of *m/z* features and the hard and soft membership maps using different fuzzifiers, distance metrics, and numbers of clusters over the sagittal mouse brain. Supplementary figure 2 summarizes the effects of all these parameters on the correlation. In general, using soft membership maps increases the correlation for any number of clusters and distance metrics, but the size of the increase depends on the fuzzifier. Fuzzifiers close to 1 produce membership maps that only keep strongly related pixels in them, like in hard membership maps, and therefore the correlations are similar between the hard and soft membership maps. On the other hand, fuzzifiers with higher values include a representation of the transitions and overlaps between clusters, increasing the correlation between clusters and *m/z* features. In terms of the distance metric, the Euclidean distance experiences a greater increase in the correlation than the cosine distance using soft membership maps. The increases in mean correlation (the increase between the tendency lines in supplementary figure 2) using fuzzifiers of 2, 1.5, 1.25, and 1.1 are, for the cosine distance, 15.93%, 7.22%, 3.6%, 1.5%; and for the Euclidean distance 30.76%, 14.46%, 6.9%, and 2.5% respectively. Finally, the cosine distance achieves higher correlations than the Euclidean distance for clustering results with a small number of clusters, while the opposite happens for the Euclidean distance. In conclusion, soft membership maps are more beneficial in determining colocalization than hard membership

maps as the correlation value between the *m/z* feature image and the colocalized cluster tends to be higher.

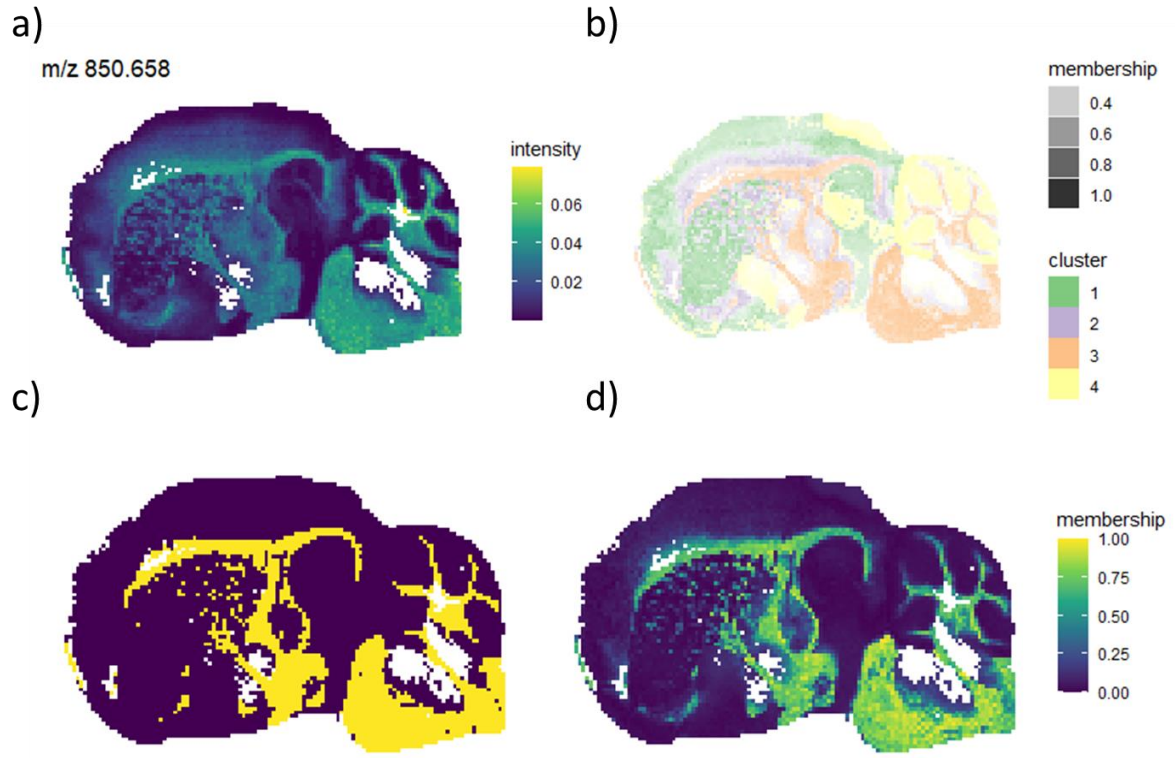


Figure 1. a) Intensity image of ion *m/z* 850.658. b) Clustering result using four clusters, Euclidean distance, and fuzzifier set to 2. The membership is displayed in the plot using the transparency of the colors. c) Hard membership map of cluster 3 consisting only of zeros and ones. d) Soft membership map of cluster 3 with a complete scale between 0 and 1.

3.2. Study of the membership distribution of clusters using the PFS

Using the distribution of the PFS over multiple clustering results, we illustrated the combined effects of the parameters of FCM in the membership distribution of the clusters. Supplementary figure 3 shows the distribution of the PFS with different distance metrics, fuzzifiers, and the number of clusters for the sagittal mouse brain. First, we observe that as the fuzzifier increases, the median PFS tends to decrease for all the other parameters, increasing the spatial overlap between clusters as expected by the definition of the fuzzifier. Second, we observe that as the number of clusters increases, the median PFS tends to decrease and decreases more as higher is the fuzzifier used. And third, the cosine distance tends to form clusters with higher PFS than the Euclidean distance and with less dispersion, except when combining fuzzifier values close to two and more than three clusters. Then, the cosine distance still produces a higher PFS median than the Euclidean distance but higher dispersion as well. These three conclusions summarize the combined behavior of the three parameters over the membership distribution of clustering results, which should be considered in adjusting the FCM to specific workflows.

Additionally, we use the PFS to compare the membership distribution of each cluster in multiple clustering results. Supplementary figure 4 shows different clustering results of the sagittal mouse brain with two to seven clusters using the cosine distance and fuzzifier set to 1.5 as an example. Using this representation, we observe that, with two clusters there is almost no

overlap between the two clusters, which indicates that the ion composition and/or the intensity levels of the ions are very different between the two regions and explains most of the variability of the tissue. With three clusters we observe that the previous cluster 2 remains stationary as the new cluster 1, while the previous cluster 1 is now split between the new clusters 2 and 3, but with lower PFS density, indicating that the overlap between clusters 2 and 3 is greater than between them and cluster 1. For the five-cluster result, we see that one of the clusters represents the same region as cluster 2 in the two-cluster result, while the other clusters overlap each other in different regions as the PFS distribution indicates. As we keep increasing the number of clusters, the overlap between clusters tends to increase in most regions of the tissue. This analysis allows visualizing which clusters occupy regions with more specific ionic composition, and which are more heterogeneous.

Finally, in MSI it is common to compare clusters pairwise to assess which ions are up-regulated or down-regulated in different morphologies. Usually, only a random subset of the total number of pixels is used as the number of pixels per cluster tends to be high, which produces p-values out of the scale. But, as we have shown in membership maps and PFS distributions, not all pixels are equally related to the cluster. Therefore, randomly selecting pixels does not discriminate between pixels. To solve this, we propose to use PFS as a pixel-selection criterion before comparing clusters. For a given clustering result, the pixels with a PFS value lower than a threshold are discarded before comparing groups by random sampling the remaining pixels. For this purpose, using PFS is more convenient than using the raw membership value, as it always ranges between zero and one, while the minimum membership for a pixel to be assigned to a cluster depends on the total number of clusters. For instance, Supplementary figure 5 compares the effects of removing pixels with PFS below 0.95 and of removing randomly the same number of pixels in the sagittal mouse brain with four clusters, cosine distance, fuzzifier set to 1.5, and comparing clusters 2 and 3 in a volcano plot. We observe a general reduction of the significance of the ions for both procedures, which is a natural consequence of the decrease in the number of observations, but using PFS as criteria, the significance of the test is higher than randomly removing pixels. Moreover, using PFS we observe a general increase in the fold change, whereas removing random pixels does not affect it. The same situation is observed in different clustering results and comparing different clusters. This result shows that PFS is useful in selecting pixels strongly related to a cluster, which leads to comparisons between clusters using pixels more representative of the differences between cluster centroids.

3.3. Effects of the Spatial Resolution on Soft Clustering Results

The spatial resolution, usually defined as the lateral length of a pixel, influences the amount of tissue averaged per sampling point in multiple-shoots acquisitions. Therefore, low spatial resolution datasets have a higher chance of pixels with mixed cell types. To study the effects of the spatial resolution on FCM results, we use two datasets of a mouse cerebellum with spatial resolutions of 10 µm and 50 µm. The cerebellum consists of two principal parts, the white matter, and the cerebellar cortex. At the same time, the cerebellar cortex is made up of three layers: the molecular layer, the Purkinje layer, and the granular layer [49]. The transition between several histological regions in the cerebellum is abrupt, making the cerebellum tissue an ideal sample to study the influence of the spatial resolution in the soft clustering, especially in the frontier between regions. We clustered both datasets with four clusters trying to replicate the cytoarchitecture of the tissue, and set the fuzzifier to 2, to promote shared pixels between clusters. We segmented the image using the cosine distance, as it tends to produce a better correlation between clusters and *m/z* features with a low number of clusters. Figure 2 shows the results of both datasets and the PFS distribution of each cluster.

