Université de Lille - Sciences et Technologies

École Doctorale des Sciences de la Matière, du Rayonnement et de l'Environnement

Thèse de Doctorat

En vue de l'obtention du grade de

Docteur de L'Université de Lille

Discipline: Chimie theorique, physique, analytique

Ruggero Guerrini

Boosting Unsupervised Analysis in Molecular and Elemental Hyperspectral Imaging Through Chemometrics

Sostenue le 26 09 2025:

Président du jury:	
Federico MARINI	Professeur, La Sapienza (Italy)
Rapporteurs:	
José Manuel AMIGO RUBIO	Professeur, UPV-EHU (Spain)
Nerea BORDEL	Professeur, University of Oviedo (Spain)
Eric ZIEMONS	Professeur, Unversité de Liege (Belgium)
Examinateurs:	
Vincent MOTTO-ROS	Maître de Conférences HDR, Université Lyon 1 (France)
Cécile FABRE	Professeur, Université de Lorraine (France)
Directeur de Thèse:	
M. Ludovic DUPONCHEL	Professeur, Université de Lille (France)
Co-encadrant	
Mrs. Nina OGRINC	Dr Researcher, Leiden University Medical Center (Netherlands)

Université de Lille - Sciences et Technologies

École Doctorale des Sciences de la Matière, du Rayonnement et de l'Environnement

Thèse de Doctorat

En vue de l'obtention du grade de

Docteur de L'Université de Lille

Discipline: Chimie theorique, physique, analytique

Ruggero Guerrini

Renforcement de l'analyse non supervisée dans l'imagerie hyperspectrale moléculaire et élémentaire grâce à la chimiométrie

Sostenue le 26 09 2025:

Président du jury:	
Federico MARINI	Professeur, La Sapienza (Italy)
Rapporteurs:	
José Manuel AMIGO RUBIO	Professeur, UPV-EHU (Spain)
Nerea BORDEL	Professeur, University of Oviedo (Spain)
Eric ZIEMONS	Professeur, Unversité de Liege (Belgium)
Examinateurs:	
Vincent MOTTO-ROS	Maître de Conférences HDR, Université Lyon 1 (France)
Cécile FABRE	Professeur, Université de Lorraine (France)
Directeur de Thèse:	
M. Ludovic DUPONCHEL	Professeur, Université de Lille (France)
Co-encadrant	
Mrs. Nina OGRINC	Dr Researcher, Leiden University Medical Center (Netherlands)

Alla mia famiglia
A mia madre, a mio padre
A mia sorella
Ad Artù, ad Aslan, Klimt e Monet

Acknowledgements

And here we are, after three years.

During my Bachelor's and Master's degrees, I loved the acknowledgements. Today, it's quite different. It has a different feel to it. This time, I don't feel comfortable writing this part. I'm pretty sure that these acknowledgements can't fully express how much you have all truly done.

First of all, I would like to thank my supervisors, Ludo and Nina. Working with me was not easy: I am quite hyperactive and hard to keep up with! Thank you both for your patience and flexibility. Ludo, you understood me and supported me through some really tough times, providing the peace of mind I needed most. I'll never forget it.

I would like to thank all the members of the PhD defence committee. Thank you for agreeing to read and evaluate my work.

Volevo ringraziare Raffaele per tutto ciò che ha fatto per me. Ci è voluto un pochino per entrare in confidenza, o meglio più per me per sentirmi a mio agio. Ma da subito mi hai preso a cuore e ti sei curato di me. Non mi hai mai lasciato solo neanche nei momenti peggiori, e mi hai sempre aiutato, sia lavorativamente che non. Le conversazioni e le discussioni lavorative mi han dato molto. Senza di te non avrei raggiunto tutto quello che ho ottenuto, soprattutto in termini di pensiero critico. Sei stato un fratello maggiore per tutto questo tempo, e anche se la vita ci ha messo vicini soltanto per un breve periodo, ovunque sarò ti considerò sempre come tale.

My lifelong friends: Zara, Nicola, Setzi, Giacomo, Arianna and Gaia. Having me as a friend is not easy: I disappear for months and then reappear out of nowhere. The hardest thing is knowing that although life brought us together, it has also put thousands of kilometeres between us. Some of you feel this distance more than others. Zara, I know you want to kill me for always moving around and never staying in one place, but deep down, I know you're happy and proud of me.

Giacomo, one day we will work out how to put our ideas into practice.

Nicola, your admiration has given me strength during my hardest moments. Although we haven't spoken much since our studies took us in different directions, the connection remains as strong as it was when I was in Pavia.

Ale, every time I'm in Milan, I need to see you. You've actually been the oldest friend mentioned in these acknowledgements since high school. I hope there will be

an opportunity for us to work together.
Ari, I miss our chats and coffees in Rome. Your determination motivated me through the worst of times.
Gaia, although we met by chance, our friendship and affection will always endure.
All the friends I met thanks to Lille: Ceci, Yesid, Christopher, Fulvio, Daniele, Alessandra, Lorenzo, Federico and Nicholas. You have all enriched my life through your experiences, conversations and advice. Being open and honest with you felt natural from the very beginning.
Quiero agradecer a Adri, mi amigo perezoso. Te lo escribo por separado solo porque quería escribirlo en español. Empecé a hablar español contigo recién en febrero, y he aprendido mucho gracias a ti.
Fede, who although we officially separated at the end of the master's degree, you were always close to me. If it were not for you, I would not be here today, and my future would not be like the one I have ahead of me.
There would be no lab 156 without you, Eugenio. I think we've known each other for three and a half years. It seems like yesterday that we first met, and we've been through a lot together. As you were always there for me, I will always be here for you.
Demi. In poco meno di un anno abbiamo vissuto tanto. Purtroppo la distanza non ha aiutato a viverci in pieno come avremmo voluto. Sei stata una parte importantissima di questo ultimo anno. Dalle mozzarelle, passando per i cinghiali ed il tuo primo volo. Le tue lettere, i tuoi biglietti al mattino, i caffé a letto. Ci sono tante cose che vorrei dire, ma dei ringraziamenti in una tesi non bastano. É anche merito tuo se sono riuscito a raggiungere questo traguardo. Posso ora aggiungere che non vedo l'ora del futuro che ci aspetta. Dovevo correggere la tesi, e non potevo farlo senza correggere questa parte. É tanto importante il dialogo che stiamo costruendo, e come ci sosteniamo a vicenda. Ti amo tanto, micetta. O se preferisci, gnappetta.
My second family, Annabella, Enzo, Siria, Gea, Buster, Hermes. Although we've seen each other very little over the past three years, you've always been there for me through thick and thin. You're like a second family to me.
Last but not least, my family: Mom, Monica, Dad (wherever you are), Manuel, Cristiano, Luca, and now Luca (x2). And I cannot forget Artù, Aslan, Klimt and Monet either. You have always believed in me and supported me through the worst of times. I hope I can make you proud every day. I'm sorry, Mum. I'm sorry,

sister. There's probably some nomad DNA in me. I need to explore, push limits and experience life, and I just can't stop. Maybe I inherited the best part from Dad.

Pe' arrivacce qui, da Roma ho fatto l'autostop
E in Francia è già 'n ber pezzo che ce sto
Ma pure da emigrato
Mica so cambiato
Io so' Romeo
Er mejo der Colosseo

Contents

Abstract

Today, hyperspectral imaging (HSI) plays a key role among the analytical techniques. The challenge is to improve both spatial and spectral resolution as well as acquisition speed. In this situation, chemometrics plays a key role as a relevant tool to extract meaningful information from the data. MALDI and LIBS can be considered the most used in molecular and elemental imaging respectively. Although they share the same problem of data dimension, the different data acquisition process, the different type of data, data storage, as well as the different purpose of analysis explain the need for different tools between LIBS and MALDI. In MALDI, as a non-target analysis, clustering has more quickly become one of the most important techniques studied due to its ability to group similar objects, e.g. similar region in a MALDI image. In cancer diagnosis, tissues are usually examined by the golden standard of histopathological and immunohistochemical analysis. However, these are targeted analyses, so they are not suitable if we are looking for new biomarkers. In fact, MALDI can provide complete spatially resolved biological information. With this amount of information it is possible to segment the image to a region with similar biochemical information. As it will be explored later, a common technique it is to follow the histological annotation to cluster the data with an interactive hierarchical clustering technique (bisecting kmeans), or even if there is no histological annotation the segmentation is mainly driven by the user experience, so it can be really biased. For this reason, one of the main purposes of this research was to find an appropriate chemometric approach to cluster the MALDI data in a less biased way, using a hierarchical approach, giving all the sheets to be studied and evaluating their division without using prior information such as histological annotation to match or premonitions.

LIBS is widely used in various fields [1]. Millions of spectra can be measured every day, which is a real challenge for data set analysis [2]. In the field of mineral

analysis, a sample can contain multiple phases and different minerals can have very close elemental compositions, making data analysis really difficult. Even if the composition does not appear to be different, they can appear to be different from a visual perspective. This discrepancy is important to investigate because it can help to really understand and explore the sample. For this reason, the HSI data fusion approach with LIBS and RGB has been explored to enhance the LIBS analysis.

In addition, the possibility of fusing elemental and molecular information in a biological context is being explored. As the most used in its field, the fusion between MALDI and LIBS can enhance the exploration of biological tissue, allowing elemental and molecular information to be correlated, with the possibility of understanding whether exogenous elemental modifications can cause molecular variations related to disease. As this is a pioneering study, we started with a rat sagittal brain sample to see the feasibility of the technique.

Given the state of the art, the work was not so much focused on developing brand new advanced methods, but rather on changing the approach and calibrating the research question, purpose and technique.

Résumé

Aujourd'hui, l'imagerie hyperspectrale (HSI) joue un rôle clé en chimie analytique. Le défi consiste ainsi à améliorer à la fois la résolution spatiale et spectrale ainsi que la vitesse d'acquisition. Dans ce contexte, la chimiométrie joue un rôle essentiel en tant qu'outil pertinent pour extraire des informations significatives à partir des données. Le MALDI et le LIBS peuvent être considérés comme les techniques les plus utilisées pour l'imagerie moléculaire et élémentaire respectivement. Bien que ces techniques partagent le même problème lié à la dimension des données, les différences dans le processus d'acquisition, le type de données, le stockage des données ainsi que les objectifs d'analyse expliquent la nécessité de recourir à des outils différents pour ces deux modalités. Dans le cas du MALDI, en tant qu'analyse non ciblée, le clustering est rapidement devenu l'une des techniques les plus importantes étudiées en raison de sa capacité à regrouper des objets similaires, par exemple des régions similaires dans une image MALDI. Dans le cadre du diagnostic, les tissus sont généralement examinés selon une méthode de référence comme l'analyse histopathologique ou l'immunohistochimique. Cependant, ces analyses étant ciblées, elles ne sont pas adaptées si l'on cherche de nouveaux biomarqueurs. En effet, le MALDI peut fournir des informations biologiques complètes avec une bonne résolution spatiale. Avec cette masse d'informations, il est possible de segmenter l'image en régions ayant des informations biochimiques similaires. Une technique courante consiste à suivre l'annotation histologique pour regrouper les données à l'aide d'une technique de clustering hiérarchique interactive (bisecting k-means), ou même, en l'absence d'annotation histologique, la segmentation est principalement guidée par l'expérience de l'utilisateur, ce qui peut introduire un biais important. Pour cette raison, l'un des principaux objectifs de cette recherche était de trouver une approche chimiométrique appropriée pour regrouper les données MALDI de manière moins biaisée, en utilisant une approche hiérarchique, en donnant toutes les

branches à étudier et en évaluant leur division sans utiliser d'information préalable comme l'annotation histologique ou des présupposés. Le LIBS[1] est largement utilisé dans divers domaines. Des millions de spectres peuvent être mesurés chaque jour, ce qui représente un véritable défi pour l'analyse des ensembles de données[2]. Dans le domaine de l'analyse minérale par exemple, un échantillon peut contenir plusieurs phases et différents minéraux peuvent avoir des compositions élémentaires très proches, rendant l'analyse des données particulièrement difficile. Même si la composition ne semble pas différente, elle peut apparaître différente visuellement. Cette divergence est importante à étudier car elle peut aider à mieux comprendre et explorer de tels échantillons.

Pour cette raison, l'approche de fusion de données HSI entre les imagerie LIBS et RGB a été explorée pour améliorer l'analyse LIBS d'échantillons complexes. De plus, la possibilité de fusionner des informations élémentaires et moléculaires dans un contexte biologique est en cours d'exploration. En tant que techniques les plus utilisées dans leurs domaines respectifs, la fusion entre le MALDI et le LIBS peut améliorer l'exploration des tissus biologiques, en permettant de corréler les informations élémentaires et moléculaires, avec la possibilité de comprendre si des modifications élémentaires exogènes peuvent entraîner des variations moléculaires liées à des maladies par exemple. Comme il s'agit d'une étude préliminaire, nous avons commencé avec un échantillon sagittal de cerveau de rat pour évaluer la faisabilité de la technique. Étant donné l'état de l'art, le travail ne s'est pas tant concentré sur le développement de méthodes avancées entièrement nouvelles, mais plutôt sur un changement d'approche et une reformulation de la question de recherche.

Chapter 1

Introduction

Analytical chemistry is the branch of chemistry concerned with characterising the chemical and physical composition of a sample. Typically, a representative portion of the sample should be selected for sampling. This can be important and efficient for some common analyses, such as water analysis. However, in some cases the heterogeneity of the sample is relevant and it is not possible to represent the sample with only a small part of it. In this case, the normal approach is not applicable.

If, for example, the aim of the study is to find the contaminants in a field and to understand their diffusion, the process will be to construct a grid over the sample and to take a sample for each cell. The number of samples is related to the density of the grid. Subsequent subsampling will involve the implementation of spectrometric analysis. This provides not only chemical and physical information, but also spatial information. An example of grid sampling applied to a square field is shown in figure 1.

S-1	S-2	S-3	S-4
S-5	S-6	S-7	S-8
S-9	S-10	S-11	S-12
S-13	S-14	S-15	S-16

Figure 1: Example of a grid-sampling

When considering the entire figure as an image of the sample, it can be observed that each previously defined subsampling corresponds to a pixel. The result is what can be described as an **hyperspectral image** (HSI) of the sample [3].

Just as grayscale or Red-Green-Blue (RGB) images provide information about color and texture, hyperspectral images can be viewed as normal images where instead of 1 (Grayscale) or 3 (RGB) channels there are hundreds of thousands. Each of these channels corresponds to, e.g., a specific lambda λ or m/z value.

Elemental and molecular analysis are two sides of the same coin. Molecular analysis finds its main application in biomedical research. On the other hand, elemental analysis is more interesting in materials science, applied to geology, environmental monitoring, cultural heritage and agriculture.

Both have benefited from important advances in recent decades, and these improvements in speed and spatial resolution have made it possible to obtain more information about samples in less time.

Among the various techniques, Matrix-Assisted Laser Decomposition (MALDI) and Laser-Induced Breakdown Spectroscopy (LIBS) can be considered the gold standards of hyperspectral imaging for molecular[4, 5] and elemental [6] analysis, respectively. Table 1 report an overview of MALDI and LIBS imaging.

The features of these techniques allowed them to became the golden standard of their fields. Due to the wide range of biomolecules that can be acquired and the tissue-type specific molecular profile that can give, MALDI became one of the most important molecular technique [7, 8] LIBS is one of the most interesting approach for elemental analysis, used also for the exploration of Mars by the NASA [1].

The amount of data generated by both techniques is impressive (up to millions of spectra) in a relatively short time (less than an hour). Considering that each spectrum is made up of thousands or more variables, the dimensions of the data generated are important. so efficient handling of this data is critical.

Chemometrics is a good candidate to address these issues. Using a proper approach based on statistical methods or machine learning algorithms, it is possible to extract relevant information from the images and study the heterogeneity among them.

Purpose of the thesis

Here, finally, the purpose of my thesis: to improve some unsupervised approach for two of the most used hyperspectral imaging (MALDI and LIBS). One of the crucial points of MALDI imaging is the segmentation (clustering). Without prior knowl-

	MALDI imaging	**LIBS Imaging**
Purpose	Molecular imaging of biomolecules	Elemental imaging
Laser role	Exciting the matrix, causing desorption and ionization	Ablation and formation of plasma
Sample preparation	Matrix is deposed uniformly to maintain the spatial distribution of molecules	Almost no pretreatment needed
Detection	Mass Spectrometry	Emission Spectroscopy
Spatial Resolution	10-100 μ	10-100 μ
Laser Speed frequency	up to 10 kHz	up to 1 kHz
Application	Biomarker discovery, tissue analysis	Material Science

Table 1: Comparison of MALDI Imaging and LIBS imaging

edge, it is not easy to find and validate the appropriate segmentation algorithm. The aim of this work is to improve clustering analysis. In addition, a first step towards the fusion of molecular and elemental information was made with the fusion of LIBS imaging and MALDI imaging.

The thesis is organized as follows: First, the basic instrumentations are presented. This is followed by a basic introduction about the basis of chemometrics. Then, for each chapter, a specific problem is introduced to clarify the state of the art in this field, followed by the solutions proposed and adopted during this thesis. The *state of the art* section as an introduction before each chapter was a choice to help the reader follow the discussion due to the gaps between the arguments. Code and processing are explained not only theoretically, but also Matlab code are presented to help the reader follow and give him directly also tools.

1.1 Mass Spectrometry

Mass spectrometry is an analytical technique that allows atoms and molecules to be ionized, separated and analyzed as an ionic gas based on their m/z ratio. It is useful for the identification of unknown samples (organic, inorganic and biomolecules), obtaining structural information and, coupled with other techniques (e.g. GC/LC chromatography), quantitative analysis in complex mixtures. A simplified diagram of the instrumentation is showed in figure 2

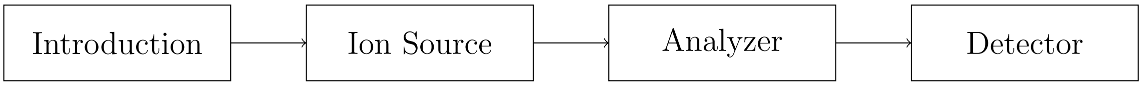


Figure 2: Diagram of a mass spectrometer

The first step is the introduction of the sample into the system. This process depends on the physical state of the sample (gas, solid or liquid). The introduction system allows the sample to be introduced into the instrument. Nowadays the introduction systems are automatized, so it can work with autosampling or coupled with separations techniques as chromatography.

Once the analytes are inside the instrumentation, the source has to ionise the atoms (or molecules) of the sample. There are different techniques such as Electrospray Ionisation (ESI), Electron Impact (EI) or Matrix-Assisted Laser Desorption/Ionisation (MALDI). During this thesis, the latter was used because it is able to detect different types of biomolecules. The mass analyser separates the ions by their *m/z* ratio. There are different types of mass analyser, such as quadrupole, time-of-flight (TOF) and ion cyclotron resonance (ICR). Each analyser uses a different technique. The detector converts the ion beam into a mass spectrum. The latter are under high vacuum to prevent the ions from being disturbed by interactions with molecules in the atmosphere.