We observe that overall, the four clusters occupy the same regions in both datasets but have slightly different PFS distributions in some clusters. First, cluster 1 (green) is located on the white matter and has the highest PFS values and similar dispersion in both datasets. The high PFS values in both datasets of the white matter indicate that it is a very homogeneous region in terms of lipid composition and intensity levels compared with the other regions. Second, cluster 2 (purple) is in the granular layer and has lower PFS values in the 10 µm dataset than in the 50 µm dataset. This indicates that FCM detects more heterogeneity in the granular layer with increased spatial information compared to the other clusters. This can be caused by the increased specificity in the acquisition of the pixels. Third, clusters 3 (orange) and 4 (yellow), located on the molecular layer and mixed with the pia mater, have slightly different levels of PFS. In the 10 µm dataset, cluster 4 has higher PFS values and less dispersion, while cluster 3 has slightly lower PFS values, compared with the 50 µm dataset. This may be a consequence once more of the increased spatial information, as the specificity of pixels allows a better representation of the frontiers between clusters 3 and 4 in the 10 µm dataset as can be seen in the PFS map of the 10 µm dataset in figure 2.

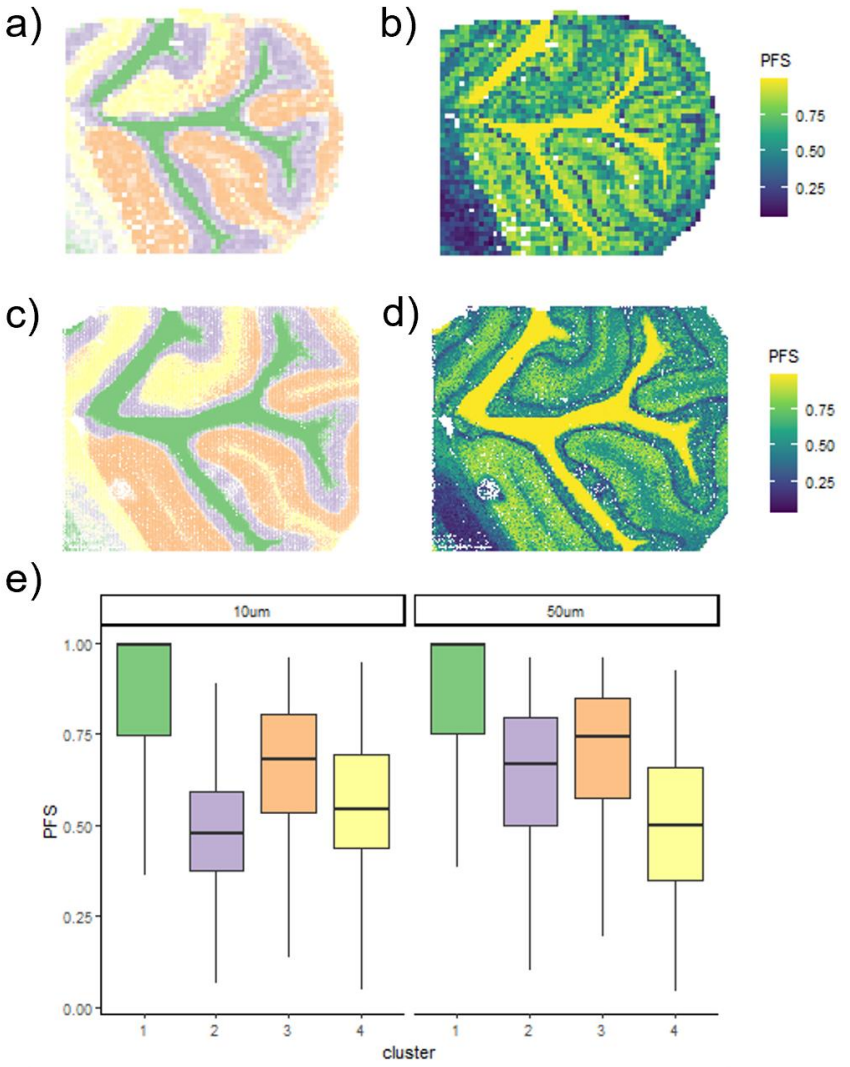


Figure 2. a) Clustering result of the 50 µm dataset. b) PFS map of the 50 µm dataset. c) Clustering result of the 10 µm dataset. d) PFS map of the 10 µm dataset. e) PFS distribution of each cluster using the Euclidean and the cosine distances of the 10 µm and 50 µm datasets.

Regarding the cytoarchitecture of the cerebellum, with four clusters we achieved a clear distinction between the white matter and the granular layer in clusters 1 and 2; and we saw a mix of the molecular layer and the pia mater in clusters 3 and 4. At the same time, FCM was not able to form an independent cluster for the Purkinje layer in any of the datasets. But we can see it indirectly in the PFS maps as a frontier of pixels between clusters 1 and 2 with very low PFS values. This frontier is easier to see in the 10 µm dataset, as the clusters 1-4 are better defined and can clearly be distinguished from these pixels of the Purkinje layer. We tried to increase the number of clusters to detect this specific region, but we could not achieve a good segmentation of it, which may be a consequence of the small number of pixels under this condition compared to other regions in the datasets. Additionally, this frontier is very difficult to see using membership maps, as it is the result of combining the low membership values of all the clusters. All these results point to a general preference for increased spatial resolution, as it helps in assessing heterogeneity in clusters, and shows the utility of PFS maps in discovering small structures hard to cluster, something impossible to achieve with hard clustering methods.

3.4. Semi-Supervised Segmentation Workflow of Head and Neck Cancer Samples

We have developed a semi-supervised soft clustering workflow to study how FCM assigns membership to the transition regions of the head and neck cancer dataset using the distinction between tumoral and healthy tissue as input. The workflow is semi-supervised in two different ways: using histologically defined ROIs and including a feature selection step. The workflow starts by clustering the tumoral samples (S1 and S2) to obtain a cluster that can be associated with the tumoral cells. This is determined manually by comparing the clustering results with ROIs annotated by a histopathologist over a replicate of slide S2, which splits the tumor samples into tumor epithelium and stroma regions. We use 2 clusters to distinguish between these regions in the clustering. After obtaining the cluster, we remove the pixels of the cluster with low PFS, to avoid pixels shared between clusters. From the remaining pixels, we selected the *m/z* features that have more discriminative power between the tumoral pixels in slides S1 and S2, and the samples containing healthy tissue (S5 and S6) using receiver operating characteristic (ROC) curves. For each *m/z* feature, we elaborate the ROC curve and compute the area under the curve (AUC). The *m/z* features that have an AUC higher than 0.9 or lower than 0.1 are used in the clustering analysis of the whole dataset. This procedure ensures that the clustering of all the samples together will focus on studying the spatial distribution of the *m/z* features with more discriminative power between healthy and tumoral pixels and will not follow other morphological structures promoted by other *m/z* features. For the clustering of all the samples, we set the fuzzifier to 2 and used the cosine distance to soften the border between clusters. Starting with the tumoral samples (S1 and S2), figure 3 shows the clustering results and the PFS over samples S1 and S2 using two clusters and the microscopy image of a consecutive tissue sample on slide S2.

We observe that cluster 1 occupies most of the regions annotated as tumor epithelium, while cluster 2 is located in the stroma and other regions. Also, the PFS images reveal a sharp frontier between clusters, which indicates big differences between tissue types and limited overlap. Following, we used the pixels of cluster 1 with PFS higher than 0.7, which is a value close to the median of the cluster, and all the pixels of the healthy samples (S5 and S6) to select the *m/z* feature with more discriminative power using ROC curves. Supplementary figure 6 shows the pixels from cluster 1 used for the ROC curve analysis and the effect that removing them has over a volcano plot, which replicates the results presented over the sagittal mouse brain dataset. With the ROC curve analysis, we selected 27 *m/z* features, which were used in the clustering

of all tissue samples and clustered all the tissue sections with two clusters. Figure 4 shows the clustering result for the whole dataset and the PFS distribution of each cluster at each tissue type.

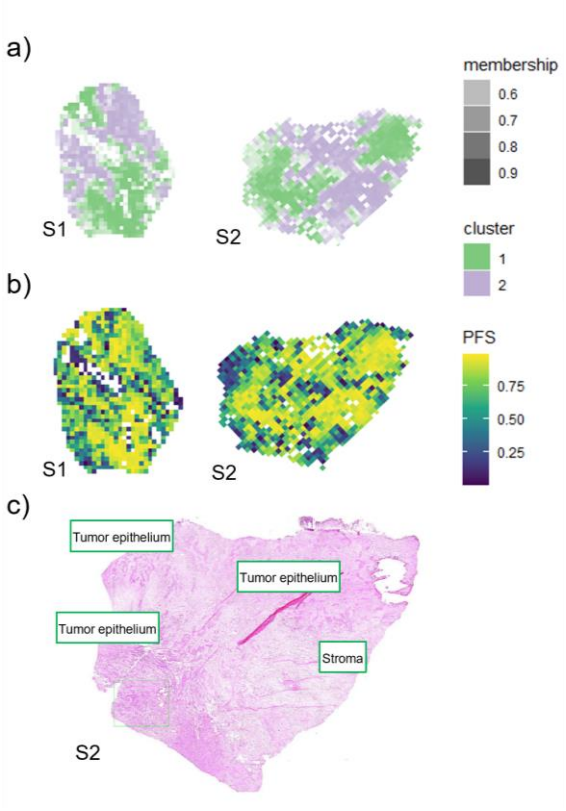


Figure 3. a) Clustering result of samples S1 and S2 using two clusters and fuzzifier set to 2. b) PFS map of samples S1 and S2. c) Optical H&E-stained image of a slide consecutive to S2.