Fundamentals In the mass spectrometer, analyte molecules are converted to ions by applying energy to them. The ions formed are separated by their mass-to-charge ratio (m/z). m/z is an unitless ratio, as it is the mass number divided to the number of fundamental charges z on the ion. It is equal to the mass of the ion only if the ion is singly charged. MS it is not a real spectroscopy, because it uses electrons (or other ionizing agents) and this interaction causes the ionization of the analytes and

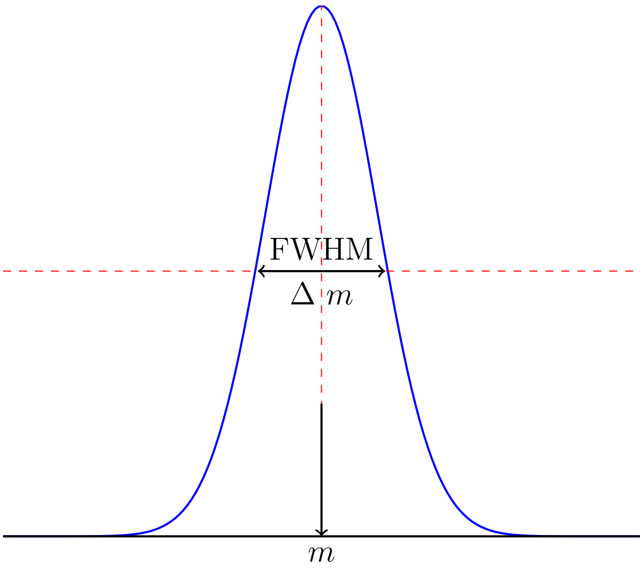


Figure 3: Representation of how the resolution is calculated

the subsequent fragmentation. It means that this technique is **destructive**. It is not possible to recover the sample after the analysis. The neutral molecules are not observed.

Resolution Resolution is the ability of a mass analyzer to obtain distinct signals for two ions with a small difference in m/z. The Marshall resolution definition is

$$R = \frac{m}{\Delta m} \tag{1}$$

A representation of how the resolution is calculated is in figure 3 where δm is the width of the peak on a specific height, usually at 50% (FWHM - Full Width at Half Maximum). This definition is important because it does not requires two peak with similar intensities and close m/z values. Based on the kinetic theory of gases, the *mean free path* is:

$$L = \frac{kT}{p\sigma\sqrt{2}} \tag{2}$$

where k is the Boltzmann constant, T is expressed in Kelvin, p in Pascal and σ is the collision section. Value of pressure and mean free path are shown in table 2 Around 10^{-9} atm is the pressure of the mass analyzer in the instruments. The *mean free path* is important because the ion should reach the detector to give a signal. A visual representation of the mean free path is shown in figure 4

Pressure	L
1 atm	$\sim$ 64 nm
0.001 atm	$\sim$ 64 μm
10^{-6} atm	$\sim$ 64 mm
10^{-9} atm	$\sim$ 64 m

Table 2: Mean free path

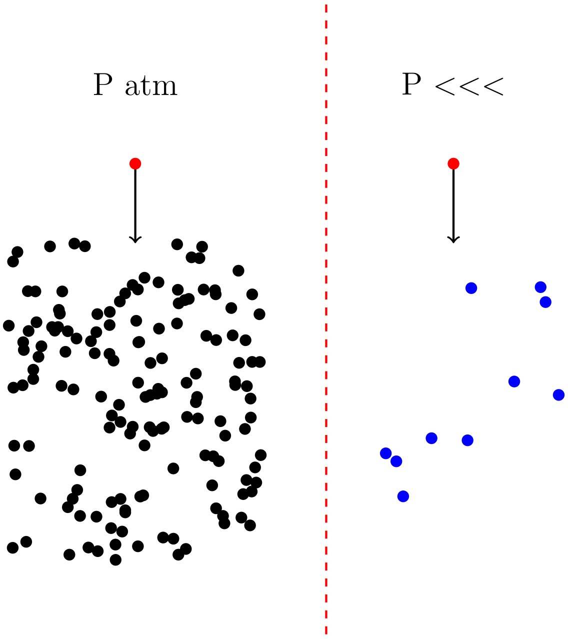


Figure 4: Representation of the differences between P atm and P <<<. As it is possible to see, in the case of P <<< the probability of collision between the ion and other molecules is lower.

1.1.1 MALDI

MALDI is a condensed-phase technique. It allows molecules to be vaporised and ionised in a single step, enabling the analysis of volatile and thermolabile compounds, even those with high molecular weights, such as whole proteins. The analytes are dispersed in the matrix, which plays a crucial role. When a laser beam is applied, the matrix absorbs the energy from the laser, desorbs and transfers some of the energy to the analytes, ionising them. There are key parameters of laser and matrix that are crucial for this technique:

Laser: Key parameters of laser are the wavelength, the speed and the beam size

Matrix: key parameters of matrix are the capability of co-crystallizing with the analytes, the stability in vacuum, high absorption for the laser wavelength, the non reactivity with the analytes, the less tendance of fragmentation and autoionisation, and the capability of transfer the energy to the analytes.

The wavelength of the laser should be optimal for the analysis, because it should be well absorbed by the matrix. The speed has an impact on the preservation of the the analytes, long beam laser could degrade the analytes. The beam size has an important role for the potential spatial resolution of the analysis.
Co-crystallisation is crucial. The analytes are distributed throughout the matrix and the molecules are rounded by matrix molecules. This facilitates energy transfer from the matrix to the analytes. The dispersion of the analytes should be as homogeneous as possible, otherwise variability would be introduced between different shots on the same sample. The size of the crystals is also important. Small crystals allow better spatial resolution because the crystallisation is better organised. The matrix should be stable in vacuum and transfer energy to the analytes without reacting with them, otherwise signals will be lost. At the same time should not ionize itself, or at least do not cover the m/z range of the analytes, otherwise can easily cover the relevant signal. All of these properties should be well balanced to optimize the analysis. To explain how the co-crystalization and the dispersion of the analytes trough the matrix can be done, example of sample preparation for MALDI analysis will be presented.

- **Dried Droplet:** the sample and matrix are first mixed and then applied to the substrate

- **Thin Layer:** the matrix is dissolved in acetone, placed on the substrate, dried and crystallized. the sample is placed on the top of the matrix layer and allowed to dry

- **Sandwich Method:** The matrix is applied, then the sample and the the matrix again. Everything is air dried.

There were just three ways to prepare the sample for MALDI analysis. For MALDI imaging sample preparation is different but the key factor of laser and matrix will remain. To conclude the introduction about MALDI, it might play an important role due to its unique advantages, including high sensitivity, a wide range of molecules (thermolabels molecules can be analyzed), molecular specificity, and the flexibility to analyze many varied analytes on a single platform.

1.1.2 MALDI Imaging

MALDI imaging is an imaging mass spectrometry technique able to combine the traditional MALDI analysis with spatial mapping capabilities, allowing the molecular analysis across a surface with high spatial resolution. For each pixel there is a mass spectrum. In fact, the spatial resolution is easily around ~ 10 μm, a cell resolution, so it is possible to see the distribution of analytes in the different cells of the tissue. On table 3 is reported a comparison between the Traditional MALDI vs MALDI imaging.

Traditional MALDI	**MALDI Imaging**
Provides a single mass spectrum for the entire sample	Acquires spectra point-by-point, creating a spatial map of molecular distributions
Sample and matrix are mixed or co-crystallized uniformly	Matrix is deposed uniformly to maintain the spatial distribution of molecules
n samples x z mz values	nxm spatial coordinates x z mz values
Commonly used for global molecular identification.	Widely applied in tissue imaging and biomarker localization.

Table 3: Comparison of Traditional MALDI and MALDI Imaging

Sample preparation and analysis

Tissue (previously frozen) is cut, typically 10-20 μm thick, deposited on a MALDI plate and a matrix is sprayed over it, typically using a spray-through machine to increase the reproducibility and uniformity of the deposited matrix. The analytes will migrate from the tissue into the matrix and co-crystallisation will occur. This process is critical to understanding some of the results that can be obtained by chemometric analysis of MALDI imaging data. The migration should be homogeneous throughout the sample, otherwise the signal will not properly correspond to

the true concentration. Once the laser hits the matrix, it desorb with the analytes, transferring energy and ionising them. The signal is usually in the low m/z range. If there are signals from the analytes that have fallen into this region, they may be masked by the matrix signal. As we will see later, it is possible to consider which region of the spectrum contains the matrix signal and which does not. In the case of MALDI imaging, a first division can be made between where there is only matrix and where there is matrix and sample. In the first case, all the energy absorbed by the matrix is used to ionise the matrix directly, as there are no analytes to transfer the energy to. If there are analytes, the energy is used to ionise the analytes, as mentioned above. Here the signal from the matrix will be lower. This causes a relevant variance of the matrix signal between the two regions of the image. This variance is so relevant that it is common to see the separation between matrix and tissue as the first principal component. This is an important tool for creating a mask that allows the next step to focus only on the pixels of the sample.
In some instruments, MALDI imaging allows the selection of an area to be scanned. Typically, biological tissues are not perfectly square. For this reason, it is not convenient to analyse a square section containing the entire sample, as a large part of it will be just matrix. This helps to reduce the dimensionality of the data.

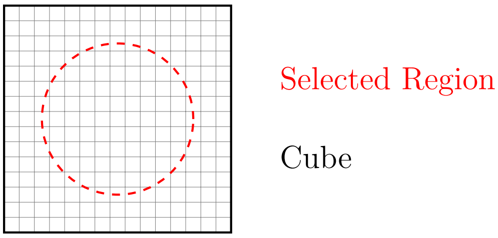


Figure 5: Visual representation of MALDI imaging. It is possible to select the region to analyse. The final output will be always a cube, with a certain amount of NAN values

In this region, for each *(x,y)* coordinate, the sample is hit by the laser, desorption, ionization, separation and detection of analytes takes place. A mass spectrum is then acquired for each *(x,y)* coordinate. The results of a MALDI analysis is a cube of dimensions n,m,z where n,m indicate the spatial dimensions and z indicate the spectral dimension. The result is complete information on the (bio)chemical composition and spatial distribution throughout the sample. The biological samples usually does not have a perfect square shape, so some instrumentation allows to acquire the mass spectra on a specific region (selecting it based on camera). It also helps to save memory, because of the reduced number of acquired spectra.

Data type

MALDI Imaging provides two types of spectra: **centred spectra** or **profile spectra**. The raw signal acquired by the instrument is the profile spectrum, which retains information about the shape and width of the peaks. The centroided spectra are an elaboration of the profile spectra, retaining only the most relevant features of the peaks (e.g. maximum intensity or integration of the peaks).

Role of MALDI imaging in tissue analysis

Histopatological staining and immunohistochemistry (IHC) are the state-of-the-art for diagnosing and staging of tumors[9], with a complementary role. HIstopatological staining focuses on the morphological study of tissue. IHC allows direct visualisation of the spatial distribution of individual proteins within tissue. However, each protein requires a specific antibody and multiplexing is very limited, typically allowing the detection of no more than two proteins at a time. IHC is a targeted method and is not suitable for untargeted studies [10]. In table 4 is presented a comparison between histopatological staining, immunohistochemistry, and MALDI imaging.
Thanks to the capability of MALDI imaging to visualize the spatial distribution of a wide type of biomolecules, such as glycans, lipids, proteins, or small molecule drugs, by their molecular masses the technique become increasingly popular. The spatial distribution of biomolecules can be directly correlated with the tissue histology.

Histological analysis and MALDI imaging

As MALDI imaging is a non-targeted analysis, the importance of this technique was clear from the first application. Once the histological analysis had been carried out, the MALDI analysis of the tissue made it possible to better characterise and study the differences observed in the histological analysis. Coupling these two techniques was therefore crucial. There are different ways of combining the information obtained from these two techniques. Later, the clustering techniques (REF) and their role in tissue analysis will be better explained. For now, it is enough to say that with MALDI it is possible to see the distribution of specific molecules that can be

	Histology	**IHC**	**MALDI imaging**
PROS	Well-estblished method	Well-established method	Label-free
	Easy to perform	highly specific	Extensive information about (bio)chemical composition and distribution
CONS	Non specific	Depends on the availability and quality of antibodies	Limited spatial resolution
	Subjective interpretation	Restricted to a few target antigens	Requires specialized instrument

Table 4: Comparison of histology, immunohistochemistry (IHC) and MALDI imaging

both known or unknown in advance, to better characterise the region, or to obtain spectra based not on all the tissue but specific selected region, thus analysing the composition of a specific selected region.
Coupling histological information with MALDI one can be done on two different approach:

- **Same tissue:** staining on the same tissue section used for MALDI-MSI analysis, either before or after the analysis
- **Consecutive tissue:** staining on the consecutive tissue section used for MALDI-MSI analysis

Staining the same section allows unambiguous correlation between MALDI-MSI images and histological images, but may be hampered by potential loss of tissue integrity after analysis or during removal of the MALDI matrix. On the other hand, the use of consecutive tissue sections avoids these problems, but introduces the uncertainty that adjacent sections may not be completely identical.

Importance of data analysis in MALDI imaging

MALDI provides a hyperspectral image of the tissue with a mass spectrum as the third dimension. Even without prior knowledge of the tissue, it is possible to extract relevant information about it using **unsupervised approaches** from chemometrics. These approaches look for latent data structures and underlying relationships that are not immediately observable by univariate statistical methods, without any prior information about dominant categories or variables. Of all the unsupervised approaches, clustering is one of the most interesting. Clustering, which will be discussed in more detail later in the thesis, is a group of multivariate data analyses that are able to cluster homogeneous elements into a group. In this domain, elements correspond to pixels, and each pixel is a spectrum, so clustering is not only the appropriate chemometric analysis that is important. There are different ways of analyzing or extracting information from the data. It is therefore important that the results of the data analysis are read by a biologist in order to properly evaluate the results, so collaboration between these two worlds is important.

An interesting review regarding the spatial multiomics analysis was published recently [11]. An in-depth analysis of the various techniques and evolutions is given. MALDI imaging has gained importance in the *omics science*, particularly in *metabolomics* and *proteomics*, due to its spatial resolving power, making them *spatial omcs sciences.* Spatial omics refers to all techniques capable of molecular spatial resolution, used to analyze biological molecules and their distribution within a tissue. Table 5 presents the spatial omics sciences

Spatial Omics Sciences

Technique	**Description**
Spatial Transcriptomics	Gene expression
Spatial Genomics	Physical arrangement of the genome
Spatial Proteomics	Proteins
Spatial Epigenomics	Modification to the DNA sequence
Spatial Metabolomics	Metabolites

Table 5: Main features, description and differences of Spatial Omics techniques

Challenges and future research perspectives

Subcellular acquisition is a key goal of recent research[12]. However, several challenges must be addressed, such as the lower ion production, which requires more sensitive instrumentation. Crystals size is also crucial as it should be smaller than the laser spot. Smaller crystals enhance desorption and ionization of analytes due to a higher laser intensity at the focal point. The ultimate goal is to study single cells in their natural tissue environment [13]. This purpose of pushing spatial resolution involves challenges related to the increase of data dimension [14]. Last but not the least, identification of molecular species has a key role in MALDI imaging analysis [15]. The m/z values obtained should be traced back to unique compound identification to have a meaningful biological and diagnostic conclusion. The untarget nature of MALDI-imaging has its own pros and cons. m/z ratios need to be identified as a unique compound. One of the most famous is METASPACE [16]. In the context of unsupervised analysis, this can helps to identify new molecular species. e.g., after this segmentation, the idea was to get a list of the most interesting features of each region of the samples. However, identification and annotation are not really part of the main purpose of this work, so it won't really be explored.

1.2 Emission Spectroscopy

Emission spectroscopy is an analytical technique based on the light emitted by excited atoms. Analyte atoms are excited by heat or electrical energy. The energy is typically supplied by a plasma, a flame, a low-pressure discharge, or a high-powered laser [17]. Atoms begin in the **ground** state and, upon absorbing energy, transition to a higher energy level, the **excited state**. Each element emits a unique spectrum, which acts like a fingerprint for identifying the elements. This makes the technique useful for qualitative analysis. Various approaches and types of instrumentation are used. Among them, Laser Induced Breakdown Spectroscopy (LIBS) will be introduced as instrumentation used in this thesis.

1.2.1 LIBS

LIBS is an emission spectroscopy technique based on a plasma source. The duration of plasma is about nanoseconds and its power causes a dielectric breakdown. This creates a highly plasma. At the end of the laser pulsation, the plasma cools down and the radiation emitted by the excited atoms (and ions) can be detected.

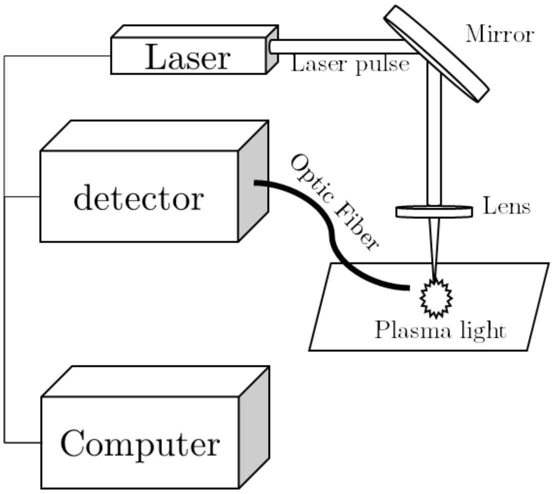


Figure 6: Representation of a LIBS system

A number of salient properties of lasers employed in LIBS must be considered, including wavelength, pulse energy, irradiance, directionality, and monochromaticity. Typical pulse energies used in LIBS range from 10 mJ to 500 mJ. Given that the energy of a visible photon is around $\sim 10^{-19}$ J, this corresponds to approximately 10^{17} to $5 \cdot 10^{18}$ photons per pulse [6]. The most interesting properties of laser is the very narrow spectral range where the majority of the output energy is concentrated, unlike conventional light sources that are broadband. However, fundamental frequency of the laser can be not the optimal one for the excitation. To overcome this problem, it is possible to pass the laser pulse through a suitable birefringent material. The complete emission spectrum of each element is unique. All the elements ablated with the laser and present in the plasma and that generate light, should be representative of the sample. To be significative, plasma should be in a thermodynamic equilibrium and the elemental composition should be the same as the one in sample. In this case there is a connection between the spectral line intensities observed and the relative concentrations of elements.

1.2.2 LIBS imaging

Traditional LIBS vs LIBS imaging LIBS Imaging is an spectrometry technique able to combine the traditional LIBS analysis with spatial mapping capabilities, allowing the elemental analysis across a surface with high spatial resolution. On 6 is reported a comparison between traditional LIBS and LIBS imaging.

Traditional LIBS	**LIBS Imaging**
Provides a single spectrum for the entire sample	Acquires spectra point-by-point, creating a spatial map of elemental distributions
n samples x λ wavelength	nxm spatial coordinates x λ wavelength
Commonly used for quick, localized measurements.	Widely applied in studies involving tissue imaging, material characterization, and environmental monitoring.

Table 6: Comparison of Traditional LIBS and LIBS Imaging

Sample preparation and analysis The diameter of the laser is directly proportional to the diameter of the resulting ablation crater. This limitation results in a reduction in the spatial resolution that can be achieved, due to the requirement that the step size (the distance between two consecutive laser shots) must exceed the diameter. The surface of the samples should be correctly polished, as the focal point of the laser should be constant and correct all over the surface.

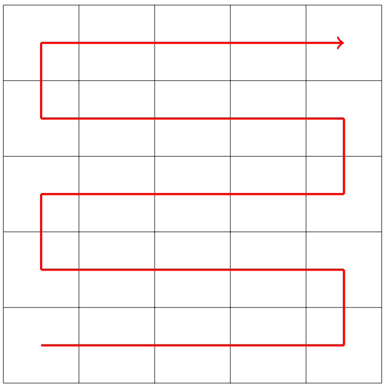


Figure 7: Possible path of acquisition

As it is shown in figure 7, the acquisition is sequential. This will cause that the odd (or even) rows are flipped. This is crucial to keep in mind once the data need to be imported. An initial step of flipping them is needed.