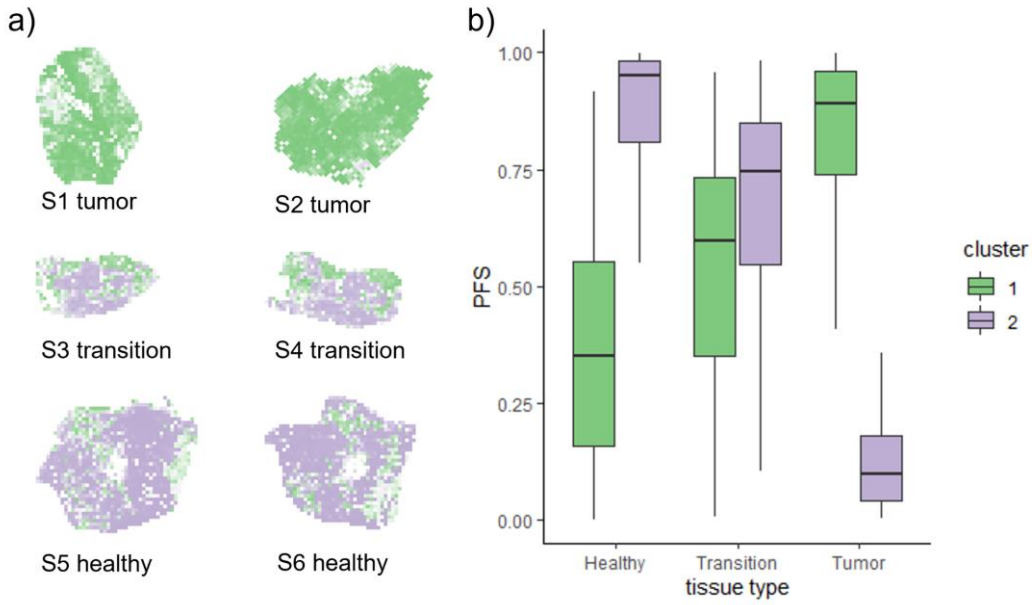


Figure 4. a) Clustering result of all the tissues using two clusters, the cosine distance and fuzzifier set to 2. b) PFS distribution per cluster and per tissue type of the clustering results of all the samples.

We see that cluster 1 is located mostly in the tumor samples and cluster 2 in the healthy samples, both with high values of PFS. The transition samples have pixels from both clusters with similar levels of PFS while the other four samples have very low values of PFS from the opposite cluster. This indicates that pixels in the transition samples are more heterogeneous compared to the tumor and healthy samples, resulting in a bigger overlap between clusters in the region. Also, in the healthy samples, we observe bigger PFS dispersion in cluster 1 than in cluster 2, which may be related to clustering together the tumor epithelium and the stroma in samples S1 and S2. This can be seen in the membership of cluster 1 over the tumor samples, where most of the regions with lower membership were annotated as stroma by the histopathologist. Finally, the workflow has successfully focused on the differences between tumor and healthy samples, and indicated regions closely related to the tumor in the transition samples. This is clearly exposed in the progression of PFS in both clusters across the different sample types and cancer progression.

4. Discussion

In this work, we have applied FCM to MSI datasets and studied the effects of the parameters in different applications. One of the main results computed by FCM is the membership matrix, in which rows represent pixels of the MSI experiment, columns represent the clusters, and the values represent the membership of a pixel to one of the clusters. In MSI, these values can be plotted over the tissue morphology to visualize which regions are more related to one cluster, increasing the interpretability of the results. Moreover, an enhanced clustering image can be obtained by plotting in colors the cluster with the highest membership for each pixel and including the membership as the transparency of the image. Using these images, it is possible to localize, in the morphology of the tissue, the pixels with a spectrum closer to the cluster centroid. Additionally, FCM enables the assessment of spatial overlap between clusters by correlating the membership maps of each cluster, as the membership matrix can be understood as the spatial centroids of the cluster. We have used this idea to enhance the colocalization of *m/z* features with membership maps and obtained higher correlations than using hard clustering masks, as hard clustering masks only indicate where the cluster is predominant. This increase in correlation is important as it allows a safer automatic discard of low correlation *m/z* features. Moreover, it indicates that soft membership maps better represent the structures behind the combination of the *m/z* features over the tissue compared to hard membership maps.

Regarding the combination of the fuzzifier, the distance metrics, and the number of clusters, we have observed a general decreasing tendency of the membership of the pixels when increasing the fuzzifier and the number of clusters. Additionally, the cosine distance formed clusters with higher membership than the Euclidean distance. But we think that there is no standard choice of parameters and, like most tools in MSI, requires the expertise and hypothesis of the data analyst to select the appropriate parameters in each experiment. For instance, the choice of fuzzifier influences the degree of overlap between clusters, and high levels of overlap might be of interest in studies related to samples with mixed tissue types or conditions like tumor margin classification.

Nevertheless, to study the combined effects of the parameters we have developed the PFS as a tool for the detection of regions where different clusters are expressed. We have used the PFS to compare the membership distribution between clusters, filter out pixels with low PFS values before comparing clusters to improve random sampling processes, and detect morphological structures not included in any of the clusters. It is also very important to notice that low PFS pixels do not always mean that the pixels contain a combination of different tissue types. It can also indicate that the pixels have a different molecular signature compared to the other clusters and therefore, they need an independent cluster. Sometimes the algorithm is not

capable of clustering these pixels alone, because their condition represents a small percentage of the total variance of the dataset or because there are a small number of pixels under the condition compared to the whole dataset. In these cases, increasing the number of clusters may be adverse, as it can result in splitting regions into too many clusters. We think this is the case in the mouse cerebellum dataset, where we showed that the Purkinje layer did not form an independent cluster but, thanks to the increased spatial resolution, we detected a low PFS area between the granular and the molecular layer, which can be associated with the Purkinje layer. Although the cluster localization was approximately the same between different spatial resolutions, the possibility of clearly distinguishing regions where the clustering misclassified pixels adds value in interpreting the results. Therefore, we showed that increasing spatial resolution is beneficial for the performance of clustering algorithms.

Finally, the combination of histologically defined ROIs and FCM in a two-step workflow allowed us to better interpret the results of the head and neck cancer dataset, in understanding the role of the transition samples. Thanks to this, we could select the specific *m/z* features that contain more information on the distinction between tumor and healthy samples and expand the clustering criterion over the other samples, leading to the discovery of mixed regions on the transition samples using the PFS distribution. We think that this is the main benefit of soft clustering methods over hard methods, as knowing the degree of overlap between clusters at different tissue samples allows us to classify the results using two criteria: cluster assignment and membership/PFS distribution.

5. Conclusion

FCM brings additional information to MSI data analysis through the dimension of membership, which allows for new ways of interpreting the results compared with hard clustering results. In our case, the study of membership through the newly developed PFS allowed easy selection of the pixels more related to a cluster, unveiled morphological regions more challenging to detect, and enhanced a tissue type classification workflow in multiple samples of a human head and neck cancer dataset.

We expect this work to contribute to attracting interest in soft clustering methods for the spatial segmentation of MSI data. Future lines of research could be study the use of bisecting k-means guided by the membership of the clusters; developing a notion of membership more intrinsically related to how biologic features spatially overlap each other in MSI; distinguishing the low PFS pixels that are a mixture of tissues from pixels that are not well classified; and trying new soft clustering algorithms used in other fields of image processing.

6. References

[1] K. Chughtai and R. M. A. Heeren, “Mass spectrometric imaging for biomedical tissue analysis,” Chem. Rev., vol. 110, no. 5, pp. 3237–3277, May 2010.

[2] J. L. Norris and R. M. Caprioli, “Analysis of tissue specimens by matrix-assisted laser desorption/ionization imaging mass spectrometry in biological and clinical research,” Chem. Rev., vol. 113, no. 4, pp. 2309–2342, Apr. 2013.

[3] R. Van de Plas, J. Yang, J. Spraggins, and R. M. Caprioli, “Image fusion of mass spectrometry and microscopy: a multimodality paradigm for molecular tissue mapping,” Nat. Methods, vol. 12, no. 4, pp. 366–372, Apr. 2015.

[4] A. Römpp et al., “Histology by mass spectrometry: label-free tissue characterization obtained from high-accuracy bioanalytical imaging,” Angew. Chem. Int. Ed Engl., vol. 49, no. 22, pp. 3834–3838, May 2010.

[5] P. M. Angel, J. M. Spraggins, H. S. Baldwin, and R. Caprioli, "Enhanced sensitivity for high spatial resolution lipid analysis by negative ion mode matrix assisted laser desorption ionization imaging mass spectrometry," Anal. Chem., vol. 84, no. 3, pp. 1557–1564, Feb. 2012.

[6] A. R. Buchberger, K. DeLaney, J. Johnson, and L. Li, "Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights," Anal. Chem., vol. 90, no. 1, pp. 240–265, Jan. 2018.

[7] G. Arentz et al., "Applications of Mass Spectrometry Imaging to Cancer," Adv. Cancer Res., vol. 134, pp. 27–66, Jan. 2017.

[8] G. Mrukwa, G. Drazek, M. Pietrowska, P. Widlak, and J. Polanska, "A novel divisive iK-means algorithm with region-driven feature selection as a tool for automated detection of tumour heterogeneity in MALDI IMS experiments," in Bioinformatics and Biomedical Engineering, Cham: Springer International Publishing, 2016, pp. 113–124.

[9] T. Alexandrov and J. H. Kobarg, "Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering," Bioinformatics, vol. 27, no. 13, pp. i230–8, Jul. 2011.

[10] S.-O. Deininger, M. Becker, and D. Suckau, "Tutorial: multivariate statistical treatment of imaging data for clinical biomarker discovery," Methods Mol. Biol., vol. 656, pp. 385–403, 2010.

[11] D. Bonnel et al., "Multivariate analyses for biomarkers hunting and validation through on-tissue bottom-up or in-source decay in MALDI-MSI: application to prostate cancer," Anal. Bioanal. Chem., vol. 401, no. 1, pp. 149–165, Jul. 2011.

[12] S.-O. Deininger, M. P. Ebert, A. Fütterer, M. Gerhard, and C. Röcken, "MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers," J. Proteome Res., vol. 7, no. 12, pp. 5230–5236, Dec. 2008.

[13] P. Ràfols et al., "Signal preprocessing, multivariate analysis and software tools for MA(LDI)-TOF mass spectrometry imaging for biological applications," Mass Spectrom. Rev., vol. 37, no. 3, pp. 281–306, May 2018.

[14] T. Hiratsuka et al., "Hierarchical Cluster and Region of Interest Analyses Based on Mass Spectrometry Imaging of Human Brain Tumours," Sci. Rep., vol. 10, no. 1, p. 5757, Apr. 2020.

[15] E. Le Rhun et al., "Evaluation of non-supervised MALDI mass spectrometry imaging combined with microproteomics for glioma grade III classification," Biochim. Biophys. Acta: Proteins Proteomics, vol. 1865, no. 7, pp. 875–890, Jul. 2017.

[16] S. M. Willems, A. van Remoortere, R. van Zeijl, A. M. Deelder, L. A. McDonnell, and P. C. W. Hogendoorn, "Imaging mass spectrometry of myxoid sarcomas identifies proteins and lipids specific to tumour type and grade, and reveals biochemical intratumour heterogeneity," J. Pathol., vol. 222, no. 4, pp. 400–409, Dec. 2010.

[17] J. G. Swales et al., "Quantitation of Endogenous Metabolites in Mouse Tumors Using Mass-Spectrometry Imaging," Anal. Chem., vol. 90, no. 10, pp. 6051–6058, May 2018.

[18] H. Bednarz, N. Roloff, and K. Niehaus, "Mass Spectrometry Imaging of the Spatial and Temporal Localization of Alkaloids in Nightshades," J. Agric. Food Chem., vol. 67, no. 49, pp. 13470–13477, Dec. 2019.

[19] D. Veličković, H. Herdier, G. Philippe, D. Marion, H. Rogniaux, and B. Bakan, "Matrix-assisted laser desorption/ionization mass spectrometry imaging: a powerful tool for probing the molecular topology of plant cutin polymer," Plant J., vol. 80, no. 5, pp. 926–935, Dec. 2014.

[20] B. A. Boughton, D. Thinagaran, D. Sarabia, A. Bacic, and U. Roessner, "Mass spectrometry imaging for plant biology: a review," Phytochem. Rev., vol. 15, pp. 445–488, 2016.

[21] K. D. Bemis et al., "Cardinal: an R package for statistical analysis of mass spectrometry-based imaging experiments," Bioinformatics, vol. 31, no. 14, pp. 2418–2420, Jul. 2015.

[22] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," Proc. Natl. Acad. Sci. U. S. A., vol. 99, no. 10, pp. 6567–6572, May 2002.

[23] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays," Statistical Science, vol. 18, no. 1. 2003. doi: 10.1214/ss/1056397488.

[24] K. D. Bemis et al., "Probabilistic Segmentation of Mass Spectrometry (MS) Images Helps Select Important Ions and Characterize Confidence in the Resulting Segments," Mol. Cell. Proteomics, vol. 15, no. 5, pp. 1761–1772, May 2016.

[25] I. Chernyavsky, T. Alexandrov, P. Maass, and S. I. Nikolenko, "A two-step soft segmentation procedure for MALDI imaging mass spectrometry data," 2012. doi: 10.4230/OASIcs.GCB.2012.39.

[26] D. Guo, K. Bemis, C. Rawlins, J. Agar, and O. Vitek, "Unsupervised segmentation of mass spectrometric ion images characterizes morphology of tissues," Bioinformatics, vol. 35, no. 14, pp. i208–i217, Jul. 2019.

[27] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," J. cybern., vol. 3, no. 3, pp. 32–57, Jan. 1973.

[28] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms." 1981. doi: 10.1007/978-1-4757-0450-1.

[29] M. Falasconi, A. Gutierrez-Galvez, M. Leon, B. A. Johnson, and S. Marco, "Cluster analysis of rat olfactory bulb responses to diverse odorants," Chem. Senses, vol. 37, no. 7, pp. 639–653, Sep. 2012.

[30] V. Schwämmle and O. N. Jensen, "A simple and fast method to determine the parameters for fuzzy c-means cluster analysis," Bioinformatics, vol. 26, no. 22, pp. 2841–2848, Nov. 2010.

[31] M. Huang, Z. Xia, H. Wang, Q. Zeng, and Q. Wang, "The range of the value for the fuzzifier of the fuzzy c-means algorithm," Pattern Recognit. Lett., vol. 33, no. 16, pp. 2280–2284, Dec. 2012.

[32] K. Zhou and S. Yang, "Fuzzifier selection in fuzzy C-means from cluster size distribution perspective," Informatica , vol. 30, no. 3, pp. 613–628, Jan. 2019.

[33] N. R. Pal and J. C. Bezdek, "On cluster validity for the fuzzy c-means model," IEEE Transactions on Fuzzy Systems, vol. 3, no. 3. pp. 370–379, 1995. doi: 10.1109/91.413225.

[34] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler, Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. John Wiley & Sons, 1999.

[35] E. A. Jones et al., "Multiple statistical analysis techniques corroborate intratumor heterogeneity in imaging mass spectrometry datasets of myxofibrosarcoma," PLoS One, vol. 6, no. 9, p. e24913, Sep. 2011.

[36] S. Sarkari, C. D. Kaddi, R. V. Bennett, F. M. Fernandez, and M. D. Wang, "Comparison of clustering pipelines for the analysis of mass spectrometry imaging data," Conf. Proc. IEEE Eng. Med. Biol. Soc., vol. 2014, pp. 4771–4774, 2014.

[37] T. Alexandrov, I. Chernyavsky, M. Becker, F. von Eggeling, and S. Nikolenko, "Analysis and interpretation of imaging mass spectrometry data by clustering mass-to-charge images according to their spatial similarity," Anal. Chem., vol. 85, no. 23, pp. 11189–11195, Dec. 2013.

[38] B. Balluff et al., "Integrative Clustering in Mass Spectrometry Imaging for Enhanced Patient Stratification," Proteomics Clin. Appl., vol. 13, no. 1, p. e1800137, Jan. 2019.

[39] E. Del Castillo et al., "rMSIKeyIon: An Ion Filtering R Package for Untargeted Analysis of Metabolomic LDI-MS Images," Metabolites, vol. 9, no. 8, Aug. 2019, doi: 10.3390/metabo9080162.

[40] P. Ràfols et al., "Assessing the potential of sputtered gold nanolayers in mass spectrometry imaging for metabolomics applications," PLoS One, vol. 13, no. 12, p. e0208908, Dec. 2018.

[41] A. Bednařík, S. Bölsker, J. Soltwisch, and K. Dreisewerd, "An on-tissue paternò-Büchi reaction for localization of carbon-carbon double bonds in phospholipids and glycolipids by matrix-assisted laser-desorption-ionization mass-spectrometry imaging," Angew. Chem. Weinheim Bergstr. Ger., vol. 130, no. 37, pp. 12268–12272, Sep. 2018.

[42] P. Ràfols et al., "rMSIproc: an R package for mass spectrometry imaging data processing," Bioinformatics, vol. 36, no. 11, pp. 3618–3619, Jun. 2020.

[43] L. Sementé, G. Baquer, M. García-Altares, X. Correig-Blanchar, and P. Ràfols, "rMSIannotation: A peak annotation tool for mass spectrometry imaging based on the analysis of isotopic intensity ratios," Analytica Chimica Acta, vol. 1171. p. 338669, 2021. doi: 10.1016/j.aca.2021.338669.

[44] S.-O. Deininger et al., "Normalization in MALDI-TOF imaging datasets of proteins: practical considerations," Anal. Bioanal. Chem., vol. 401, no. 1, pp. 167–181, Jul. 2011.

[45] D. Eddelbuettel, Seamless R and C++ Integration with Rcpp. Springer Science & Business Media, 2013.

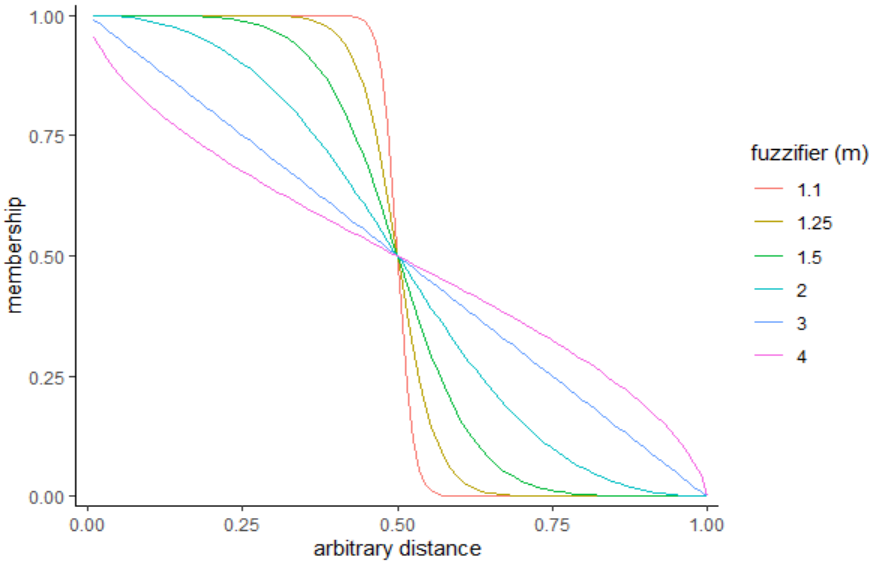
[46] Z. Cebeci, "Comparison of internal validity indices for fuzzy clustering," J. Agric. Inform., vol. 10, no. 2, Dec. 2019, doi: 10.17700/jai.2019.10.2.537.