Importance of data analysis in LIBS imaging As with other hyperspectral techniques, the instrumentation for LIBS imaging has evolved greatly in recent years, allowing more data to be generated in less time. The advent of novel techniques has rendered the attainment of kHz acquisition frequencies a possibility. This facilitates the analysis of large areas in a reduced timeframe. As an example, with this technology is it possible to acquire 1 million of spectra in 16/17 minutes. However, this acquisition speed brings other problems. First of the, the strong reduction of Signal-To-Noise ratio (SNR). Regarding this, a chapter of this thesis is dedicated to (see chapter 4). Regardless the acquisition speed, the data dimension is a challenge in LIBS imaging analysis, with hundreds of thousands or even millions of spectra. This amount of data is not easy to manage, not only in terms of data storage, but also in terms of the power of calculations required. Depending of the research question, there can be different approaches to reduce the dimensionality. Feature selection approaches or dimensionality reduction approaches can be useful to reduce the spectral dimensionality. However, this can be not enough. For example, in the case of clustering for a huge dataset, the high number of spectra is problematic, as showed by Duponchel et al. [18]. This shows how borrowing from other fields can help analysis.

1.3 Chemometrics

Chemometrics is the field of analytical chemistry that uses statistics, and more recently artificial intelligence (AI), to solve complex problems in chemical data analysis and interpretation. Contemporary technologies enable the acquisition of substantial volumes of data within a comparatively brief timeframe. Effective management of this data is then necessary. Having a lot of data is like having a lot of Lego: analyzing a sample without chemometric tools gives a fairly limited view - like identifying that certain Lego bricks are present - whereas with these tools, it is possible to highlight specific associations between the bricks, helping us better understand the complexity of the sample (Figure 8).

After a spectroscopic analysis, for each sample a spectra is given. In imaging spectroscopy, a pixel is a spectrum. A spectrum it can be represented as a vector:

$$X = [x_1,\ x_2,\ ...,\ x_p] \tag{3}$$

Figure 8: Visual representation of the data and the importance of chemometrics. Without chemometric tools even with good data (left) it is not possible to really understand the complexity of the sample, or to appreciate inside structure among them (right)

each value of the vector X is the intensity acquired at specific λ (or m/z e.g.). The wavelength is also a vector:

$$\lambda \ = [\lambda_1, \ \lambda_2, \ ..., \lambda_p] \quad \text{or} \quad m/z \ = [m/z_1, \ m/z_2, \ ..., m/z_p] \tag{4}$$

Considering n samples with p spectral variables, X is defined as a matrix:

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \vdots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}$$

Now is it possible to plot the spectra using the λ (or m/z) as X and the intensities as Y. This is the way a chemist usually thinks of a spectrum: figure 9

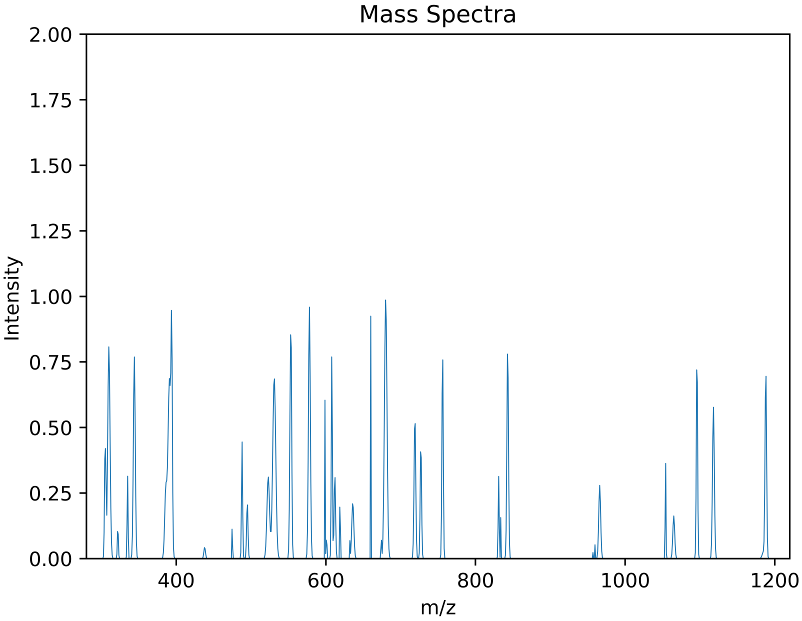


Figure 9: A Mass Spectrum representation, simulated.

This is just one representation of a spectrum. From a chemometric point of view, a spectrum is also a point in a multidimensional space where each variable corresponds to a dimension. To illustrate this, consider the following example of 5 spectra with only two m/z values. They are plotted as shown in figure 10 below
Figure 10 (B) shows 5 different mass spectra with 2 m/z values. Considering each m/z as a dimension, the spectra can be also represented as showed in (A). Representation in real-world data (thousands of m/z) is impossible. having explained the chemometric representation of the spectra is explained, it is important to see what can be done with the data.
Chemometrics can be summarized in three steps:

1) **Pre-processing:** This is the first step in data analysis. It is important to clean the data and remove unwanted information that may interfere with the relevant information. It includes algorithms for outlier detection, normalization, baseline correction, denoising, and dimensionality reduction.

2) **Processing:** The heart of the analysis. Once a suitable method for the problem has been chosen, it can trained on the data. Includes algorithms such as regression, classification and clustering. Other examples are exploratory

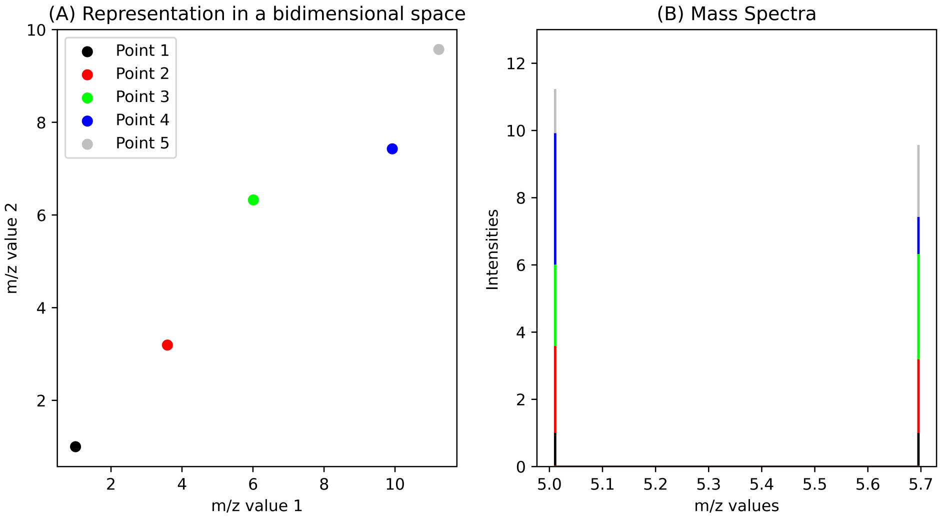


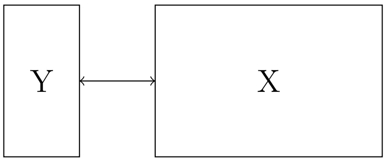
Figure 10: Spectra representation in a bidimensional space

analysis such as PCA.

3) Post-processing: Final stage of the analysis. The results are interpreted, validated and and properly presented. It focuses on validation, visualisation and interpretation.

A classification of approaches can be made taking into account the information available on the data and the purpose:

Supervised analysis: There is prior knowledge about the samples, such as labels or specific properties. The data and information are used to build models that can be used to predict the unknown samples to predict their properties or labels. The predictions are made based on the relationships learned during the training process.



Unsupervised analysis: There is no prior knowledge of the samples. The main aim is to understand the samples and the relationships between them, as well as relations between variables.

There are many algorithms in chemometrics that can be used directly to address the challenges. However, it is important to continue research to improve the known techniques or to implement new approaches. This can be done by bringing in well-established techniques from other fields, as has been done by Duponchel et al. [18] and Alexandrov T. et al. [19]. Chemometrics can then be divided into two main areas: the application of existing chemometric methods and the development of new chemometric methods. The latter aims at developing innovative methods to address and solve new challenges, such as the ever increasing size of data. It may also involve tackling old problems and trying to find new methods to solve them, with robust results and easier interpretation. In this area, it is also important to integrate machine learning and deep learning techniques that can help. It is important for industries to be able to solve or monitor situations with existing methods, but they are still interested in new methods that may be cheaper or have a lower error rate, allowing better control and lower costs. In the medical field, it is important to speed up the interpretation of biological samples in order to improve healthcare.

Chemometrics for this thesis

Improving the extraction of information from unknown samples has been the main objective of this thesis, then the unsupervised methods are more interesting. As mentioned in the section on MALDI imaging, in the case of tissue analysis, the ability to extract information from tissue, to identify tumour and non-tumour tissue by tissue without a priori knowledge, and also to identify intermediate stages, is crucial to the study and understanding of tumours. In particular, clustering techniques have been investigated with the aim of combining method optimisation with biological validation. It is important to have a method that is not only optimal and significant from a statistical point of view, but also easy to read and as real as possible from a biological point of view. This introduction will cover clustering from the basics to state-of-the-art clustering in MALDI imaging.

The work in this thesis has touched on topics other than just clustering, but with the ultimate aim of increasing useful information that can help in the clustering process. The work can be divided into two macro topics: spatial extraction and hyperspectral data fusion. As mentioned above, LIBS is a powerful technique for elemental analysis. However, as will be explained later, sometimes the information contained in these images is not sufficient. Complementary HSI techniques, such as

MALDI or even just RGB images, can enhance the exploration of unknown samples. This section presents the basics of the algorithms used.

1.3.1 Pre-processing

Preprocessing is an essential step in chemometrics. The purpose of preprocessing is to remove unwanted information from the data that can interfere with the important one and affect the models. The raw signal can be defined as

$$S(x) = k \cdot s(x) + \varepsilon(x) + \beta(x) \tag{5}$$

where k is the multiplicative effect, $s(x)$ is the signal of interest, $\varepsilon(x)$ is the error (random noise) and $\beta(x)$ is the baseline effect. These artefacts can be due to instrumental (electronic), environmental (chemical or physical) or operator artefacts. They can all be reduced as much as possible, but they can never be eliminated. In addition to the noise, baseline shifts and multiplicative effects already described, there is another effect that can occur: peak shifts. The origin of this effect can be different. Usually in MALDI imaging peak shifts can be caused by the sample properties (local composition, thickness, matrix deposition) and the calibration method are sources of mass shift[20]. In LIBS imaging, peak shifts can be caused by the Stark shift[21].

1.3.2 Clustering

Clustering is an unsupervised technique that combines samples into homogeneous groups. The samples in a cluster must be as similar as possible, but as dissimilar as possible to the samples in other clusters. It uses the similarity between samples to identify hidden patterns or latent relationships. It works only on the characteristics of the samples, no labelling is required.

For chemical data, clustering divides the data into groups with similar chemical information. As a primary classification, there are four main categories of clustering [22]: (1) Partitioning methods; (2) hierarchical methods; (3) density-based methods; (4) and grid-based methods. Partitioning methods try to find k clusters in n objects. They are distance-based and after an initial partitioning there is an iterative relocation technique that improves the partitioning. This improvement is proved

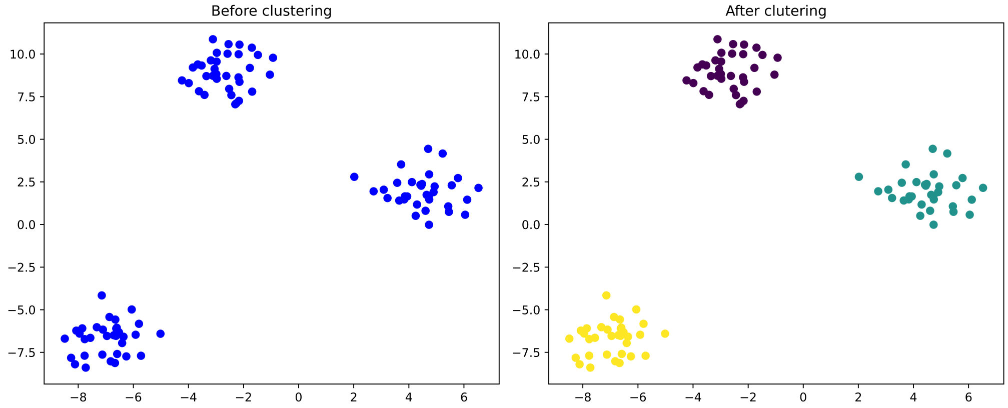


Figure 11: Clustering example

by a function that represents the quality of the partitioning and the algorithm tries to minimise it. Hierarchical methods generate a hierarchical decomposition of the samples, in a *top-down* or *bottom-up* approach. Density-based methods use density as a criterion to generate clusters. If a sample has at least n neighbors within a given radius r, it can be considered a cluster. It is not necessary to specify k in advance. As not all the points are assigned, this algorithm can be used for outlier detection. The main problem is that it depends on the density parameter, which is not easy to determine. Grid-based method divides the space into cells based on a grid structure, and it perform a clustering for each cell. This allows efficient handling of large dimensional data. Among the different clustering methods, kmeans and hierarchical clustering will be presented and explained in more detail as part of the thesis.

Before doing so, it is important to define how spectra are considered similar, so the similarity measure is defined. The distance in space where the spectra are represented can be thought of as a measure of similarity (or dissimilarity). Consider the example in the figure 12. It is possible to distinguish three groups (characterised by different colours). A-B, C-D, E-F are close to each other. This small distance in coordinates, where each axis refers to the wavelength (or m/z), means that the chemical information within them is not so different. On the other hand, these three groups are far from each other in this space, and it means that also the chemical composition is different.

When there are thousands (or more) of variables, even relevant differences in a few dimensions can be masked by the others, making samples look similar when they

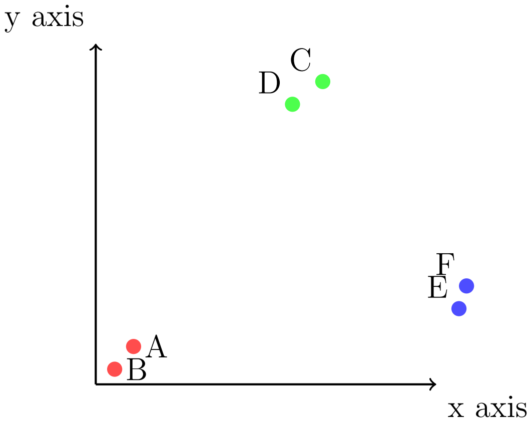


Figure 12: Example of 6 samples that correspond to 3 groups: A and B, C and D, E and F.

are not (the curse of dimensionality[23]). There are different way to calculate the distance In this work, the correlation distance was used because it is independent of normalization, as it is shown by the following equations (NB $a, b > 0$):

$$\begin{aligned}\rho_{aX,bY} &= \frac{\mathrm{Cov}(aX, bY)}{\sigma_{aX}\sigma_{bY}} \\ &= \frac{ab \cdot \mathrm{Cov}(X, Y)}{|a| \cdot \sigma_X \cdot |b| \cdot \sigma_Y} \\ &= \frac{ab}{|a||b|} \cdot \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \\ &= \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \\ &= \rho_{X,Y}\end{aligned}$$

1.3.3 kmeans

The kmeans algorithm is a well-known partitioning clustering algorithm. It is centroid based, which means that it uses a centroid to represent the clusters. The centroid can be thought of as the barycenter of the cluster and can be calculated in several ways. The most common way is to calculate the centroid as the mean of the elements in the clusters (*kmeans*). However, it is also possible to calculate it as the median (*kmedian*) or with the medoid (*kmedoid*) of the elements in the clusters. In kmedoids clustering, the medoid is the data point within a cluster that has the minimal total distance to all other points in the same cluster. The latter two are more robust, less sensitive to outliers and, especially the latter, more representative, but more computationally expensive than kmeans. The algorithm tries to find the solutions with the lowest overall error, expressed in the equation 6, where $M_i^{\{k\}}$ is the sample i in cluster k and $G^{\{k\}}$ is the centroid of cluster k.

$$ERR = \sum_{k=0}^{K} \Big(\sum_{i \in I_k} ||M_i^{\{k\}} - G^{\{k\}}||^2 \Big) \tag{6}$$

The algorithm is explained by the table 7 below. There are different problems with this algorithm. First at all, it does not give global optimum, but the results depends strongly on the initial estimation of the centroids. This can be solved by multiple running of the algorithm, and then keep just the best result. *kmeans++* [24] algorithm was developed to improve the initialisation, but it still need multiple running. kmeans assumes equal variance among all the dimensions, so spherical clusters are predilected. This cause problems in real samples, where the clusters has usually a different shape. Additionally, the criteria is overall error minimization, and this cause unbalanced clusters to be not well described. Last but not the least, k number of clusters must be provide in advance, as implicit assumption due to the algorithm itself. Because the k number of cluster is not always available, multiple running and optimization step are required, to find the optimal number of k in the sample that is analyzed. This process is done with the quality cluster criteria, that

Algorithm: kmeans

INPUT	**METHOD**	**OUTPUT**
D dataset k number of clusters	1) Generate k centroids 2) Calculate the distance between the points and the centroids 3) Assign each points to the cluster of the nearest centroids 4) Calculate the new centroids as the mean of the spectra in the cluster 5)If the label are not changed from the previous iteration, then stop the calculation, otherwise repeat points 2-4	k clusters

Table 7: Overview of kmeans algorithm

will be explained in the next session.
Table 13 shows how to apply kmeans in matlab.

Matlab

```
[idx,C] = kmeans(x, k, ...
           'Distance', 'Correlation', ...
           'MaxIter', 800, ...
           'Replicates', 30);
```

Figure 13: Kmeans script for matlab. It can work with squared euclidean, cityblock, cosine, correlation.

1.3.4 Quality Cluster Criteria

In the majority of cases where clustering algorithms are used, it is not possible to determine the exact number of clusters in advance. Consequently, it is not possible to apply an external criterion to validate the results. Therefore, an alternative approach to validate the results is required, or more precisely, to compare the results and propose the one that is statistically stronger. In the field of clustering, this can be achieved by using quality clustering criteria. Quality cluster criteria refer to algorithms that are able to generate a "quality index" for a given data partitioning. The validation of a data partitioning without external validation, e.g. without tagging the data, is called internal validation.

The internal validation criteria use the characteristics of the clusters, mainly compactness and separation, but also density and overlap, to give this quality index, which is used to compare the results. Figure 14 shows a visual representation of compactness and separation.