[47] M. Ferraro, Maria, B. Ferraro, P. Giordani, and A. Serafini, "fclust: An R Package for Fuzzy Clustering," The R Journal, vol. 11, no. 1. p. 198, 2019. doi: 10.32614/rj-2019-017.

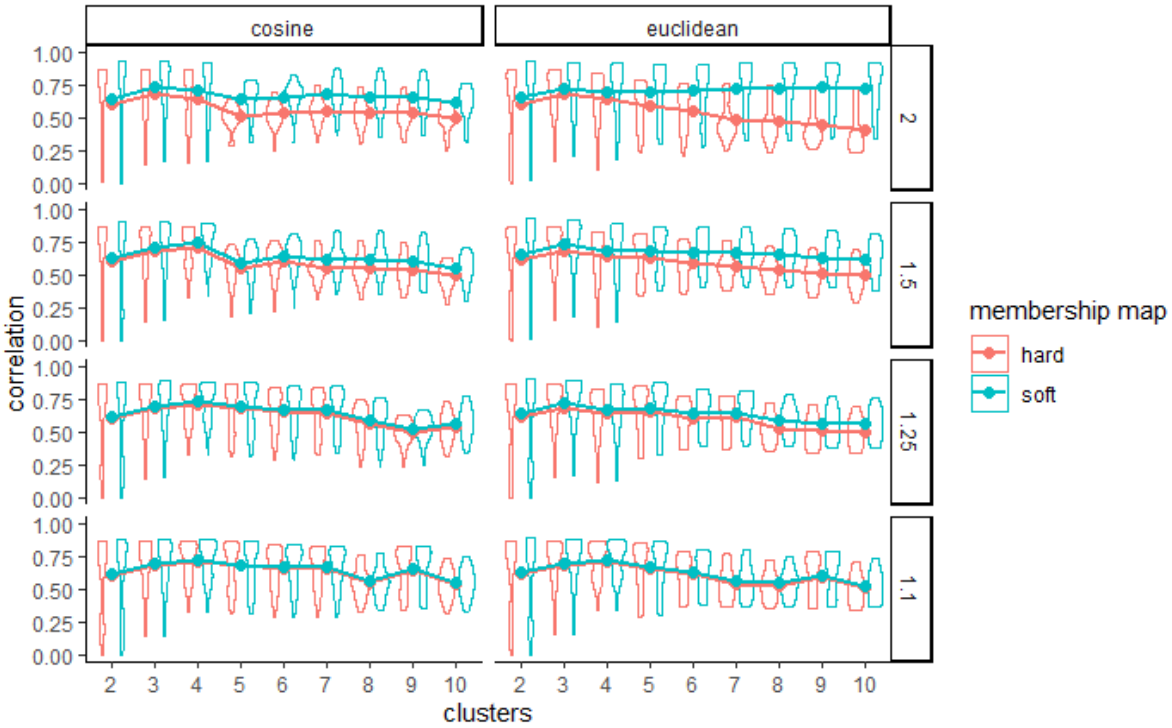
[48] T. Smets et al., "Evaluation of Distance Metrics and Spatial Autocorrelation in Uniform Manifold Approximation and Projection Applied to Mass Spectrometry Imaging Data," Anal. Chem., vol. 91, no. 9, pp. 5706–5714, May 2019.

[49] D. Rokni, R. Llinas, and Y. Yarom, "The Morpho/Functional Discrepancy in the Cerebellar Cortex: Looks Alone are Deceptive," Front. Neurosci., vol. 2, no. 2, pp. 192–198, Dec. 2008.

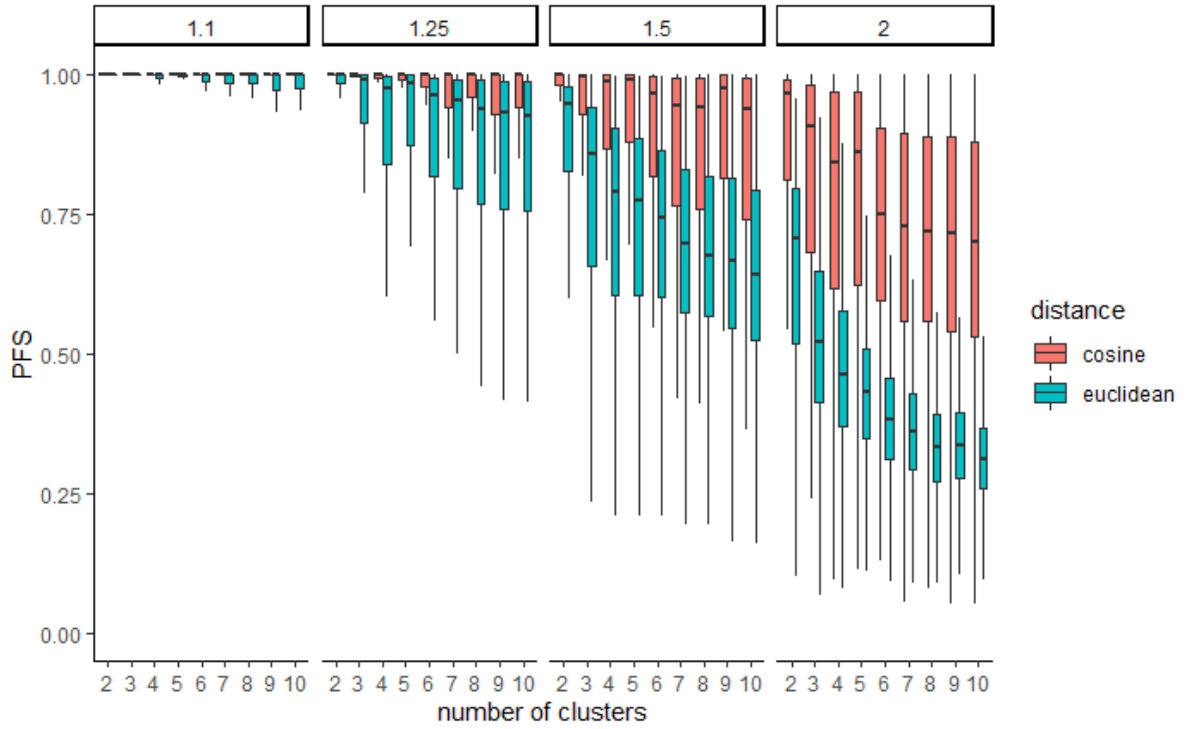
7. Supplementary figures



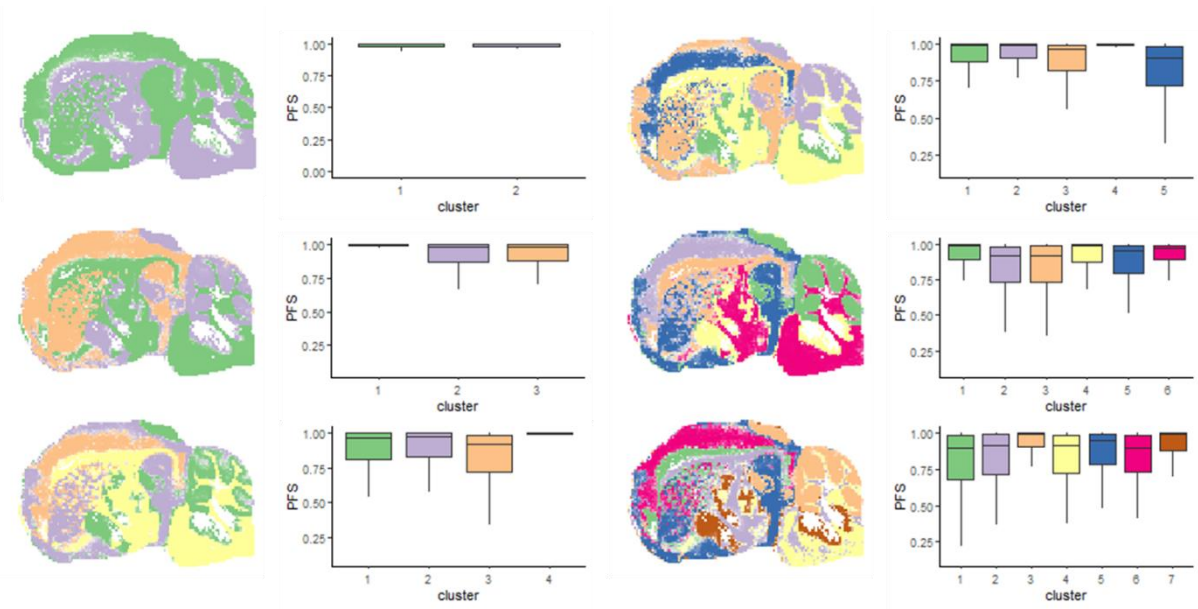
Supplementary figure 1. Membership curve of FCM using different fuzzifiers. The y-axis represents the membership of a pixel to a cluster. The x-axis represents, between 0 and 1, the dynamic range of the multidimensional distance between the pixel and cluster centroid; being 0 the closest and 1 the farthest. Fuzzifier values close to 1 create sharp borders between clusters, quickly changing between high membership and low membership, while values higher than 2 soften the borders between clusters.



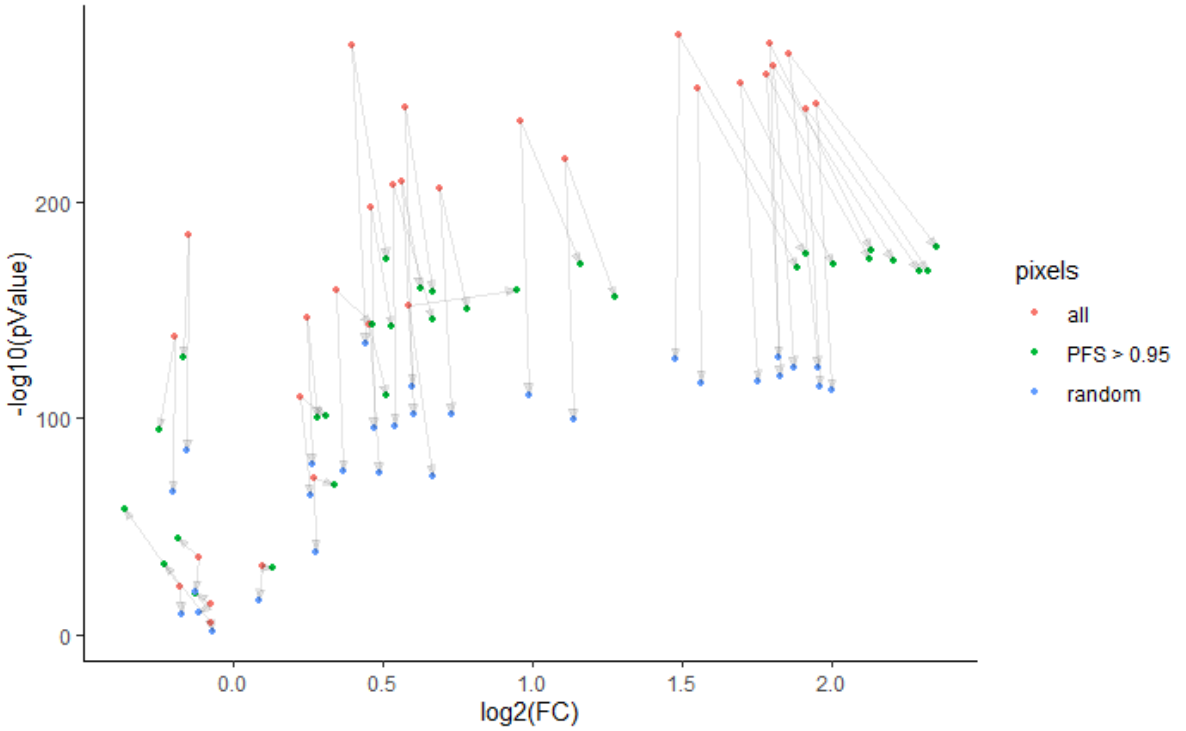
Supplementary figure 2. Colocalization between *m/z* features and cluster membership map images under the influence of the fuzzifier (rows), distance metric (columns), type of membership map, and the number of clusters. The tendency lines indicate the mean correlation for each case.



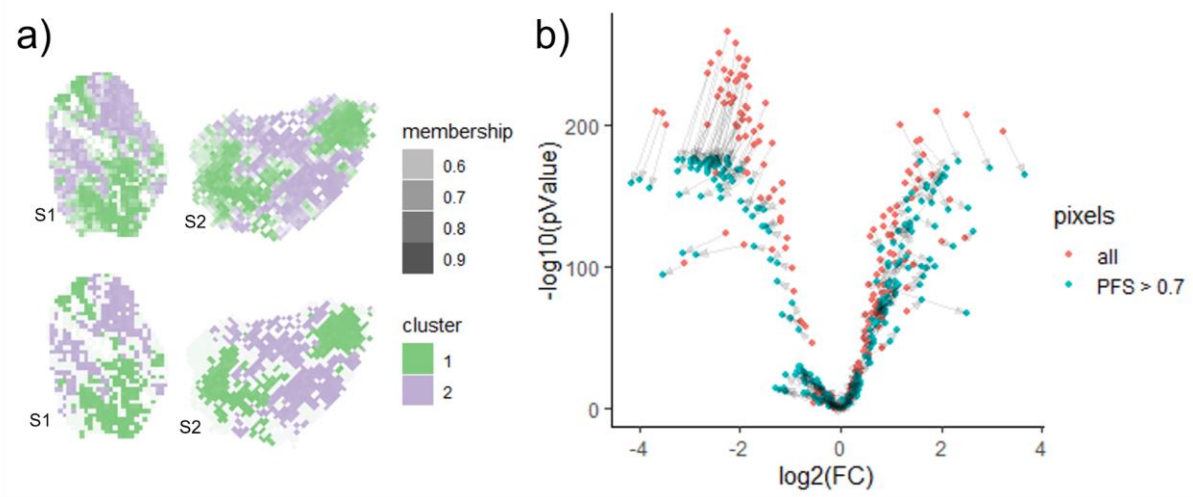
Supplementary figure 3. Boxplot of the effects of the fuzzifier (columns), distance metric, and the number of clusters over the PFS. The distribution represents the PFS of all the pixels of the image at each subset of parameters.



Supplementary figure 4. Morphological representation of six clustering results of the mouse cerebellum and the PFS distribution of the clusters at each clustering result.



Supplementary figure 5. Volcano plot of the sagittal mouse brain comparing clusters two and three from the four-cluster result using the cosine distance and fuzzifier set to 1.5. The color scale indicates which pixels are used to compare clusters. The total number of pixels removed in both procedures is the same. The arrows indicate how the fold changes and the significance of an m/z feature change after removing pixels using each method.



Supplementary figure 6. a) Clustering result of samples S1 and S2 with all the pixels (top) and with only pixels with a PFS greater than 0.7. b) Volcano plot of samples S1 and S2 before and after removing pixels with PFS below 0.7. We observe a general increase in the absolute value of the FC and a reduction in statistical significance like in the sagittal mouse brain.

CHAPTER 6

Final discussion and conclusions

The work included in this doctoral thesis can be clearly separated in two parts.

Firstly, the development of an isotope and adduct peak annotation tool suited to facilitate the identification of the low mass range compounds. We can now easily find monoisotopic ions in our MSI datasets thanks to the rMSIannotation software package.

Secondly, the development of software tools for data analysis and spatial segmentation based on soft clustering for MSI data. In this thesis, we have developed tools and methodologies to search for significant ions (rMSIKeyIon software package) and for the soft clustering of tissues (Fuzzy c-means algorithm).

We believe that the goals proposed in this thesis have been successfully accomplished since we have developed and validated tools that attempt to solve part of these problems. Still, multiple challenges remain open and future lines of research can be proposed.

1. Peak annotation: a necessary step for ion identification of MSI data

The relationship between ions and their localization over a tissue section allow the discovery of spatial features on the sample but, without molecular identity, the work remains incomplete and can only be considered as cytoarchitecture, not spatial metabolomics. Therefore, the identification of metabolites is the most important step in any spatial metabolomics study. However, the identification of metabolites in MSI has multiple challenges to overcome in terms of spectral acquisition and bioinformatic strategies.

Starting with the spectral acquisition, the principal drawback of MSI for the identification of metabolites is the simultaneous ionization of all the compounds at each sampling point, which produces a chain of negative effects. It starts with the simultaneous in-source fragmentation of many compounds, complicating the connection between parental ions and fragments, and producing a superpopulation of ions in the low mass range (metabolites). The superpopulation of fragment ions makes the identification of the metabolites difficult, especially in the low mass range. Comparing the situation with LC-MS spectra, in which the signals are associated with a *m/z* and retention time, the confusion between fragments and parental ions in MSI is harder to solve. But, even using LC-MS, for untargeted analysis, *da Silva et al.* estimated that only 1,8% of the spectra can be annotated[1], associating the term “dark metabolome” to refer to all the unknown signals.

Ion superpopulation in MALDI-MSI also happens due to the presence of matrix ions in the spectra, as organic matrices (the most widely used) have high fragmentation and adduct formation rate.[2,3] The existence of this so high number of ions results in many of them sharing very close *m/z* values. Even using high resolution mass analyzers most of them are hard to split in different peaks resulting in overlapping peaks. Moreover, this not only happens with fragments and some small metabolites but can also occur with parental ions of different molecules. This problem can be addressed in some extend depositing inorganic matrices, like gold nanoparticles.[4,5] Most inorganic matrices produce cleaner spectra than organic matrices and have carbon-free isotopic patterns, which can be used to easily identify their signal background, but tend to enhance the fragmentation of analytes in the tissue.

Overlapping peaks is a challenging problem specially in MALDI-TOF datasets for spatial metabolomics, as the mass resolution of most TOF detectors is not high enough in the low mass range of the spectra to resolve them. Additionally, overlapping peaks usually pass without notice leading to erroneous annotations by distorting the real *m/z* and intensity of the compound. During the development and validation of the rMSIannotation software for isotope and adduct annotation we found that many of the confidently identified compounds in the MALDI-TOF dataset, suffer from overlap. For instance, we found overlap in the isotopic patterns of the sodium adduct of phosphatidylcholine (32a:1), with molecular formula $[C_{40}H_{78}NO_8P+Na]^+$ and *m/z* 754.535; and the potassium ion of sphingomyelin (d18:1/C17:0)

with molecular formula $[C_{40}H_{81}N_2O_6P+K]^+$ and *m/z* 755.546. Their monoisotopic and first isotope ions have differences in *m/z* of 0.0073 Da (9 ppm), which could not be resolved by the used TOF analyzer[6]. Using this example we studied how overlapping affected our algorithm using *in silico* patterns at different mass resolutions and abundance ratios between ions. The conclusions of the study determined that rMSIannotation can annotate some pairs of overlapped compounds depending on the abundance ratio of the compounds and the mass resolution of the peaks. As far as we know, this is the first time a peak annotation tool has been tested over overlapped peaks and proven to annotate them. We did not check the results for more than two overlapping ions, but we assume that the situation gets more challenging in this case. Additionally, in MALDI-TOF datasets, this same case of overlap in even smaller ions would result in lower annotation possibilities as the mass resolution decreases. These two facts, in-source fragmentation, and peak overlap between others, like mass accuracy, are important reasons for changing MALDI-TOF instruments for MALDI-Orbitrap or MALDI-FT-ICR instruments in spatial metabolomics experiments, as they suffer less their bad effects.