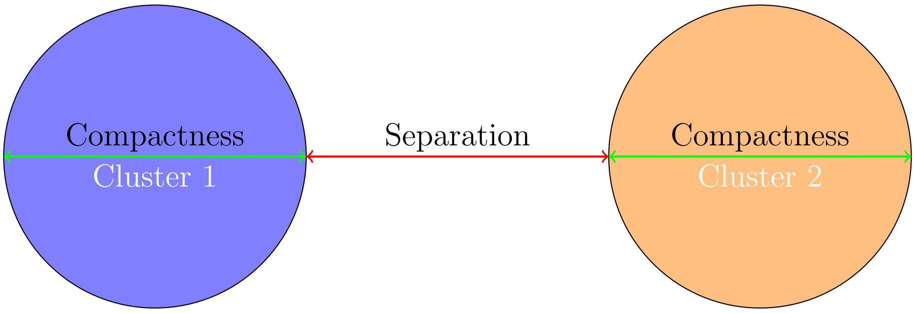


Figure 14: Visual representation of Compactness and separation

In cases where the optimal number of clusters is not known in advance, a method is required to calculate the results of the kmeans algorithm with values of k ranging from, e.g., 2 to 10. The quality cluster criterion can then be applied to each resulting partitioning. The resulting segmentation map for each data set will then have a "quality" index. The segmentation map with the higher quality is then selected as the most probable partitioning.
As mentioned above, the quality cluster criteria are based on the measures of compactness, separation, density and overlap. Not every criterion uses all of these measures. There are many criteria that use only compactness and separation. A comprehensive review of the index has recently been published, showing the comparison of 68 quality cluster criteria [25].
However, there is an important point to be made about these criteria. A review of the measurements used (compactness, separation, density and overlap) shows that these criteria cannot be applied when k=1. Although this is generally not important as normally when clustering algorithm are applied we are not looking to observe one cluster among the data, so we want to see more than one cluster, it can be disadvantageous, as will be shown in Chapter 1: Automation (2.2). In this section the reader will see how this issue has been addressed in the course of writing this thesis. The following three criteria will be discussed and explained: PBM [26], Silhouette [27] and Calinski-Harabasz [28]. These criteria played an important role in the present thesis.

Calinski-Harabasz The Calinski-Harabasz index is expressed by the equation 7

$$C = \frac{N-K}{K-1} \cdot \frac{BGSS}{WGSS} \tag{7}$$

where BGSS is expressed by the equation 8 and WGSS is expressed by the equation 9

$$BGSS = \sum_{k=i}^{k} n_k ||G^{\{k\}} - G||^2 \tag{8}$$

Equation BGSS is the sum of the weighted sum of the squared distances between the centroids $G^{\{k\}}$ and the global centroid G.

$$WGSS = \sum_{k=0}^{K} WGSS^{\{k\}} = \sum_{k=0}^{K} \Big(\sum_{i \in I_k} ||M_i^{\{k\}} - G^{\{k\}}||^2 \Big) \tag{9}$$

Equation WGSS is the sum of the squared distances between the observation $M_i^{\{k\}}$ and the centroid of the cluster $G^{\{k\}}$.

Pakhira, Bandyopadhyay, Maulik (PBM) The PBM index is expressed by the equation 10

$$C = \Big(\frac{E_t \cdot D_B}{K \cdot E_w}\Big)^2 \tag{10}$$

where E_T is expressed by the equation 11, E_w is expressed by the equation 12, and D_B is expressed by the equation 13

$$D_B = \max_{\mathrm{k<k'}} d(G^{\{k\}}, G^{\{k'\}}) \tag{11}$$

D_B is the largest distance between two cluster centroids.

$$E_w = \sum_{k=i}^{K} \sum_{i \in I_k} d(M_i, G^{\{k\}}) \tag{12}$$

Equation E_w is the sum of the distances of the points of each cluster to their centroid

$$E_T = \sum_{i=1}^{N} d(M_i, G) \tag{13}$$

Equation E_t is the sum of the distances of all the points to the centroid of the entire dataset

Silhouette The Silhouette index is expressed by the equation 14

$$C = \frac{1}{K} \sum_{k=1}^{K} s_k = \frac{1}{K} \sum_{k=1}^{K} \Big(\frac{1}{n_k} \sum_{i \in I_k} s(i)\Big) \tag{14}$$

where $a(i)$, $b(i)$, and $s(i)$ are expressed by the following equations:

$$a(i) = \frac{1}{n_k - 1} \sum_{\substack{i' \in I_k \\ i' \neq i}} d(M_i, M_i')$$

$$b(i) = \underset{k' \neq k}{min} \Big(\frac{1}{n_k - 1} \sum_{i' \in I_{k'}} d(M_i, M_{i'})\Big)$$

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Equation $a(i)$ is the mean distance of the point M_i to the other points of the cluster it belongs to. Equation $b(i)$ is the smallest of the mean distance of the point M_i to the points of eachof the other clusters. Finally, equation **??** is the silhouette quotient for each point. The global silhouette index, equation 14, is the mean of the mean silhouettes through all the clusters.

Considerations

The table 8 gives the final equation of the index.

	Calinski-Harabasz	**PBM**	**Silhouette**
	$C = \frac{N-K}{K-1} \cdot \frac{BGSS}{WGSS}$	$C = \left(\frac{E_t \cdot D_B}{K \cdot E_w}\right)^2$	$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$
COMPACTNESS	$WGSS$	E_w	$a(i)$
SEPARATION	$BGSS$	D_B	$b(i)$

Table 8: Compactness and separation for each index

For the silhouette, instead of the global silhouette, the silhouette coefficient for each point is given for convenience. For each of these, the optimal value is the highest in the list. Looking at the structure of the equation, we can see that maximizing the separation and minimizing the compactness influence this. However, as mentioned above, the minimum number of clusters required to apply this algorithm is 2.

As said previously, in this thesis clustering was performed using kmeans with correlation distance. While kmeans is traditionally associated with Euclidean geometry, correlation distance offers a shape-based similarity measure, particularly suited to high-dimensional, normalized data such as mass spectra. To evaluate clustering quality, it was employed the silhouette index. Despite its common association with Euclidean distances, the silhouette can be computed using any dissimilarity measure, including correlation-based ones, provided that the triangle inequality is not essential for interpretation. The main advantage of the silhouette lies in its interpretability at multiple scales: it yields a global summary of clustering performance, while also enabling the identification of poorly clustered data points via negative or near-zero values. Compared to centroid-based indices such as Calinski-Harabasz or PBM, the silhouette does not assume convex cluster shapes and is more flexible with respect to the underlying geometry of the data. This makes it particularly useful in applications like MALDI imaging, where cluster boundaries may not be strictly globular or equally dispersed. Moreover, since correlation distances tend to reduce the influence of absolute intensity and highlight relative peak patterns, using the silhouette allows to assess how well such relative profiles are grouped without enforcing rigid geometric constraints.

1.3.5 Hierarchical clustering

Hierarchical clustering algorithms organize data into a hierarchical structure. The approach can be done in two different ways, which are common in other analytical problems:

Top-Down: As the name suggests, the direction goes from ”top” to ”bottom”. Starting from the complete dataset, it is progressively split into smaller and smaller subsets. This approach is also known as *divisive hierarchical clustering.*

Bottom-Up: As the name suggests, the direction goes from ”bottom” to ”top”. Starting from individual samples, they are iteratively merged into increasingly larger groups until the entire dataset is unified. This approach is also known as *agglomerative hierarchical clustering.*

Even if agglomerative algorithms are used in the MALDI imaging domain [29], the divisive algorithm will be more explored and discussed. Theoretically, both divisive and agglomerative algorithms are more expensive than flat clustering, considering all possible divisions ($\mathcal{O}(2^n)$ vs. $\mathcal{O}(n^2)$). However, in the case of the divisive hierarchical algorithm applied to MALDI imaging, deeper levels are usually not reached, so the computational cost decreases, making it more convenient compared to agglomerative approaches, where distances between all samples must be calculated (which some computers may not be able to handle). Additionally, as will be shown later, during the splitting step of a divisive partitioning algorithm, clustering methods such as kmeans can be used, making the algorithm more efficient [30].

Chapter 2

Pushing the boundaries of clustering in MALDI imaging

One of the most common and straightforward strategies in MALDI imaging studies is the spatial visualization of individual ions or groups of ions. However, this is a targeted analysis as it requires prior knowledge of the most relevant ions. While this approach may have been valuable in the early stages of MALDI imaging, it does not exploit the full potential of the technique. In fact, MALDI imaging can also be used for biomarker discovery, identification, and unsupervised image segmentation.

State-of-the-art

Clustering in MALDI imaging is important because of its ability to group similar spectra together. Considering a pixel as a sample, a clustering algorithm can group pixels with a similar biochemical composition. By labelling these groups, it is possible to print a map (spatial segmentation map) showing the similarity of the different zones of the tissue. The spatial segmentation map can also be interpreted using the *m/z* values. Each cluster can be represented as its centroid, e.g. the mean spectrum of the pixels forming the cluster. The goal of classical clustering research in MALDI imaging is to retrieve the biologically relevant regions highlighted by histological annotation. However, segmentation based on MALDI imaging can potentially provide more information compared to histopathological analysis or immunohistochemistry. In fact, the latter two are target analysis techniques, so they are based on a reaction and the results are based on human perception, so they are not as sensitive to a small variation in the reactants involved and are blind to variation in the reactants not involved. MALDI imaging gives the full biochemical information of the sample, it is possible to delve deeper into the differences between tissues, leading to the possibility of extracting areas with a small difference that cannot be separated from the normal analysis. There are many different clustering techniques used in MALDI imaging, each with its own advantages and disadvantages [31]. Among these, *bisecting kmeans* plays an important role [30].

An interesting review [32] is focus on the unsupervised algorithm for exploratory analysis in MSI, and present better explanation about the different clustering techniques and the application in MALDI imaging. Among them, bisecting kmeans had an important role. Bisecting kmeans is a divisive hierarchical clustering algorithm where each division is done with kmeans algorithm with *k=2.* The first application of this algorithm in MALDI imaging is reported in [33]. This algorithm is explained in more detail in the chapter 2.1. Another clustering algorithm used in MALDI imaging domain is High-Dimensional Data Clustering (HDDC). The goal of HDDC is to overcome limitations of classical algorithms such as kmeans, which assume equal variance in every dimension [34]. HDDC relies on a Gaussian Mixture Model (GMM) framework that allows each cluster to have its own orientation and variance structure, making the method more suitable for high-dimensional datasets. Its first application in MALDI imaging was reported by Alexandrov et al. [35]. The use of spatial information for clustering in MALDI imaging is discussed in more detail in the chapter 2.3. The methods presented so far are defined as *hard clustering* because the pixel can only be assigned to one cluster. This can be reasonable because the pixels are not Schrodinger's cat, but sometimes it can be interesting to look at the similarities between pixels and the different clusters to which they are not assigned. A real case can be where a pixel is a mixture of chemical properties from two different clusters. In this case, assigning it to only one cluster may not reflect the true complexity of the sample. There are techniques that assign a pixel to multiple clusters. These techniques are known as *soft clustering.*

Unlike the hard clustering techniques, the result of a soft segmentation technique is not a single segmentation image, but different maps with different probabilities of a particular cluster. The interpretation is more complicated, but it can provide a broader view of the data.

An example is *Fuzzy c-means clustering* [36]. It is similar to kmeans as it tries to minimize the following function:

$$ERR = \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ik}^{m} ||x_i - G_k||^2 \tag{15}$$

where: N is the number of samples; K is the number of clusters; x_i is the i-data point; G_k is the centroid of cluster k; u_{ik} is the membership degree of x_i in cluster k $m > 1$ is the fuzziness parameter, controlling the level of cluster overlap. The goal is to find both the optimal cluster centroid and membership values.

The algorithm works as follows: (1) Randomly assign membership values u_{ik}; (2) Update centroids; (3) Update membership values; (4) Repeat 2-3 until the convergence is reached. The convergence in fuzzy clustering is usually a threshold on the value expressed by the equation 15. The membership values are updated at each iteration according to the equation 16

$$u_{i,k} = \frac{1}{\sum_{j=1}^{K} \left(\frac{||x_i - G_k||}{||x_i - G_j||}\right)^{\frac{2}{m-1}}} \tag{16}$$

The centroids are updated as the weighted mean of the data points

$$G_k = \frac{\sum_{i=1}^{N} u_{ik}^{m} x_i}{\sum_{i=1}^{N} u_{ik}^{m}} \tag{17}$$

Other soft segmentation techniques are AMASS (MSI analysis by semisupervised segmentation) [37](Bruand et al. (2011a)), Latent Dirichlet Allocation[38] (Chernyavsky et al. (2012)) and Spatial Shrunken Centroids (Bemis et al. (2016)) [39].

2.1 Bisecting vs QDD kmeans

The aim of this chapter is to present and explain the first part of this thesis. Learning that bisection kmeans is one of the most common and popular algorithms in the MALDI imaging domain, the study of its popularity was interesting. However, while exploring the algorithm, some potential problems came up. For this reason, the focus was on trying to overcome these problems. The next chapter 2.2 deals with the continuation of this work. As introduced earlier, bisecting kmeans is a divisive hierarchical clustering algorithm where each division is done with kmeans algorithm with *k=2*. In some software, this algorithm is interactive, allowing the user to easily move through the hierarchical tree. However, the partitioning does not always consider statistical parameters. In fact, it is driven more by histological annotation, if available, or by user experience. This was the first problem found. Using the histological annotation or the user experience could limit the analysis, e.g. not dividing a cluster that should be divided or dividing what should not be divided, just because of some bias. So the need to use a true unsupervised hierarchical divisive clustering algorithm was quite interesting from our point of view. This way, all potential divisions would be tested and validated by statistical criteria or a metric, e.g. quality index. However, this is better explained in the following chapter 2.2, as it is related to the stop criteria and was not easy to face in the first instance. In fact, there was a second relevant problem found in this approach: for the division into two subclusters. For unbalanced clusters, it can be proved that hierarchical clustering algorithms are more suitable compared to the original *kmeans*, but this does not mean that the division by 2 is always the optimal choice. This can lead to some biased conclusions. Consider the toy example has to be described in figure 15 As shown in the figure, forcing the division into two sub-clusters can lead to artifacts in the final segmentation. This can be clearly seen in a toy example like this, but in a real sample, where the true structure of the data is unknown and it is not possible to properly represent the samples in a lower-dimensional space, this can lead to mislabeling of pixels and different interpretations of the data, which can have a negative effect in the case of cancer tissue exploration.

Quality-Data-Driven kmeans (QDD-kmeans) was developed during this PhD to overcome the potential artifacts arising from the bisection nature of kmeans. QDD kmeans is a divisive hierarchical clustering algorithm where each split is done with kmeans algorithm but the number k is not fixed in advance but need a validation step. For each split, the branch that should be divided pass this process:

1. Calculate the results of kmeans with k = *2,3,...,10*
2. For each segmentation map calculate a quality cluster criteria
3. Select the optimal solution and keep that division

Comparing bisection and QDD, the latter can potentially provide the similar results of bisection with less depth or, in the best case, show clusters that have been split by the forced division of bisection and cannot be explored.

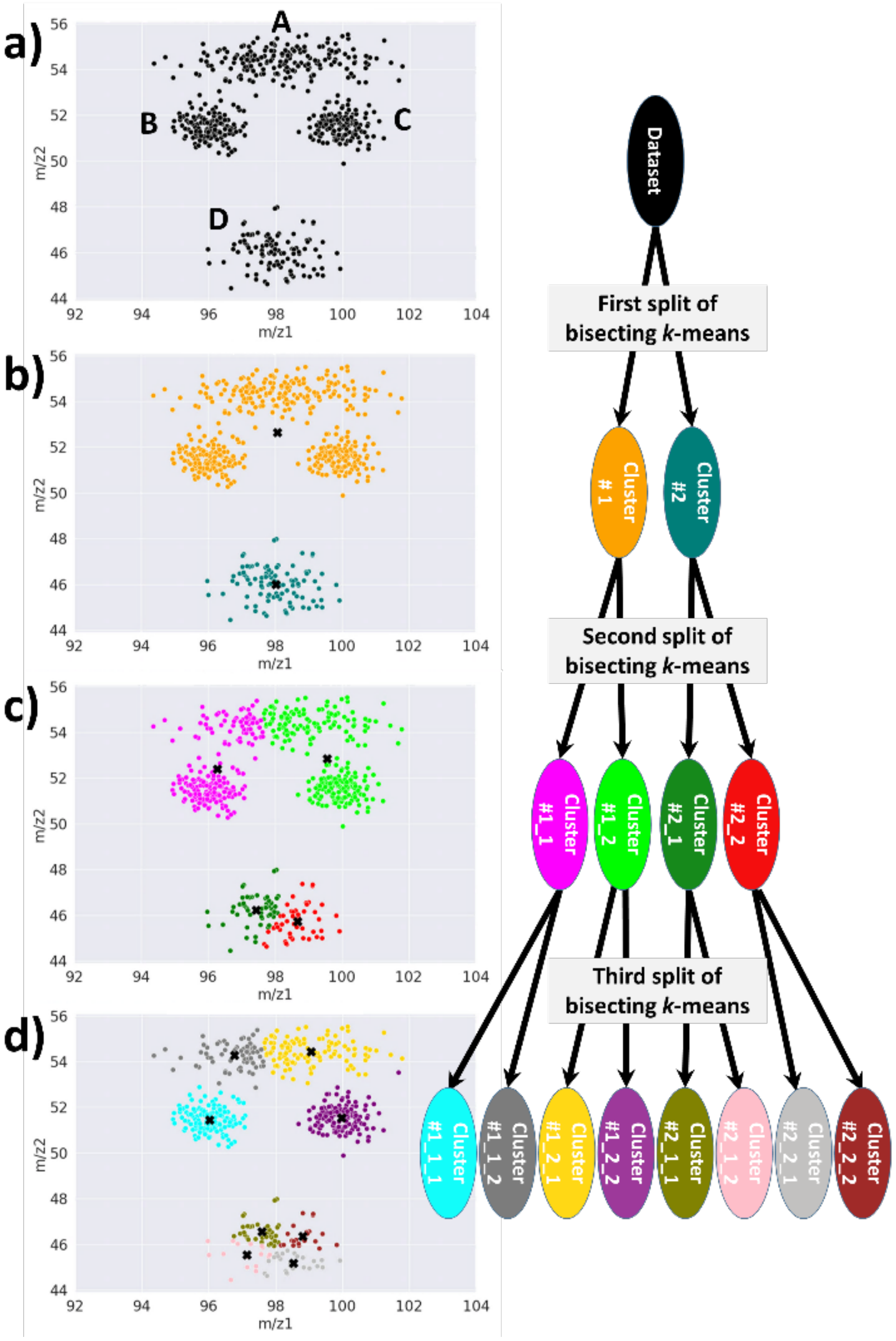


Figure 15: The bisecting kmeans approach. a) Representation of the dataset in the two dimensional data space. b-d) Successive partitioning of the dataset by applying a kmeans algorithm with two clusters to each cluster identified in the previous step

2.1.1 Bisecting kmeans vs QDD kmeans

This section presents the application and comparison between bisection and QDD kmeans on two squamous cell carcinoma samples. The approach was to systematically partition level by level (e.g., systematically partition each branch) and compare the results obtained. The results were compared in two ways to be as fair as possible: **level by level** and **total number of clusters by total number of clusters**. In the first case, the results are compared considering the same level of the tree to see the ability to reach deep information. In the second case, the results are compared considering more or less the same number of clusters. Comparing a spatial segmentation map with a large difference in the number of clusters is unfair because it doesn't allow the algorithm to display the same information.

A few considerations can be made beforehand. For bisecting kmeans, the number of possible clusters is 2^n. So for the first level there are 2 clusters, for the second level each cluster is split into two new clusters, for a total of 4 clusters. For the third level there are 8 clusters, and so on. Embedded is a different story. As mentioned above, for each split there is an optimization step regarding the number of clusters. So there is no limit to the number of clusters that can be obtained with this approach of systematically splitting level by level.