A line of research attempting to solve this problem could be the spectral deconvolution of the overlapped peaks based on lineshape fitting algorithms, a technique often used in NMR spectra.[7] The mission of these algorithms in MSI could be first, assess the mass resolution profile of the experiment (the variation of mass resolution over the *m/z* axis), later, localize mass ranges where the spectra present abnormally wide peaks according to the expected mass resolution, and finally, try to solve the fitting problem of different peak shapes in the mass range in conflict. In some spatial regions, some of the overlapped peaks may appear isolated in the spectra, which could help in deconvoluting the spectra. Some works proposed an alternative approach to detect overlapping peaks by decomposing the whole MSI spectra into Gaussian Mixture Models (GMM).[8,9] The main drawback of this method is setting the initial conditions of the algorithm for fitting such a big number of components, and the apparition of non-spectral peaks due to artifacts of the algorithm, which require special processing to detect and discard all of them. Finally, we purpose a new approach based on the peak annotation scores provided by the rMSIannotation algorithm. The scores contain information on why a peak has not been assigned to an isotopic pattern and this information can be used to detect overlapped peaks by, for instance, comparing the correlations and the expected isotopic intensity ratios of an annotated isotopic pattern to different adduct distances. This information could be used to initiate a second iteration of data processing targeted on the spectral regions where the peak annotation algorithm has scored low with parameters aiming to split overlapped peaks. We consider this an interesting line of research to follow.

Metabolite annotation and identification is also a challenge in terms of bioinformatic strategies. The challenge originates from the kind of information produced in an MSI experiment, which consist basically of *m/z* values (exact mass) with intensities localized over a two-dimensional surface. Therefore, compared with other MS methods like LC-MS, we have a very limited source of information regarding ion identity. There are some works trying to expand the *m/z* information combining MS scans with MS/MS scans over the same tissue sample but with the limitations of the available tissue material and the low concentration of precursor.[10] On the other hand, MSI benefits from a huge number of sampling points, which is one of the most valuable features of MSI experiments compared to other MS methods that enhance multiple tools. For instance, image correlation is a common procedure in most peak annotation tools for MSI that benefits from having more observations. During the development of rMSIannotation we recognized this opportunity and used it in multiple ways. We used it to correlate isotope ion images, as their spatial distribution must be almost the same, and to precisely determine the intensity ratio between monoisotopic and isotopic ions, extracted using linear models and used to determine the number of carbon atoms of the compounds, which as far as we know, we are the first to implement. The annotation of isotope ions benefits more of

these strategies compared with the annotation of adduct ions. First, the image correlation between adducts is not as strong as in the case of isotopes because is affected by the natural abundance distribution of elements in the tissues (i.e. K^+, Na^+, etc.)[11] and by the homogeneity of the matrix application[3]. And second, there are no rules regarding the intensity ratios between monoisotopic ions with different adducts that need to be tested.

The gold standard for peak annotation algorithms in MSI metabolomics datasets is the METASPACE[12] platform, based on fitting spectral library models. High resolution spectral data with high mass accuracy is required by METASPACE to achieve confident annotation to minimize the chance of overlapped peaks and minimize the mass errors. We opted for a radically different approach, consisting of annotating the isotopes and the adducts with only the acquired spectra. Once the monoisotopic peaks have been determined, the molecular annotation is done by searching externally in exact mass libraries. The principal novelty of our approach was the modeling of the HMDB[13] into one equation which relates the *m/z* of a monoisotopic ion with a range of common isotopic intensity ratios for the metabolites close to the *m/z*. This, together with the possibility of extracting very precisely the intensity ratios between isotopes using the huge number of pixels allow for a fast annotation of MSI datasets in the format of peak matrix. Developing such mathematical models could be a completely new line of research. Extending the modeling to more families of compounds and subgroups would result in a collection of multiple equations working together. With the equations and some heuristics (for instance, only allowing to test models that cover the mass range of the *m/z* of interest), the algorithm would compute multiple scores and reduce the number of candidate annotations. However, even with all the possible bioinformatic efforts, measurements of the same sample with an orthogonal technique (like MS/MS or ion mobility) are required to validate a molecular annotation.[14]

2. Ion selection strategies in MSI

The development of this thesis has covered many aspects of the untargeted analysis of MSI data. One important topic has been finding, between the huge number of ions, those that contribute more to the answers of a MSI study. For that, we have used two different strategies based on different criteria. The first, is a method to select the ions with a higher contrast between predefined regions using statistical criteria, and the second, is a method to detect the monoisotopic ions in the spectra using chemometric criteria. Both approaches are meant to facilitate the discovery of relevant molecular signatures by reducing the number of variables under study and therefore, are expected to be combined.

The first strategy, implemented through rMSIKeyIon, selects the ions with the highest contrast between predefined regions evaluating statistical significance. The main drawback of this method is MSI data present multiple incompatibilities with statistical hypothesis testing overall. First, most intensities of the ions do not follow a normal distribution, requiring data transformations or the use of non-parametric tests like rank sum tests. Additionally, using the enormous amount of data points in most MSI datasets produces extremely low p-values. Therefore, the use of random sampling before applying tests is recommended. But the difference between the spectra of pixels of the same region can be very important, as multiple ion distributions overlap each other, resulting in tests with less statistical power if the pixels of a region are not chosen carefully. During the development of rMSIKeyIon we noticed this problem and included a metric to account for it consisting in the percentage of pixels an ion does not appear in a ROI. After identifying the empty pixels of an ion in the ROI, they are removed before computing p-values and FC to not bias the real difference between regions. Clustering algorithms like k-means tend to produce this type of ROIs (clusters in terms of clustering algorithms) and is one of the main reasons to pursue workflows including soft

clustering methods. Moreover, statistical hypothesis tests require p-value adjustment for simultaneous testing, being this more restrictive as bigger is the number of ions under investigation. Finally, MSI data suffers from spatial autocorrelation, which violates the assumption of independence between pixels.[15] The pixel autocorrelation appears due to the interconnected nature of tissue morphologies and due to experimental artifacts (the acquisition of a pixels affects the acquisition of its neighborhood). For instance, in MALDI-MSI is required the application of a matrix, which can delocalize some compounds and promote their ionization over regions where they were not naturally present. Moreover, laser oversampling during the acquisition transfers information from one pixel to their vicinity by having an ablation diameter bigger than the pixel size. Conditional Autoregressive Model (CAR) has been proposed to account for spatial autocorrelation, decreasing the number of significant features after applying it. However, the main drawbacks of this method are the need for extensive computational resources and determining the range of the neighborhood of a pixel.[15]

Discarding ions should not be of major concern in MSI, as experiments consist of a huge number of redundant ions which difficult the elaboration of conclusions due to the 'curse of dimensionality', responsible of reducing the accuracy of various statistical methods like distance metrics in high-dimensional spaces.[16] The second strategy, implemented through rMSIannotation, uses chemometric information to attempt to annotate as many ions in isotopic patterns, which are the principal source of redundancy. As all the ions in an isotopic pattern express the same information (unless the experiment includes isotope labeling or measures precise isotope fluctuations), the most intense peak should be the only included in the analysis. we have shown in chapter 4 that monoisotopic ions produce almost identical results by spatial segmentation and component analysis as with including all the other ions, indicating that most of the tissue morphology is retained by only a few ions and that a huge number of ions contain redundant information for the data analysis. This is not a problem in metabolomics, as the most intense peak corresponds to the monoisotopic, but with other molecules with more than 94 carbon atoms, like peptides this is not the case, and the peak selected should be the highest in the isotopic pattern. This could be addressed using the relationship between carbon atoms of a molecule and the *m/z* value of it.

The study of isotopic patterns also leads to the fact that the number of peaks per molecule is not the same for all molecules. The number of isotopic peaks of a molecule in a spectrum increases with the *m/z* of the monoisotopic ion. Therefore, the morphology of molecules with more peaks will have more votes on upcoming multivariate procedures, with the possibility of hiding some morphological structures due to peak underrepresentation. A line of research to attempt to solve this problem could be transforming the ion space into the molecule space, where all the ions coming from the same molecule are grouped in unique variables. The main problem of this is the fluctuations in morphologies influenced by the adduct elements. Currently is not clear how to remove the adduct natural abundance effect.[11] Other interfering ions like ion fragments and matrix adducts have been proposed to be removed to enhance statistical analysis following the same strategy.[17] Finally, depending on the study goals, removing identified ions of compounds that are known to not influence the biological problem under study could also be beneficial.

3. Soft clustering as the future of the spatial segmentation of MSI data

We have explored the use of soft/fuzzy clustering for the spatial segmentation of MSI data, particularly, the fuzzy c-means algorithm. The principal necessity for this is distinguishing between pixels inside a cluster by their similarity to the cluster centroid. Most ions do not have a spatial distribution that ends abruptly in a specific region but progressively fades transitioning to other tissue morphologies. This indicates that some regions may contain a mixture of

different tissue type, and with soft clustering we attempt to represent better this phenomenon than with hard clustering. In chapter 5 we have shown that, by identifying and removing transition pixels from the ROIs before comparing them, resulted in higher fold changes and better significance testing, which reinforces the argument of transition pixels between clusters. Some previous works have noticed this and applied soft clustering to MSI data. For instance, *Bemis et al.* developed a clustering framework combining spatially aware clustering, statistical regularization, and probabilistic segmentation,[18] The main drawback of the tool is that it has many abstract parameters to tune, and not only soft clustering is involved in the results, which limits the possibilities of studying only its effects. This is one of the reasons for studying the applicability of simpler soft clustering algorithms like the fuzzy c-means in MSI.