The optimization step is crucial for QDD kmeans. Finding the optimal number of clusters has always been a challenge in clustering. As we said in the section Quality Cluster Criteria (1.3.4), there are many different criteria to define the quality of a segmentation. The final work reported in this thesis uses only one of them. However, several approaches were used during the thesis work. The original idea was to use not only one criterion, but a list of them, and to decide by a voting procedure. That is, with a list of 20 criteria, the majority decides the number of clusters. However, if the voting showed really different clusters (e.g. 11 votes for 4 and 9 votes for 8), this could mean that the true value was in the middle, so the possibility of using the rounded mean and not just the mode was considered. The initial list was large ($\sim$ 20 indices), and when some criteria showed strange behavior, they were removed. However, the voting process was complicated and problematic, so in the end it was decided to keep only Silhouette, which is also the more conservative for the structure and because the almost the same results were obtained with two other metrics (PBM and CH).

There is no official Matlab function for bisection and QDD kmeans, so the functions were home-made. An example of the possible hierarchical tree that can be obtained by bisecting kmeans in this case is shown in the figure 16

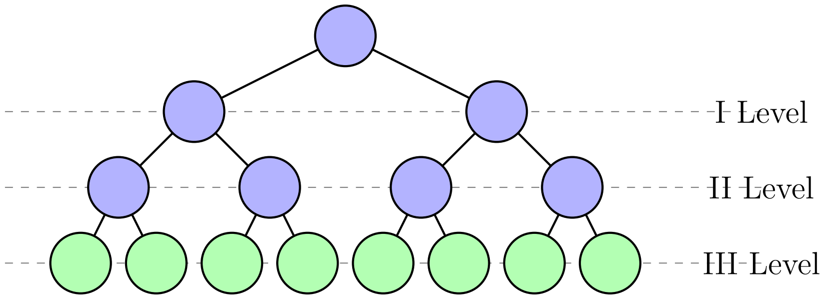


Figure 16: Hierarchical bisecting tree example. The green circle will correspond to the final clusters

2.1.2 Experiment part

The samples used in this study were taken from a previously published investigation of oral tongue squamous cell carcinoma (SCC) [40]. Of all the samples analyzed in the article, only two were focused on, referred to as patients A and C. Representative fresh tissue fragments (2 cm × 2 cm × 0.5 cm) were selected and placed on a pre-cooled specimen disc (−20°C) with a drop of water in a cryomicrotome (Cryostat, Thermo Fisher Scientific CryoStar NX70, MMFRANCE). Two consecutive sections were then prepared: a 7 μm section for H&E staining and a 12 μm section mounted on a conductive glass slide coated with indium tin oxide (ITO) for mass spectrometry imaging (MSI). Figure 17 shows the optical and TIC images of the two samples examined.

Preprocessing

The spectra was primary smoothed using Savitzky-Golay algorithm (window size = 15, second-order polynomial) to reduce noise. Top-Hat algorithm (window size = 200) was used for baseline correction. Spectra alignment was necessary due the comparison of two different datasets from different samples due to instrumental and spectral shifts [8]. The alignment was done starting with the mean spectrum of the two datasets. A peak selection was then applied to this mean spectrum using a local maximum strategy [41]. Among these peaks an interval of ± 0.3 Da for each spectrum was considered: the highest intensity value observed within this interval was assigned as the intensity at the corresponding m/z value.

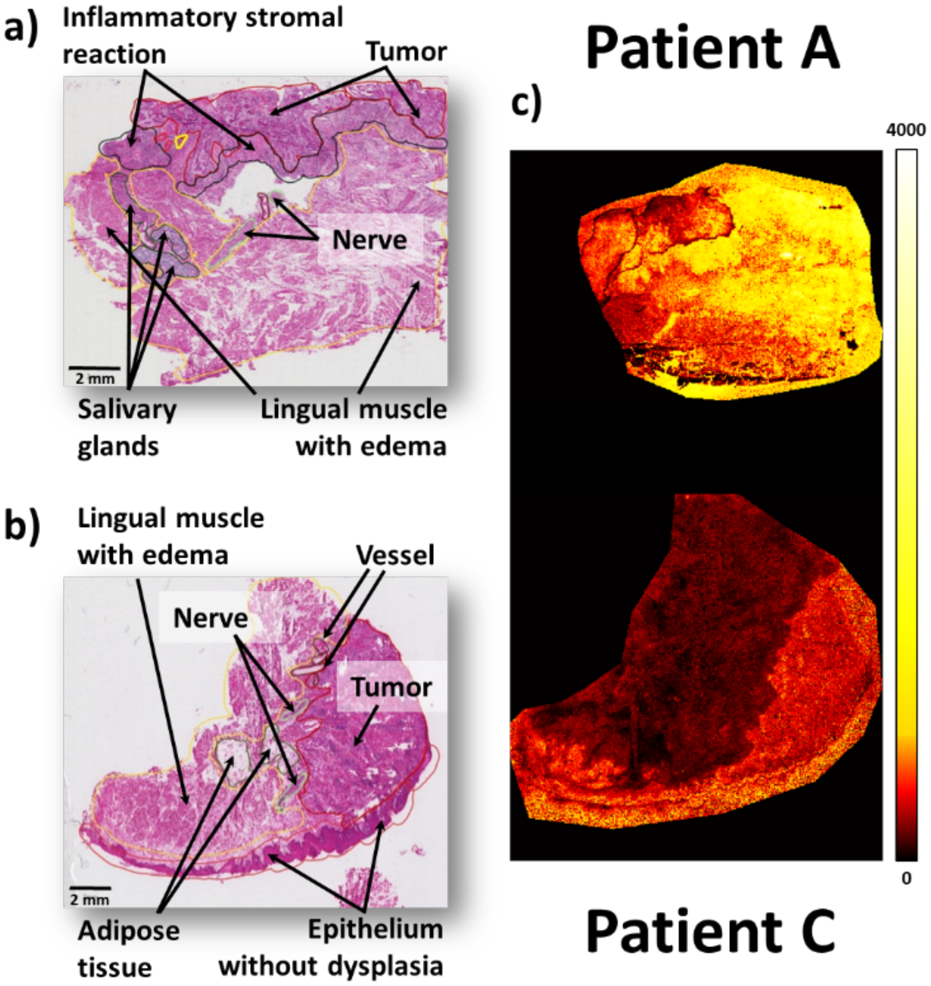

a)
Inflammatory stromal reaction
Tumor
Nerve
2 mm
Salivary glands
Lingual muscle with edema
b)
Lingual muscle with edema
Vessel
Nerve
Tumor
2 mm
Adipose tissue
Epithelium without dysplasia
Patient A
c)
4000
0
Patient C

Figure 17: Sample examinated. a) and b) corresponds to the optimal image. c) corresponding to TIC images generated from the acquired MALDI imaging datasets

Results

The purpose of this section is to highlight the differences that may arise when using bisecting kmeans or QDD kmeans in the context of unsupervised exploration of MALDI imaging data. Since the biological tissues of two patients with tongue cancer are considered in this study, we adopt a classical machine learning strategy based on pooling all spectra into a single dataset, upon which clustering will be applied.

This approach facilitates the capture of inter-sample variability, enabling the detection of both common and patient-specific patterns. Moreover, aggregating spectral data enhances the robustness of the clustering process by integrating a richer set of information, there-by improving, for instance, the identification of weak signals that might otherwise be obscured in a patient-specific analysis. Given that both considered approaches are inherently hierarchical, their behavior will be compared at each level of partitioning offering insight into their respective drawbacks and advantages. Figure 1a displays the initial results obtained from bisecting kmeans, corresponding to the first level of partitioning (the initial bisection) of the original spectral dataset. Two clusters (designated BIS_LV1_C1 and BIS_LV1_C2) emerge naturally, this number intrinsically defining this approach. This cluster naming convention will be maintained throughout the article: m_LVn_Cp, where m designates the approach used (BIS for bisecting kmeans and QDD for QDD kmeans), n represents the partitioning level, and p denotes the cluster number within that level. A swift comparison of Figure 19-a with Figure 17 suggests that the BIS_LV1_C1 cluster (in pink) is primarily associated with biological tissue, while the BIS_LV1_C2 cluster (in blue), located on the exterior, is most likely linked to the MALDI matrix. The application of the QDD kmeans approach to obtain an initial level of partitioning on the same dataset requires determining an optimal number of clusters, kopt. To this end, the kmeans algorithm is applied to the dataset for a number of clusters ranging from 2 to 10. The silhouette score is then computed for each clustering configuration. Figure 19-c illustrates the evolution of this quality metric, with the maximum silhouette score indicating the optimal number of clusters to be used in the partitioning process — 3 in this specific case. The figure 19-b displays the clustering map obtained for this set of three clusters. The QDD_LV1_C1 cluster (in pink), predominantly located at the center of each sample, corresponds to the biological material, while the QDD_LV1_C3 cluster (in blue) represents the MALDI matrix, as previously identified using bisecting kmeans. However, a third cluster, QDD_LV1_C2 (in magenta), is specifically observed in the lower region of patient C's tissue, though only faintly detected in patient A. Upon revisiting the histological images in Figure 17, it is observed that the cluster QDD_LV1_C2 (in magenta) actually corresponds to an area where epithelium without dysplasia is present. Two clusters thus define the biological material here, while a third is associated with the MALDI matrix. This is an opportunity to observe that using two clusters in bisecting kmeans effectively recovers a significant portion of the biological material within cluster BIS_LV1_C1. However, the cluster BIS_LV1_C2 presumed to represent the MALDI matrix also contains spectra from the epithelium without dysplasia.

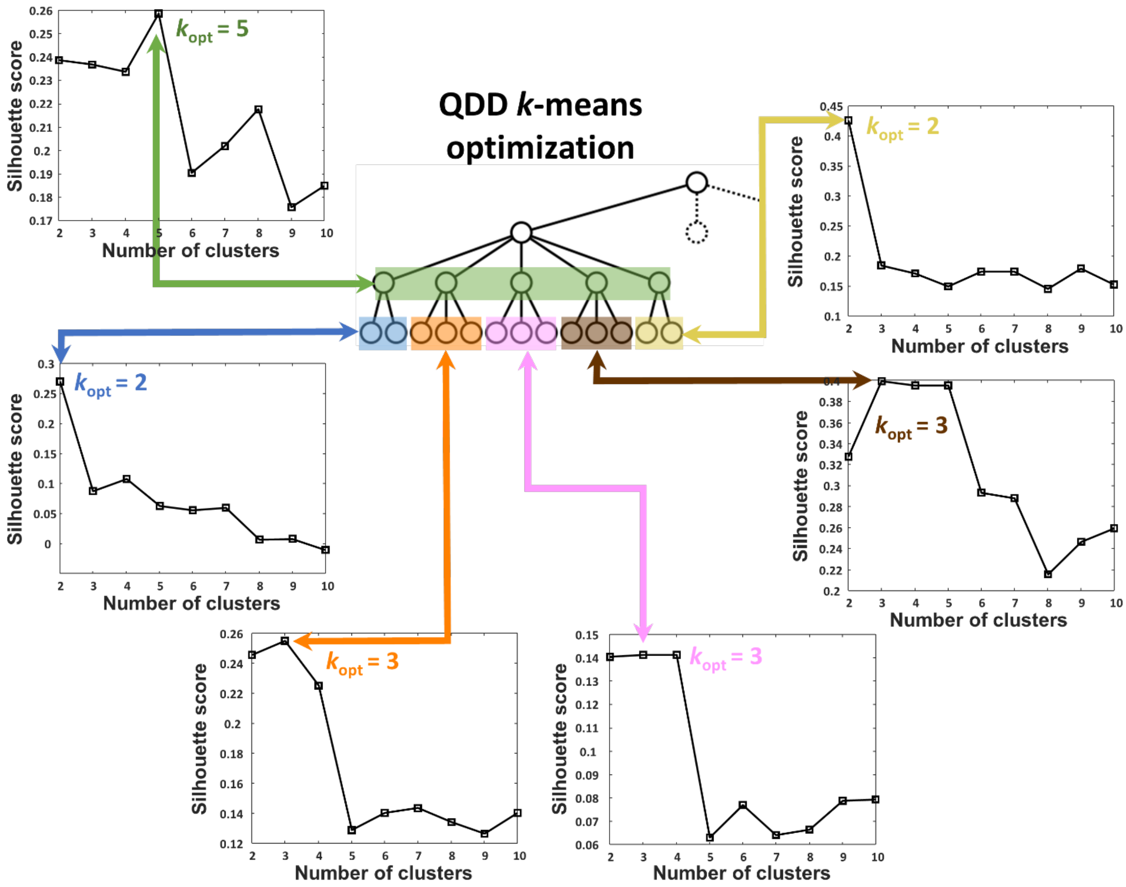


Figure 18: Silhouette scores calculated for each partitioning level and its constituent clusters within the embedded kmeans approach. The optimal number of clusters, k_{opt}, is then determined for each of them.

This observation regarding bisecting kmeans might seem trivial at first glance, as we might assume that the upcoming partitions of cluster BIS_LV1_C2 would inevitably allow us to distinguish between the MALDI matrix and the epithelium. Our attention is not currently directed at this level but rather at a crucial step in the preprocessing of imaging data, namely, the generation of a mask. Indeed, the contour of a tissue section does not have a predefined shape. Our intent to analyze its entire surface leads to acquire data over a broader area than the tissue itself because the area of acquisition is manually defined by the user in the control interface of the instrument.

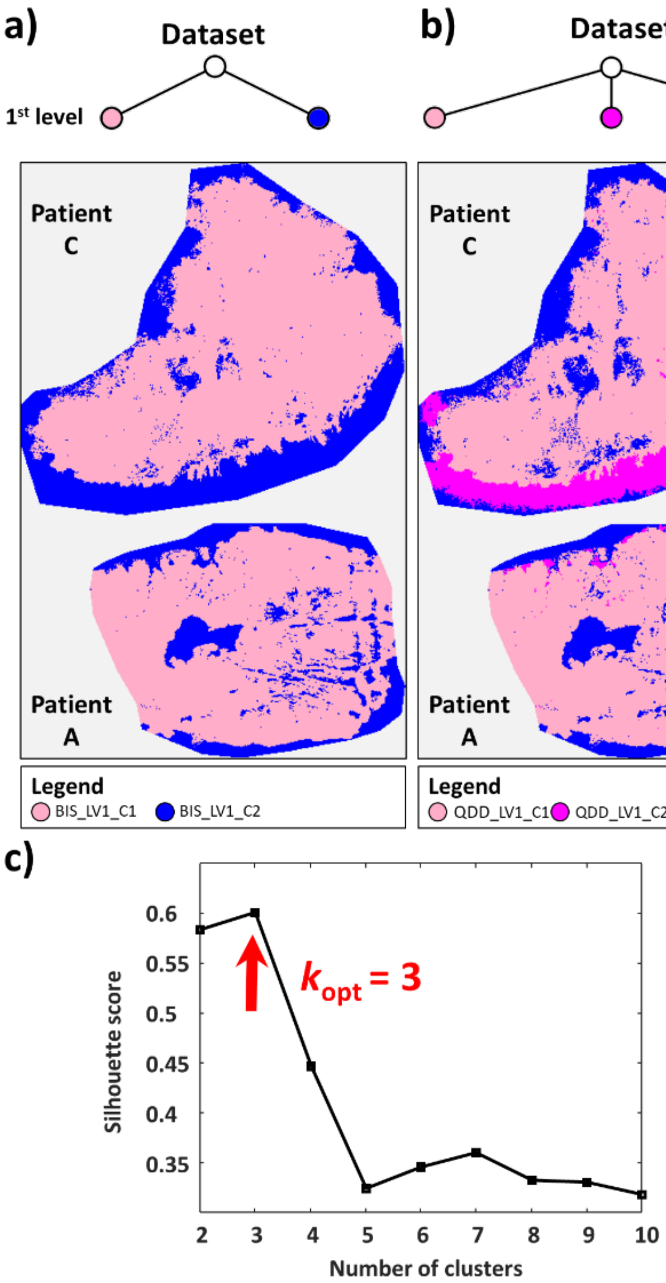


Figure 19: The first partitioning level of clustering. a) Clustering map obtained with bisecting kmeans while considering this level. b) Considering the QDD kmeans approach that optimizes the number of clusters. c) Silhouette score curve obtained on the dataset highlighting the optimal number of clusters.

The dataset thus acquired includes spectra of the biological material as well as spectra devoid of it, containing, for instance, only the MALDI matrix. It is unfortunately observed that the variance introduced by these matrix spectra significantly complicates the detection of biologically relevant variations. To ensure optimal conditions for biological observations, it is essential to generate a mask on the spectral dataset—an automated method that selectively retains only biologically derived spectra. Observing Figure 19, it becomes evident that clustering can be leveraged to generate this mask, a common practice in spectroscopic imaging. As shown in Figure 19-b, QDD kmeans proves highly effective, as the initial partitioning alone allows us to discard the MALDI matrix while retaining only the spectra of the clusters QDD_LV1_C1 and QDD_LV1_C2. As previously indicated, this does not mean that we could not apply the same procedure with bisecting kmeans; however, additional levels of partitioning would be required, thereby increasing the complexity of the process. Unfortunately, even though the bisection approach appears feasible on paper, it is clear that the community generally only considers the first level of division using bisecting kmeans to generate this mask, which naturally results in a loss of biological information throughout the entire tissue under investigation. That being said, to streamline our observations and focus on the regions of interest within biological tissues, only the spectra from cluster BIS_LV1_C1 (for bisecting kmeans) and QDD_LV1_C1 (for QDD kmeans) have been retained for the remainder of the study. Strictly speaking, the in-depth exploration of these biological tissues begins at the second level of partitioning. One could even say that, from this level of partitioning onward, increasing differences can be observed between bisecting kmeans and QDD kmeans. The results of the bisecting kmeans approach for the second level of partitioning are presented in Figure 20-a. Two new clusters (BIS_LV2_C1 and BIS_LV2_C2) naturally emerge, as they result from the bisection of the single cluster obtained at the previous level (e.g. BIS_LV1_C1). When compared to histological images, this partitioning appears very coarse, inevitably distorting the perception of the various biological structures. Thus, the BIS_LV2_C1 cluster (in red) simultaneously contains both zones with tumors and lingual muscle with edema for both patients. The second level of partitioning was also performed using the QDD kmeans approach on the spectra of the previous cluster (e.g. QDD_LV1_C1). Silhouette score analysis demonstrated that the optimal clustering of this previous cluster required considering five clusters.