In our work, we have used the standard definition of membership to a cluster of fuzzy c-means, but a future line of research could be a definition of membership that considers how MSI data acquisition translates the interconnection between tissue types, and the effects of the autocorrelation between pixels. For instance, the spatial resolution and the amount of tissue sampled per area unit are key elements in unveiling the transitions between regions in detail. By defining a new score of pixel fidelity to a unique cluster we were able to study the effects of the algorithm's parameters and more importantly, reveal regions impossible to cluster with fuzzy c-means under normal circumstances. This was the case of a very specific region in the cerebellum studied in chapter 5. Even though we could not replicate the identification of the structure of the Purkinje layer by increasing the number of clusters, being able to notice secondary structures like this, shows how beneficial having a membership dimension is for the evaluation of clustering results. The soft clustering methods used in the future for MSI should account for these regions, which we suspect to be small groups of pixels sharing a particular condition. Clustering algorithms allowing very different cluster sizes (number of pixels per cluster) can contribute to the solution of this problem.

Hard clustering visual evaluation of results is usually limited to replicate structures found in histopathologic images. Histopathological imaging structures are always desirable to look for in an MSI experiment due to the bast amount of knowledge behind them,[19] but molecular imaging methods should attempt to reveal other structures complementary to histopathology. Additionally, it is difficult to assess how many of the structures should be possible to cluster at the same time. This leads to the problem of the determination of the optimal number of clusters, which is still very challenging, but we think we have shaded some light upon it. Using membership to try to determine when a clustering result has a most optimal solution provides an additional criterion to help on the number of clusters decision and in the overall assessment of clustering quality. These possibilities are the most beneficial points of soft clustering algorithms compared to their hard clustering alternatives. Still, from our investigation we conclude that multiple morphologies can coexist within the same tissue section. Depending on the ions included in an unsupervised spatial segmentation procedure, some of the morphologies will dominate over the others in the results. This can lead to interferences between morphologies, making the decision of a correct number of clusters ambiguous. Therefore, mechanisms of accounting for all the morphologies present in a tissue section should be used to improve the selection of the number of clusters.[20] An idea is establishing a hierarchy between morphologies based on the overall intensity of the distribution and the number of peaks representing them, with the aim of normalizing them to a common scale before clustering. But at the end, the unsupervised spatial segmentation of a MSI dataset does not have a unique solution as it includes a variety of biological questions. Therefore, only a supervised or semi-supervised spatial segmentation, targeting a specific condition (known molecule signatures or ROIs), can attempt to find a ‘correct’ number of clusters.

4. Conclusions

Conclusion 1: rMSIKeyIon contributes to the untargeted analysis of ion distribution in MSI datasets with ROIs.

The first objective of this thesis involved the development of an automatic workflow for the statistical analysis of ion abundancy distributions in MSI datasets. This has been achieved with the development of the R package rMSIKeyIon, based on the combination of three parameters: the non-detected ion concentration ratio, the Mann–Whitney U ion concentration test, and the FC in the ion concentration. The tool discovers automatically up and down-regulated ions between previously defined ROIs. The ions found by rMSIKeyIon are the ones of most interest to be identified, as they are responsible of the main variations between molecular signatures of different regions.

Conclusion 2: rMSIannotation is an excellent peak annotation tool for MSI experiments in the low mass range and offers the benefits of a new approach based on modeling libraries of compounds.

The second objective of this thesis involved the peak annotation of MSI datasets in the low mass range. To accomplish it, we have developed rMSIannotation, a tool that annotates carbon isotopes and adducts easily integrable in any MSI workflow. At the same time, we have developed and proven useful a new approach to compound annotation based on modeling libraries. The results of rMSIannotation show that our approach can automatically extract valuable information from both high (TOF) and ultra-high (FT-ICR) resolution spectrometers. The presented algorithm demonstrated a high performance and annotation confidence when compared to the established metabolomics MSI annotation platform METASPACE and to manual annotation approaches. Additionally, the annotations produced by rMSIannotation can be used in variable reduction strategies.

Conclusion 3: rMSIannotation can be used as a method for finding redundant features (isotopes) and discarding them before the analysis of MSI data.

We studied the effects of removing redundant data by only keeping annotated peaks. Our findings show that most information is retained by the monoisotopic peaks and therefore, removing isotopes and other non-annotated features can be beneficial in untargeted analysis. rMSIannotation facilitates this process by detailing the annotation of each peak and facilitating the data manipulation of them. This procedure can be combined with other variable reduction procedure based on different criteria.

Conclusion 4: The Fuzzy c-means algorithm allows a better evaluation of spatial segmentation results of MSI dataset using the membership to the clusters.

The third objective of this thesis aimed to evaluate the performance of the soft clustering algorithm fuzzy c-means with MSI datasets. We have shown that the study of membership defined by fuzzy c-means allows for new ways of interpreting the results compared with hard clustering results. In our case, we have approached the study of the membership through the newly developed PFS, which allows an easy selection of the pixels more related to a cluster. Thanks to the score, we were able to unveil morphological regions hidden behind more ion-rich regions and enhance a multi-sample tissue type classification workflow using a human

head and neck cancer dataset. From our work we anticipate soft clustering to be an indispensable tool for the spatial segmentation of MSI datasets.

5. References

1. da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences of the United States of America* vol. 112 12549–12550 (2015).

2. Baquer, G. *et al.* rMSIcleanup: an open-source tool for matrix-related peak annotation in mass spectrometry imaging and its application to silver-assisted laser desorption/ionization. *J. Cheminform.* **12**, 45 (2020).

3. Janda, M. *et al.* Determination of Abundant Metabolite Matrix Adducts Illuminates the Dark Metabolome of MALDI-Mass Spectrometry Imaging Datasets. *Anal. Chem.* **93**, 8399–8407 (2021).

4. Ràfols, P. *et al.* Assessing the potential of sputtered gold nanolayers in mass spectrometry imaging for metabolomics applications. *PLoS One* **13**, e0208908 (2018).

5. Abdelhamid, H. N. Nanoparticle-based surface assisted laser desorption ionization mass spectrometry: a review. *Mikrochim. Acta* **186**, 682 (2019).

6. Bertevello, P. S. *et al.* Lipid Identification and Transcriptional Analysis of Controlling Enzymes in Bovine Ovarian Follicle. *Int. J. Mol. Sci.* **19**, (2018).

7. Cañueto, D., Gómez, J., Salek, R. M., Correig, X. & Cañellas, N. rDolphin: a GUI R package for proficient automatic profiling of 1D H-NMR spectra of study datasets. *Metabolomics* **14**, 24 (2018).

8. Polanski, A., Marczyk, M., Pietrowska, M., Widlak, P. & Polanska, J. Signal Partitioning Algorithm for Highly Efficient Gaussian Mixture Modeling in Mass Spectrometry. *PLoS One* **10**, e0134256 (2015).

9. Polanska, J., Plechawska, M., Pietrowska, M. & Marczak, L. Gaussian mixture decomposition in the analysis of MALDI-TOF spectra. *Expert Syst.* **29**, 216–231 (2012).

10. Hansen, R. L. & Lee, Y. J. Overlapping MALDI-Mass Spectrometry Imaging for In-Parallel MS and MS/MS Data Acquisition without Sacrificing Spatial Resolution. *J. Am. Soc. Mass Spectrom.* **28**, 1910–1918 (2017).

11. Hankin, J. A. *et al.* MALDI mass spectrometric imaging of lipids in rat brain injury models. *J. Am. Soc. Mass Spectrom.* **22**, 1014–1021 (2011).

12. Palmer, A. *et al.* FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nat. Methods* **14**, 57–60 (2017).

13. Wishart, D. S. *et al.* HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res.* **50**, D622–D631 (2022).

14. Baquer, G. *et al.* What are we imaging? Software tools and experimental strategies for annotation and identification of small molecules in mass spectrometry imaging. *Mass Spectrom. Rev.* e21794 (2022).

15. Cassese, A. *et al.* Spatial Autocorrelation in Mass Spectrometry Imaging. *Anal. Chem.* **88**, 5871–5878 (2016).

16. Palmer, A. D., Bunch, J. & Styles, I. B. The use of random projections for the analysis of mass spectrometry imaging data. *J. Am. Soc. Mass Spectrom.* **26**, 315–322 (2015).

17. Garate, J. *et al.* Influence of Lipid Fragmentation in the Data Analysis of Imaging Mass Spectrometry Experiments. *J. Am. Soc. Mass Spectrom.* **31**, 517–526 (2020).

18. Bemis, K. D. *et al.* Probabilistic Segmentation of Mass Spectrometry (MS) Images Helps Select Important Ions and Characterize Confidence in the Resulting Segments. *Mol. Cell. Proteomics* **15**, 1761–1772 (2016).

19. Patterson, N. H. *et al.* Next Generation Histology-Directed Imaging Mass Spectrometry Driven by Autofluorescence Microscopy. *Anal. Chem.* **90**, 12404–12413 (2018).

20. Picard de Muller, G., Ait-Belkacem, R., Bonnel, D., Longuespée, R. & Stauber, J. Automated Morphological and Morphometric Analysis of Mass Spectrometry Imaging Data: Application to Biomarker Discovery. *J. Am. Soc. Mass Spectrom.* **28**, 2635–2645 (2017).

UNIVERSITAT
ROVIRA i VIRGILI