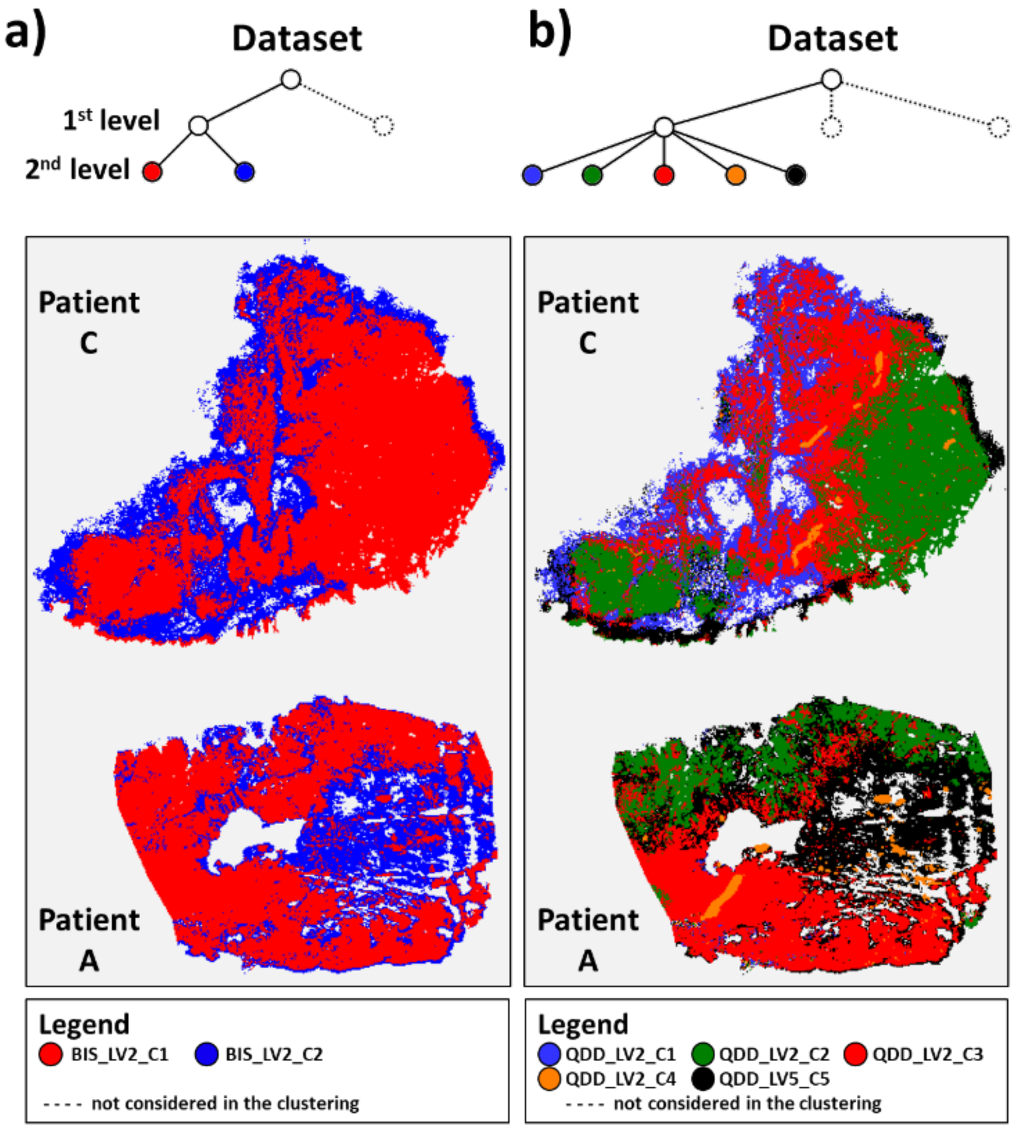


Figure 20: The second partitioning level of clustering. a) Clustering map obtained with bisecting kmeans while considering this level. b) Considering the QDD kmeans approach that optimizes the number of clusters.

The results of this new partitioning are presented in Figure 20-b. This partitioning thus corresponds quite well to the annotations of the two histological sections. The QDD_LV2_C2 cluster (in green) thus corresponds to the various tumor regions in both patients, just as the QDD_LV2_C4 cluster (in orange) is associated with nerves, and the QDD_LV2_C3 cluster (in red) effectively represents lingual muscle with edema. More precisely, the lingual muscle is also represented by a second cluster, QDD_LV2_C5, for patient A, while its presence is relatively limited in patient C. The fifth and final cluster, QDD_LV2_C1 (in blue), is primarily observed in patient C, surrounding the lingual muscle. Yet, no histological annotation is available to correlate with this cluster which demonstrates that the molecular information obtained by MALD-MSI based on molecular information enables seperating cells which show no difference for the morphological annotation. Identification of the lipid markers specific to this cluster by comparison to the others would be necessary to better understand the cell molecular phenotypes associated. At first glance, the clustering from this second level of partitioning for QDD kmeans might seem sufficient. However, a clearly identified region in patient A histological image, where an inflammatory stromal reaction is expected, is not distinguished in this clustering. In fact, the spectra of this region are included in the QDD_LV2_C3 cluster (in red), which primarily represents the lingual muscle. From our perspective, there is no valid reason why this inflammatory response should not be ob-served in the MALDI-MSI dataset. Consequently, a third level of partitioning using the QDD kmeans approach will be considered later in this study. Taking a broader perspective on these results obtained with two levels of partitioning, one could justifiably argue that comparing the two clustering approaches at a given partitioning level is not entirely fair. The systematic binary partitioning in the bisecting approach appears to mechanically slow down the clustering process in describing the dataset. A third level of partitioning was therefore considered for bisecting kmeans, with clusters BIS_LV2_C1 and BIS_LV2_C2 further sub-divided, resulting in a total of four new clusters. In a way, it could be said that this approach brings the partitioning closer to that obtained at level 2 of QDD kmeans, but the nerve structures remain entirely undetected. Moreover, as in the QDD approach, the inflammatory region is also not highlighted. Thus, the results of the fourth partitioning level from bisecting kmeans, which generated eight new clusters from the four at the previous level, were compared with the third partitioning level of QDD kmeans. For this latter method, computing the silhouette score for each of the five clusters obtained at the second partitioning level allowed the detection of a total of 13 clusters at this third partitioning level, with clusters QDD_LV2_C1, QDD_LV2_C2, QDD_LV2_C3, QDD_LV2_C4 and QDD_LV2_C5 subdividing into 2, 3, 3, 3, and 2 clusters, respectively. Figure 22 thus presents the partitioning results of both approaches. To facilitate comparisons, we attempted in this last figure to use identical colors for clusters from both clustering methods whenever they appeared to cover fairly comparable spatial regions. This was not truly feasible for the previous images due to the significant discrepancy in the number of clusters between the two approaches.

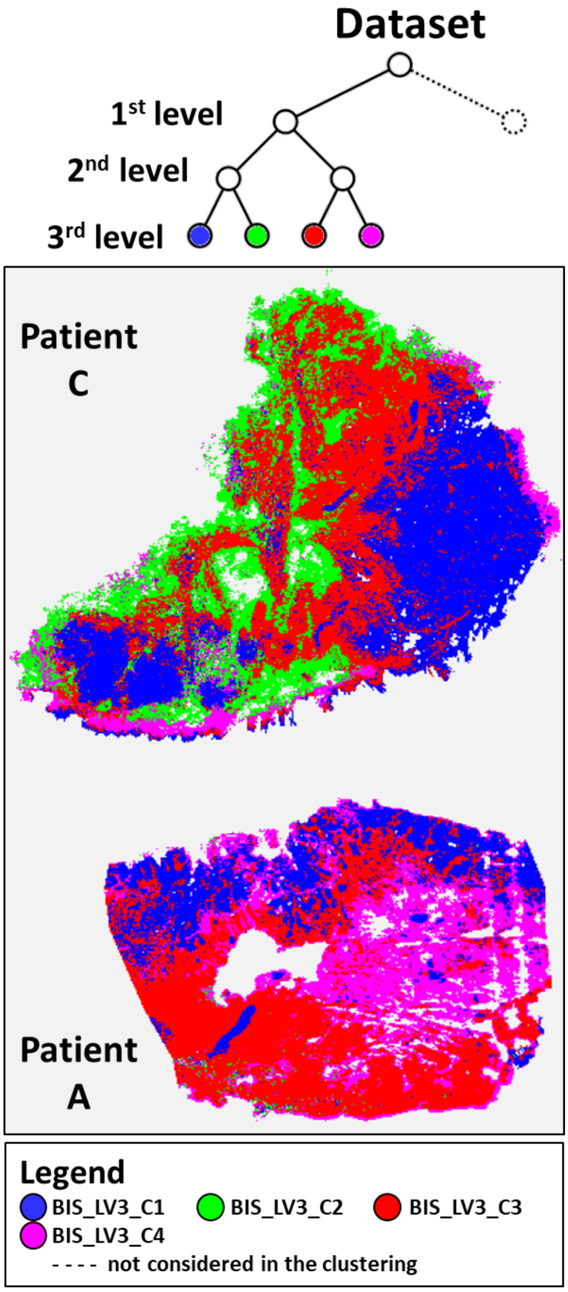


Figure 21: The third level of partitioning for bisecting kmeans.

At first glance, Figure 22, which compares the clustering of bisecting kmeans and QDD kmeans, might lead us to believe that the results are fairly comparable. In both cases, we can observe the annotated regions on the histological images, including tumor areas, lingual muscle, inflammatory zones, and nerves. Nevertheless, a more de-tailed analysis reveals that the QDD kmeans approach enables a more efficient exploration of the biological complexity of such samples. If we focus specifically on the tumor region, which lies at the very heart of this issue, we observe that it is represented by a single cluster (e.g., BIC_LV4_C3 in black) in the bisecting kmeans method. This is, of course, not a poor result when comparing its spatial distribution with histological images. What is very interesting, indeed, is that the QDD kmeans approach partitions the tumoral regions through three distinct clusters (QDD_LV3_C3 in neon green, QDD_LV3_C4 in orange and QDD_LV3_C5 in black). The superimposition of the high-resolution histological image onto the clustering map further demonstrates that cluster QDD_LV3_C4 (in orange) is specifically associated with cancerous cells. This further reveals inter-patient variability, with patient A having approximately 3.8 times more cancer cells than patient C. Another singularity is observed at the lingual muscle level with edema. Indeed, this area is characterized by a unique cluster (e.g., BIS_LV_C2 in green) with bisecting kmeans, whereas QDD kmeans presents two clusters (QDD_LV3_C7 in green and QDD_LV3_C8 in light blue). There is, therefore, still a discrepancy among patients regarding this area. Finally, when examining the inflammatory region, we observe that cluster BIS_LV4_C1 (in red), which defines it in bisecting kmeans, extends excessively over nearly the entire tissue of patient A, whereas it should be more localized around the periphery of the tumor area, as is the case with QDD kmeans (with cluster QDD_LV3_C6 in red). Given Figure 22, one might be tempt-ed to argue that the comparison between bisecting kmeans and QDD kmeans is again somewhat unfair, as these methods yield significantly different numbers of clusters—8 and 13, respectively. However, it is important to recall that the fourth portioning level for kmeans and the third one for QDD kmeans were specifically chosen because they allowed for the retrieval of annotated regions in histological images. To dispel any doubts regarding these results, figure 23 presents the outcomes of bisecting kmeans at the fifth partitioning level, which mechanically results in 16 clusters.

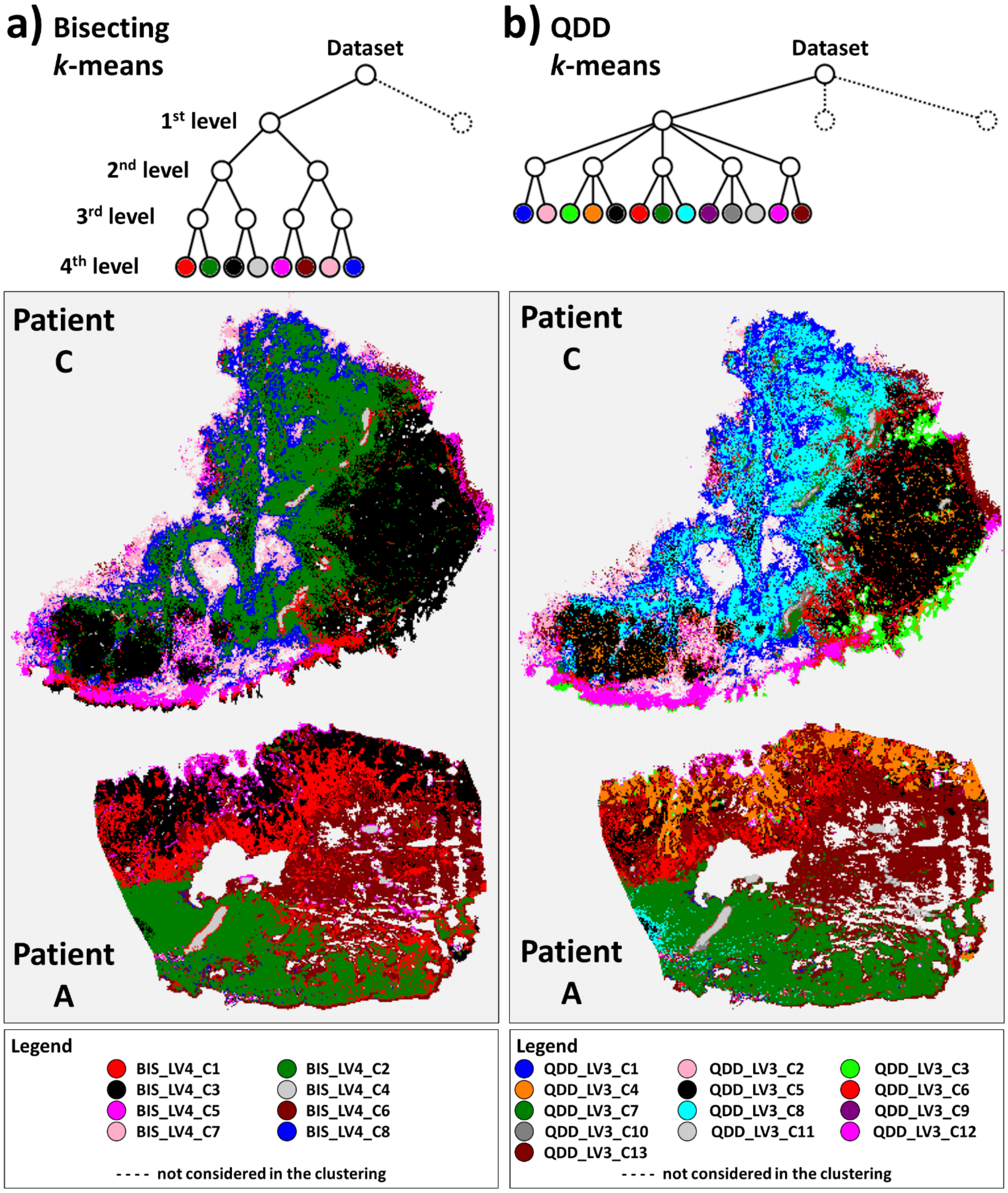


Figure 22: The final clustering maps a) considering bisecting kmeans with four levels of partitioning and b) considering QDD kmeans with three levels of partitioning.

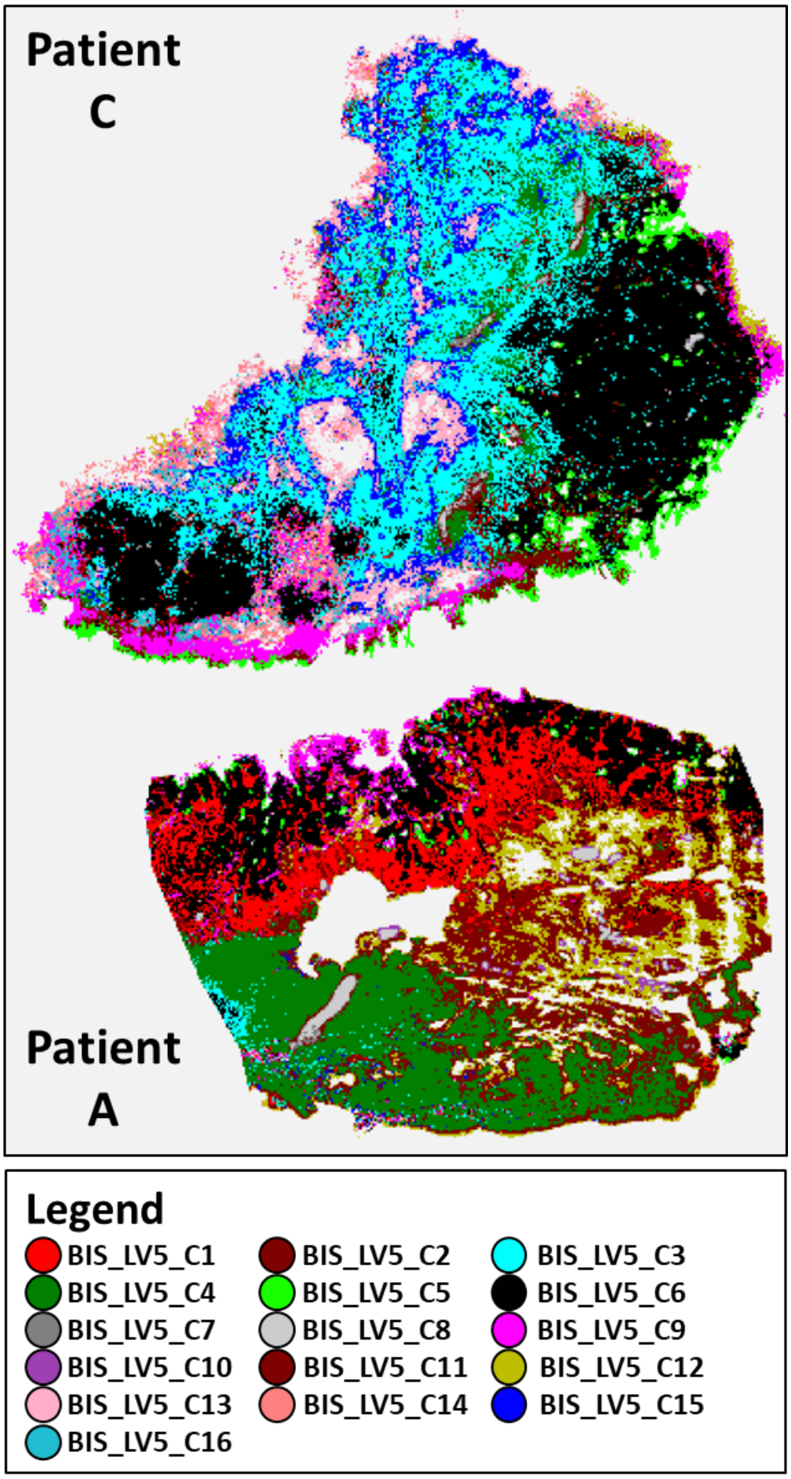


Figure 23: The fifth level of partitioning for bisecting kmeans.

Nevertheless, despite a number of clusters even exceeding the 13 obtained with QDD kmeans, we still observe certain inaccuracies. Indeed, the cancerous cells are not visible with bisecting kmeans, whereas they were in the case of QDD kmeans with the cluster QDD_LV3_C4. A certain lack of coherence can also be noted in the bisecting kmeans results regarding the inflammatory stromal reaction (corresponding to cluster BIS_LV5_C1 in red), which is entirely absent in patient C. However, it was weakly detected using QDD kmeans, as might be expected in proximity to cancerous regions. Considering the overall results, the QDD kmeans approach appears to offer an effective and relevant clustering method by leveraging a limited number of partitioning levels, which helps to substantially mitigate the effects of over-partitioning and, consequently, reduce the risk of generating clustering maps with unrealistic details.

2.1.3 Conclusions

In summary, this study has introduced the Quality-Driven Divisive (QDD) kmeans approach, a novel, adaptive clustering strategy that significantly enhances the analysis of MALDI-MSI data and provides a more accurate, less biased characterization of complex biological tissues. Traditional bisection kmeans, despite its hierarchical appeal, is inherently constrained by its fixed binary partitioning. This rigidity can lead to arbitrary merging or over-splitting of clusters, ultimately distorting the spatial distribution of biologically relevant regions. In contrast, QDD kmeans dynamically determines the optimal number of clusters at each partitioning level by using robust partition quality metrics, such as the silhouette score. This merit-based strategy ensures that the clustering process is intrinsically data-driven, thereby accounting for the inherent heterogeneity of MALDI spectral datasets. Our application of QDD kmeans to oral tongue squamous cell carcinoma tissues has demonstrated its superior performance over bisection kmeans. Specifically, QDD kmeans was the only approach capable of identifying a dedicated cluster for cancer cells. In addition, this method allowed for a more accurate delineation of inflammatory response zones. The ability of QDD kmeans to generate precise masks for biologically relevant regions ensures that extraneous spectral noise, such as signals from the MALDI matrix, is effectively minimized. This selective retention of meaningful spectra facilitates a clearer correlation between the molecular data and the underlying histology. The centroids of the 13 clusters obtained by QDD and the list of the peaks selected are showed in figures 88-100, in Appendix (5). From our perspective, the QDD kmeans approach holds great promise for advancing the broader field of spectroscopic imaging. Its ability to adaptively resolve spatial and molecular heterogeneity paves the way for more nuanced exploration of the tissue microenvironment, potentially impacting both research and clinical practice.

2.2 Automating hierarchical clustering

The purpose of this chapter is to propose a method that attempts to segment the data using a truly unsupervised approach. Before explaining the approach used, an introduction and a discussion of the stopping criteria on the divisive hierarchical clustering algorithm will be presented.

2.2.1 Divisive hierarchical clustering algorithm

The divisive hierarchical clustering algorithm can be summarized with this structure. The starting point is to place all samples into a cluster. Iteratively, the algorithm divides the cluster into smaller clusters. This process is repeated until a stop criterion is reached. Examples of common stop criteria [42] are summarized in the following list: (1) number of clusters, (2) maximum clustering tree depth, (3) minimum number of instances in a cluster, and (4) minimum intra-cluster dissimilarity. The stop criteria can also be a mixture of these criteria, which are expressed in the list [43]. The choice of the stopping criteria is crucial and may depend strongly on the domain in which this algorithm is applied. For example, consider an example shown in figure 24 below. The goal is to retrieve 4 clusters from these data.

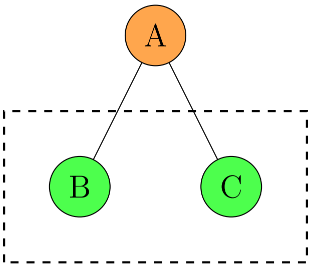


Figure 24: After the first division there are two cluster: B and C. Which one should be the next to divide?

Which cluster should be divided next? Cluster B or cluster C? Deciding which cluster to split can have a significant impact on the results of the segmentation. For example, splitting cluster B would result in the situation shown in the following figure 25 below:

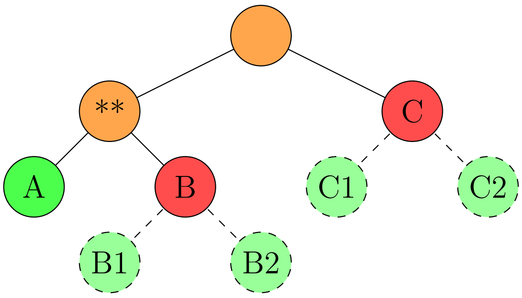


Figure 25: Cluster ** (B in the previous figure) is divided into cluster A and B. The actual situation present cluster A, B, and C. There is one last cluster to divide. Depending on which cluster (e.g. B or C) the results will change

There are 3 clusters A, B and C. Since the stop criterion is 4 as the number of clusters, the choice of which cluster to divide can totally affect the results. For example, there can be two possible solutions: dividing cluster B, the final map will be A-B1-B2-C, while dividing cluster C, the final map will be A-B-C1-C2. In real samples the correct number of clusters is not known in advance, but even in this case it is clear that this kind of criteria it is not the perfect one. Setting a maximum depth of the clustering tree in advance, without any other criteria, will lead to a systematic splitting of all branches. To avoid this, an additional stop criterion or number of iterations is needed, and in this case a criterion is needed to select which cluster should be split. This will allow to obtain a more realistic hierarchical structure. However, it is quite difficult to give a predetermined maximum depth tree, as some relevant clusters may be at a deeper level than expected. Selecting a threshold for the number of instances (spectra) in a cluster can limit the analysis as there may be very small clusters (few spectra), but with biological significance. Maybe they represent the beginning of the tumor, but just because of the threshold values, this cluster will not be considered. Even if the cluster appears to be homogeneous, splitting it may reveal two clusters with biological significance.

2.2.2 Research on the stop criteria

The aim of this part is to find the most effective way to address the issue of the stopping criteria. Several approaches were considered, and various methods were tested. One approach is worth mentioning briefly, even though it was immediately discarded due to its prohibitive computational cost: comparing all possible combinations. Given n samples and k clusters, the total number of possible combinations is given by the following equation 18:

$$S(n,k) = \frac{1}{k!}\sum_{j=1}^{k}(-1)^{k-j}\binom{k}{j}j^n \tag{18}$$

Now it is easy to see that this is prohibitive. Even with just $n = 1000$ and $k = 2$ we exceed the number of atoms in the universe. In fact, this equation with $k = 2$ became $2^{n-1} - 1$, so with $n = 1000$ the total number of possible combination is 2^{1000}. So giving an evaluation for all possible combinations is impossible.

The initial idea was to conduct a comprehensive comparison between two consecutive levels of the hierarchical tree. This comparison is made using a quality cluster criterion. If this criterion selects a situation with no further division compared to the previous iteration, the algorithm will stop. Considering kmeans as the algorithm used to divide each branch, for each branch (n) there will be k state: not divided, divided with 2 clusters, divided with 3 clusters, and so on. This case is a full factorial (grid search) case, so the number of possible combinations is k^n Such a combinatorial explosion quickly becomes computationally intractable as n increases. To reduce this complexity, we can consider a simplified binary decision for each branch: either the branch is not divided, or it is divided. Even in this binary case, the number of possible combinations at each step remains exponential: 2^n To further simplify the process, we assume that the "divided" case corresponds to a fixed split into 2 clusters, rather than the optimal number of clusters based on a clustering quality criterion. This assumption is aligned with the bisecting kmeans approach and allows us to reduce computational cost while maintaining a reasonable approximation of hierarchical clustering behavior. Consider the example shown in the figure 26. After the first division of the data set (orange), there are two clusters (blue). Each has two possible states: *divided* or *not divided.* So in this case there are 4 possible situations: A, B, C, D.

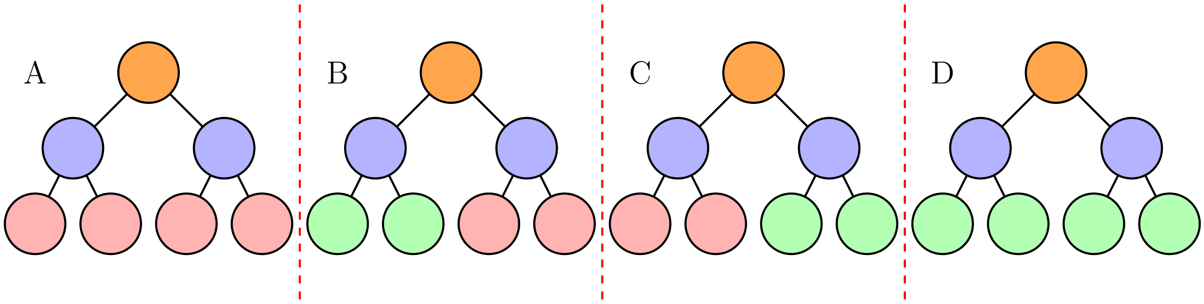


Figure 26: All the possible combination of two clusters division.

Case A is the one where there are no further splits. In case B, only the first cluster is split. Case C only the second cluster is split. Case D both clusters are split. This seems nice at first, but it is still computationally too expensive, as for k increased the number of segmentations to test is increased, and with criteria, e.g. silhouette, the time needed to test all combinations is not acceptable. Also, only two consecutive levels are considered. This means that even if further subdivisions could lead to optimal solutions, the approach relies on a ”level-by-level” comparison and makes decisions based solely on the current set of clusters. As a result, deeper divisions are not tested, which may prevent finding the truly optimal solution. So the work focused on partitioning each branch independently. Ideally, the goal was to determine whether a cluster of data should be split further, without making that decision dependent on other clusters. Consider the figure 27

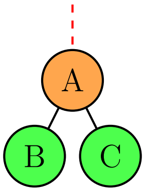


Figure 27: Generic branch (A) with his division (B-C)

How can I determine whether a cluster should be split or not?
The initial idea was to compare the orange circle (A) and its divisions, represented by the green circles (B and C). As previously explained, it is not possible to use the normal cluster criterion to compare A alone due to the separation measurement present in the algorithms. Therefore, the initial comparison was based solely on the compactness (or density) of the clusters, comparing (1) the compactness of A with the mean compactness of B and C, and (2) the density of A with the mean

density of B and C. The results showed degeneration of the tree because the levels of compactness and density were consistently high. Then a more restrictive criterion was adopted, stating that cluster A would not be split if the improvement was lower than the threshold. Even then, the tree degenerated, and if the chosen threshold did not allow the tree to generate, the clusters were 2 or 3, which was unreasonable. So, rather than comparing A with B and C, the idea was to look at B and C and try to characterize them statistically. Always keeping in mind the figure 27, if there is a relevant overlap of the two clusters (B and C), it means that A should not be divided. For example, the standard deviation of each cluster was compared with the distance between the two centroids. If the sum of the standard deviations, interpreted as the theoretical radius of the cluster, was lower than the distance between the centroids, it was considered to be overlapping. Even this approach did not work entirely, as it considered the cloud to be a perfect hypersphere, which was not the case in reality.

Another attempt was made using a 'mirror'. Since it is not feasible to compute a clustering quality index when only a single cluster is present, an alternative strategy was considered. The question arose as to whether it might be possible to introduce synthetic data that, while similar in structure to the original dataset, would be positioned far enough in the feature space to avoid interfering with the actual clustering process, yet still reflect comparable internal dynamics. To test this hypothesis, an additional matrix X' was appended to the original data matrix X. This synthetic matrix was specifically designed to invert the correlation structure relative to the original data, while preserving consistent internal relationships within X' itself. The underlying idea was that, by applying a standard clustering quality criterion to this extended dataset, an indication that two clusters were optimal could be interpreted as implicit validation of the original dataset forming a coherent, well-defined single cluster. This result suggested a potential avenue for automating quality-based decisions within the clustering process. However, in practice, the method proved to be unstable, with outcomes that varied significantly across different runs and datasets. Given these inconsistencies, the approach was set aside and not included in the subsequent phases of the study.

After several attempts to find the optimal parameter, the results always fell somewhere between no division and hundreds of clusters. It was clear that a change of approach was necessary. An attempt was made to approach the challenge from a different angle. In essence, when optimizing the number of clusters in a dataset and the stop criteria in hierarchical algorithms, the question is:

"Which segmentation is better?

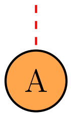
But the question remains:

"Are we sure that the optimal segmentation from a statistical point of view corresponds to the optimal from a biological point of view?"

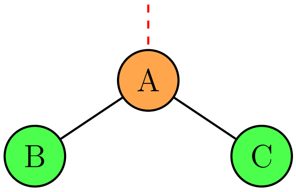
How can this question be expressed in mathematical terms? How can you prove whether the extracted information is biologically different? Proper biological validation would be required, which can be tricky and lengthy, so it does not fit well with the need for a fast algorithm. The idea was therefore not to focus on the optimal shape or statistical properties of the clusters, but to look at them more from a biological than a statistical point of view. It is important that the clusters are as representative as possible of the bioheterogeneity of the sample. So the idea was to extract information from B and C (always referring to the figure 27) and try to understand if they show the same kind of information or different ones. This can be done with two different approaches: **Univariate:** Use the feature extraction algorithm to obtain a list of the most important metabolites; **Multivariate:** Use algorithms such as PCA to extract more information at the same time. Both approaches have been tested, but the second is presented here due to its more manageable characteristics. The first approach will be discussed in the conclusion and perspective section.

At the outset, the algorithm for each branch can be summarized as follows:

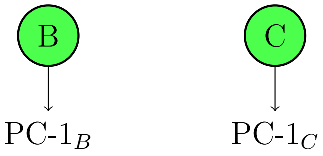
1. Consider a cluster A



2. Divide into two-subclusters



3. Extract the PC-1 for each sub-clusters (PCA performed on each cluster separately)



4. Calculate the absolute value of correlation between the two PC-1

$$\text{PC-1}_B \longleftrightarrow \text{PC-1}_C$$

5. If the absolute value of correlation is lower than a threshold then divide, otherwise not

This way, the division of all branches is tested. The decision can be interpreted as follows: ”if the main information is not highly correlated, then the information that can be extracted from the two clusters is different and the partition makes sense”. Otherwise, if the information is similar, the partition does not make much sense. Reading this, you might ask, ’How can you know the optimal threshold in advance?’ The answer is quite simple: **you can’t**. It is necessary to use and test different thresholds. A segmentation map is obtained for each threshold used. Ultimately, you obtain different segmentation maps corresponding to the various depths of the same tree (the higher the threshold, the less restrictive it is, and the deeper you go). It is necessary to identify the optimal segmentation map. Although the normal clustering criteria can be used, since our approach is based on a different question, we tried to maintain the same approach for finding the optimal segmentation map. For each map, PC-1 is extracted from each cluster. Then |D(correlation)-1| (or |corr|) is calculated for each combination. The number of distances needed to be calculated using the equation 19

$$\binom{n}{2} = \frac{n!}{2!(n-2)!} = \frac{n(n-1)}{2} \tag{19}$$

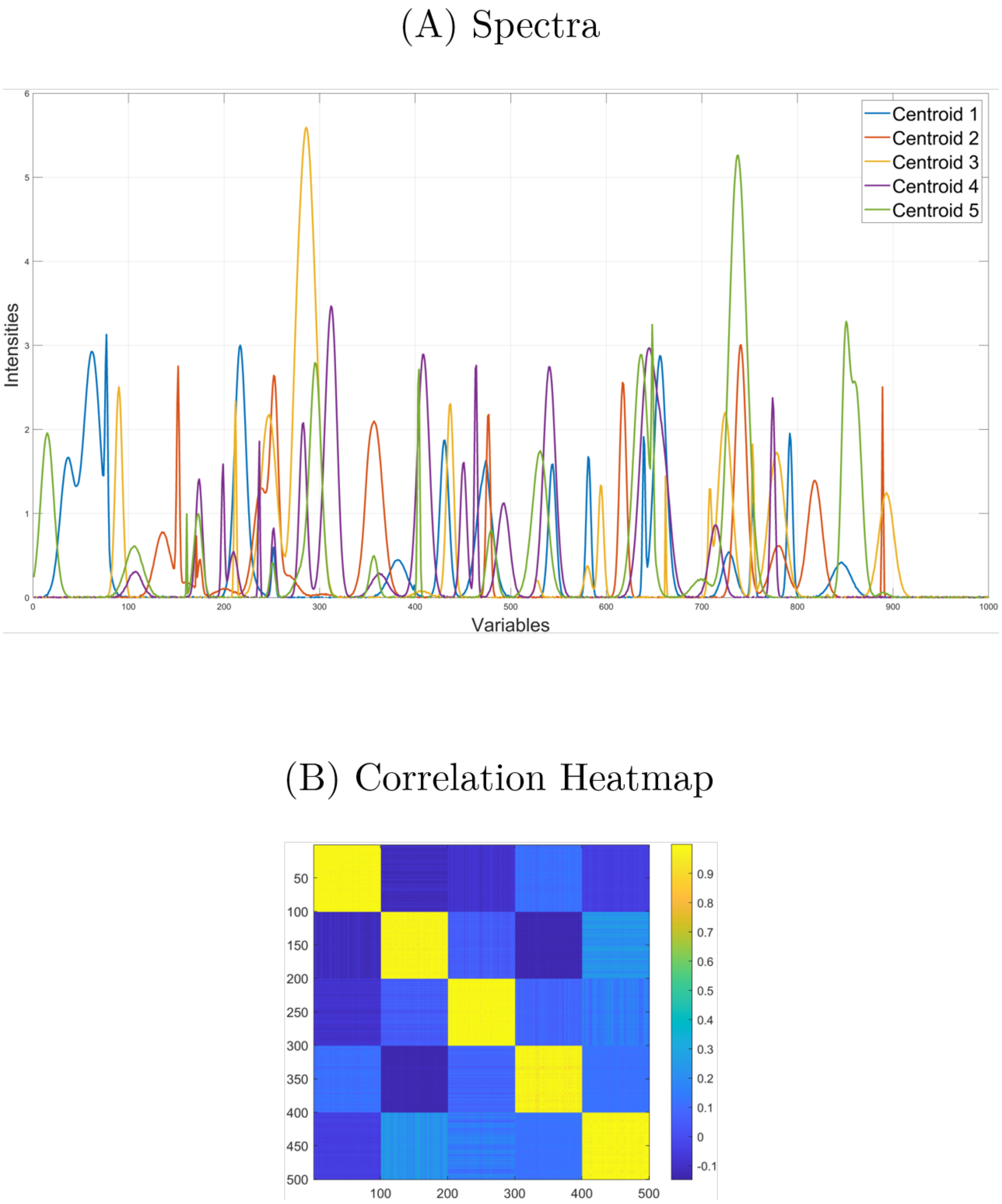
Calculating the median of this value provides an indication of the quality of the segmentation. Repeating this process for all the segmentation maps and plotting the results provides an evaluation of their quality. There will be an indicator of the total absolute value of the correlation for each segmentation map. The map with the minimum value can be considered the one with the lowest overall absolute correlation between the clusters, and therefore the most interesting one from our perspective. The only inputs are the spectral data cube and a list of different thresholds. Therefore, there are no other input options or operator choices that can influence the division. A toy example was simulated to test the algorithm. Five clusters were simulated, each containing 100 spectra with 1000 spectral variables. The code for generating the spectra is shown in the figure 28 below.

Matlab

```
num_points = 1000;
x = linspace(0, 1000, num_points);
n_peaks  = 10;
mu       = randperm(800, n_peaks) + 10;
for j = 1:10
    mu      = randperm(900, 20) + 10;
    sigma   = rand;
    spectrum = zeros(size(x));
    for i = 1:n_peaks
        sigma     = rand * 10;
        amplitude = rand * 10;
        spectrum  = spectrum + amplitude * exp(-((x - mu(i))
           .^2) / (2 * sigma^2));
    end
    spectra(j, :) = spectrum;
end
n_cluster = 100;
sp     = [];
labels = [];
rng(1);
for k = 1:5
    base1 = spectra(2 * k - 1, :);
    base2 = spectra(2 * k, :);
    w  = linspace(0.2, 0.4, n_cluster)';
    noise = 0.05 * randn(n_cluster, num_points);
    cluster = w(randperm(100, 100)) .* base1 + w(randperm
       (100, 100)) .* base2 + noise;
    sp      = [sp; cluster];
    labels  = [labels; k * ones(n_cluster, 1)];
end
```

Figure 28: Matlab code for the generation of the spectra

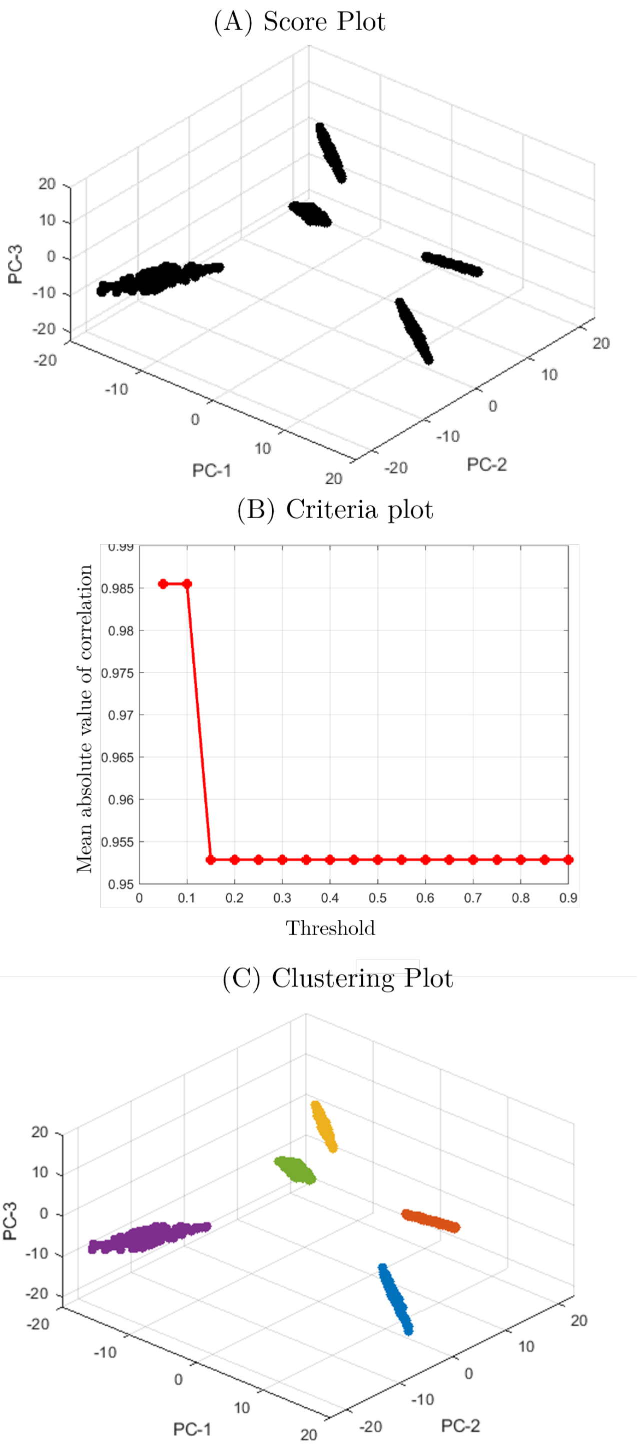
In this way, there is a high level of correlation within clusters, while the level of correlation between clusters is low. Figure 29 below shows the centroid of the cluster and the correlation heatmap of all spectra.



(A) Spectra

(B) Correlation Heatmap

Figure 29: Visual representation of the five clusters. Each color represents a cluster.



(A) Score Plot

(B) Criteria plot

(C) Clustering Plot

Figure 30: (A) scores plot of the data with the original labels. (B) plot of the criterion value versus the threshold. (C) scores plot of the data with the labels from the clustering algorithm.

The algorithm is applied. Different thresholds were tested: from 0.05 to 0.9 with a step of 0.05, for a total of 18 values. The results are shown in Figure 30. Figure 30-A shows the score plot of the data after a PCA. Figure 30-B shows the plot of the criterion value versus the threshold. With the first two thresholds, the algorithm stopped at 4 clusters. By increasing the threshold, the number of clusters reached 5. Even when reaching a threshold of 0.9, there was no further increase in division. This means that when attempting to split, the generated sub-clusters were always too strongly correlated. Figure 30-C shows the score plot of the data with the labels from the clustering algorithm. The labeling matches 100% the true labeling, proving that in this toy example the algorithm and the criterion work properly together.

Real Sample

The sample used in this part was one of the 16 samples used in a previous article [44]. The collection, preparation and analysis of the sample will be reported as expressed in the article. The tissues are colorectal cancer resection that comes from the department of pathology from Leiden University Medical Center. Tissues were Formalin-Fixed Paraffin-Embedded (FFPE) according to routine protocols of the department. H&E-stained sections were annotated by two experienced gastointestinal pathologist. FFPE tissue blocks were sectioned with a microtome (Leica Biosystems RM2245 Microtome) at 6-μm thickness. Tissues sections were mounted in pairs onto poly-L-lysine- and indium tin oxide-coated glass slides (Bruker Daltonics). Before tissue mounting, glass slides were cleaned in 70% ethanol for 10 min and coated with a 0.05% poly-L-lysine solution in mQ. All sections were dried overnight at 37°C, stored at 4°C, and then prepared following the procedure by Holst et al. [45]. In brief, paraffin was removed by heating the slides for 1 h at 65°C followed by two consecutive washes in xylene (10 min and 5 min, respectively). Tissues were rehydrated in ethanol baths (100% ethanol, twice for 2 min), followed by water baths (twice for 5 min), and dried for 10 min in a vacuum desiccator. On-tissue derivatization was performed by incubating the tissue slides in derivatization solution (250 mM 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide, 500 mM 1-hydroxybenzotriazole hydrate, and 250 mM dimethylamine in dimethylsulfoxide) for 1 h at 60°C, followed by addition of a 25% ammonia solution (1:0.4 v/v derivatization solution-ammonia) and further incubation for 2 h at 60°C, both protected from evaporation. After derivatization, tissue sections were rinsed thoroughly with

100% ethanol followed by sequential washes in 100% ethanol (2 × 2 min) and water (2 × 5 min). Slides were dried in a vacuum desiccator (10 min). On-tissue digestion was performed applying 10 layers at 10 μl/min of peptide N-acetyl-beta-glucosaminyl asparagine amidase (0.1 μg/μl in Tris buffer; from N-Zyme Scientifics) using a SunCollect sprayer (SunChrom). N-glycans were released overnight at 37°C in a humid environment. After incubation, slides were dried in a vacuum desiccator for 10 min, followed by matrix application (5 mg/ml α-Cyano-4-hydroxycinnamic acid in 50:49.9/0.1 (%v/v) ACN:mQ:TFA) using the SunCollect sprayer (6 layers at (1) 10 μl/min, (2) 20 μl/min, (3) 30 μl/min, (4+) 40 μl/min). N-glycan MALDI-MSI was performed in positive-ion reflectron mode on a rapifleX MALDI-TOF/TOF-MS instrument (Bruker Daltonics). A m/z range of 900 to 3300 was used, with 1000 laser shots per pixel, and a 50 × 50 μm^2 pixel size. MSI data acquisition was enabled by the flexImaging software (flexImaging 4.0 Build 32, Bruker Daltonics). After the MSI analysis, excess MALDI-matrix was removed by washing twice in 70% ethanol (5 min each). Tissues were stained with H&E after routine histopathological procedures. As preprocessing, was used the one applied in the article. The average spectrum was processed in mMass[46] using the following parameters: baseline subtraction with 15 precision and 25 relative offset; smoothing with Savitzky-Golay smoothing, window size: 0.05 m/z and four cycles; internally recalibration Peak picking was performed with a signal-to-noise (S/N) threshold of 3 (S/N $\geq$3) followed by deisotoping (maximum charge: 1, isotope mass tolerance: 0.15 m/z, isotope in tensity tolerance: 50%).

Here, the idea was to find an automatic clustering algorithm by testing the division of each branch individually. The criteria for splitting or not splitting the branch is based on the correlation between the first principal component of one branch and the first principal component of the other branch. The purpose here is not to propose the perfect automatic clustering algorithm, but to modify the approach slightly by changing the reasoning behind the criteria. Previously, this approach was tested on a simulated dataset to prove the potential of this approach and to check that it does not lead to inconsistent results. When applied to a real sample, the goal was not to retrieve the biological annotation, but to study the segmentation map, compare the different clusters, and study the differences. In fact, the goal was to obtain an automatic clustering algorithm that can obtain a segmentation map based on fixed criteria, ensuring that the main information that can be extracted from the clusters of the segmentation map obtained are the less correlated compared to the other segmentation maps.

Results

During the acquisition, some holes are present in the tissue, clearly visible in figure 31-left. For this reason a mask (figure 31-right) was generated, to avoid artifacts. The cube was unfolded and once the mask was applied, only the spectra related to the sample were kept. The threshold we tested: 0.86, 0.87, 0.88, 0.89, 0.90, 0.91, 0.92, 0.93. Higher values lead to results with more than 20 clusters, which can be considered as oversegmented. All segmentation maps are shown figure 32. Once the different segmentation maps have been obtained, it is necessary to find the optimal segmentation map. Applying the criteria previously explained, the segmentation obtained with a threshold equal to 0.91 is the optimal one, as shown in figure 33.

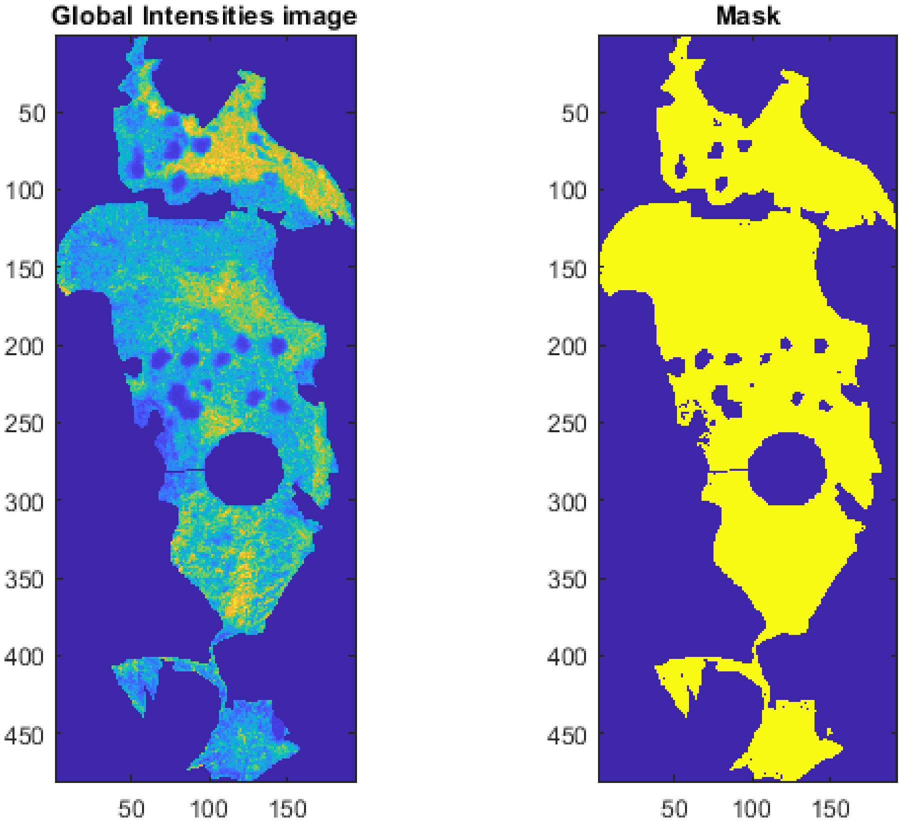


Figure 31: Mask generation

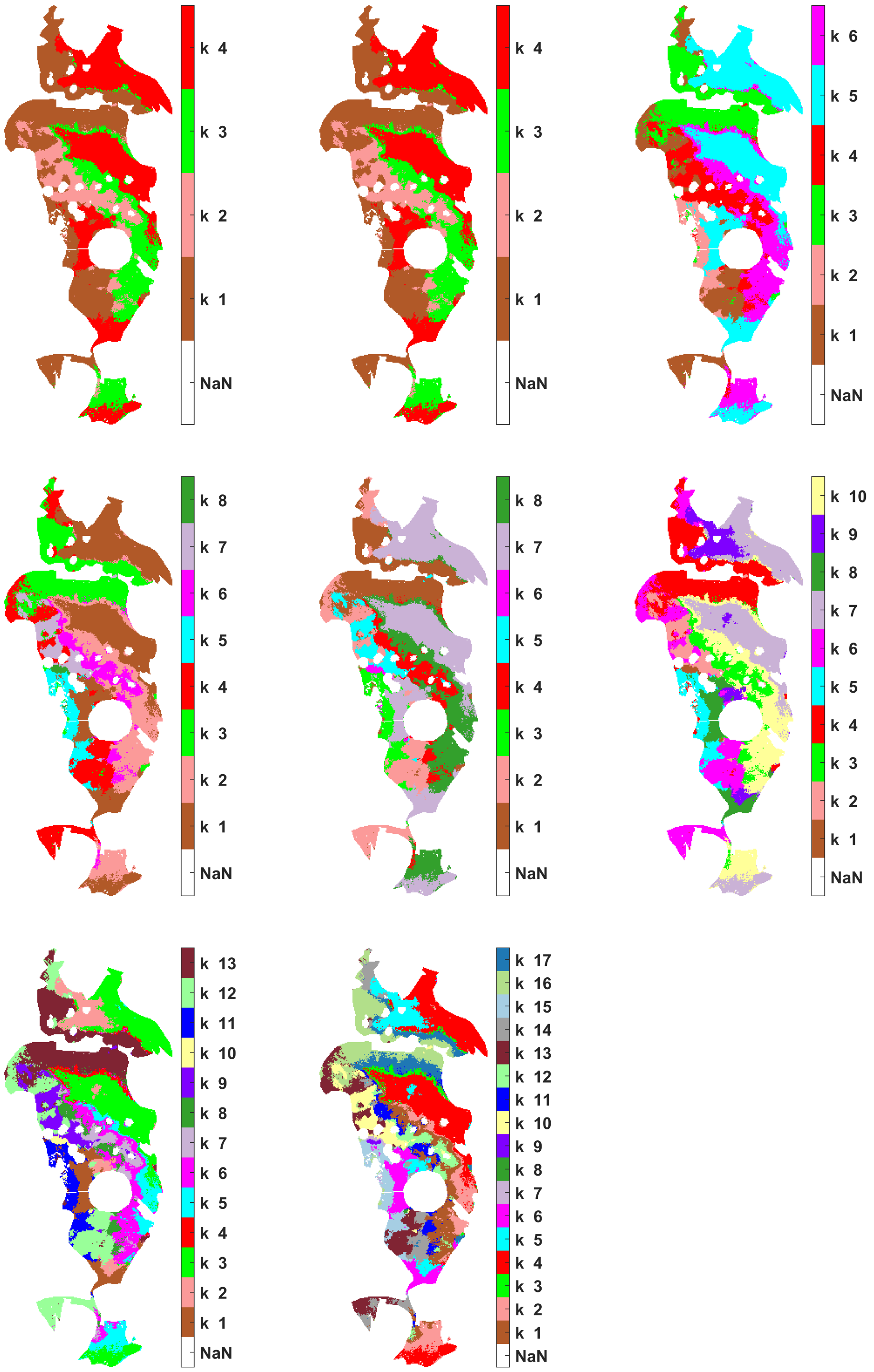


Figure 32: Segmentation maps with the different thresholds. The thresgold are: 0.86 (k=4), 0.87 (k=4), 0.88 (k=6), 0.89 (k=8), 0.90 (k=8), 0.91 (k=10), 0.92 (k=13), 0.93 (k=17)

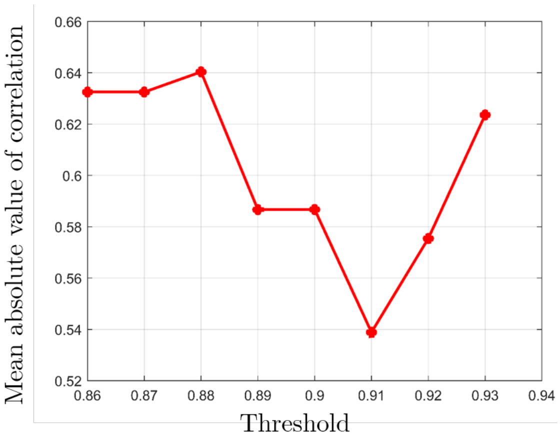


Figure 33: Validation plot: value of criterion vs threshold.

Once the optimal segmentation map is obtained, it is relevant to see the structure of this segmentation map, remembering that this is a hierarchical divisive clustering algorithm. The tree is shown in figure 34

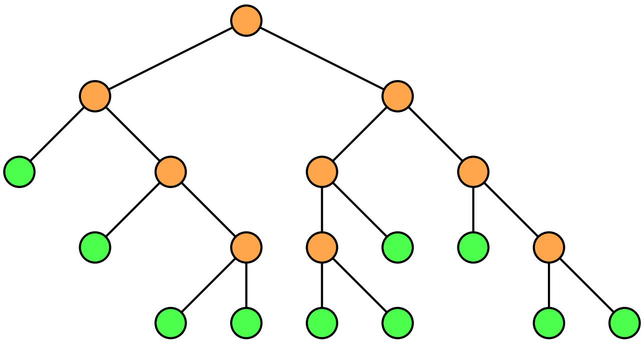
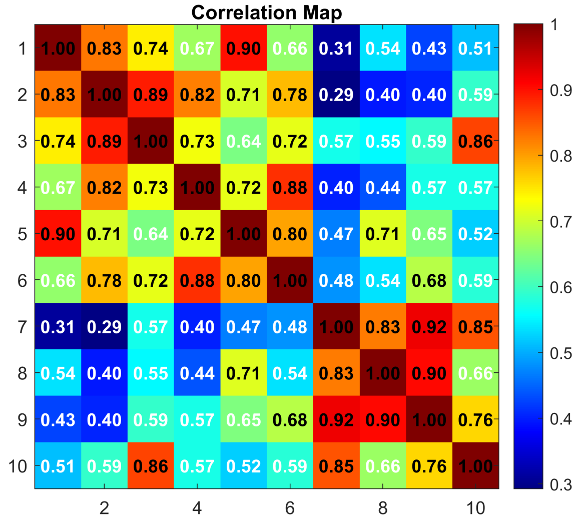


Figure 34: Hierarchical tree of the sample. The green circles represent the final clusters

The first consideration of the data was done by comparing the segmentation map with the histological annotation (figure 36). At first sight, the segmentation maps seem to show more clusters than the one annotated in the histological image. The yellow cluster corresponds to the stroma in the histological annotation (figure 36-left) seems to be explained by the cluster 7 in figure 36-right, looking at the top of the tissue. However, when looking at the middle or the bottom of the tissue, it seems to be present in the region associated with the tumor (cluster red in figure 36 - left). The cluster 1 in figure 36-right seems to be related to the central stromal region. To properly read the comments on this correlation map, it is important to look at the hierarchical tree shown in Figure 34. In the final map, clusters 1-4 come from cluster 1 of the first division, while clusters 6-10 come from the second cluster of the first division. Looking at cluster 1, the highest correlation is with cluster 5, which was separated from cluster 1 during the first division. The higher concentration between clusters 2 and 3 than between clusters 3 and 4, which are the division of the same branch, seems to indicate that during the formation of the parent cluster of 3 and 4, the division of this cluster and cluster 2 was not quite correct. In addition, cluster 3 has a high correlation with cluster 10, even though it is quite far away in the hierarchical tree. The same discussion can be made with clusters 4-6. The same comments regarding the pattern found in clusters 2-3-4 can be observed in clusters 7-9-10. The higher correlation of 9 and 10 with 7 than with each other. Cluster 7, looking at the histologic annotation, covers those annotated as ”hyperplastic” and ”muscle/healthy”. Since cluster 7 was not split, it means that the clusters obtained from its split are not considered to be sufficiently different based on the criteria and threshold. This means that these two regions are very close from a spectral point of view. In addition, cluster 7 shows a higher concentration compared to the other part of the tissue, with clusters 8, 9 and 10, which are located in the center of the tumor region, around the holes. It is important to emphasize that the focus of this approach was to test the division criteria. As mentioned in the introduction, bisection was chosen as the easiest to apply because, unlike the QDD approach, there is no optimization step, so there are no other parameters to test.



1	1.00	0.83	0.74	0.67	0.90	0.66	0.31	0.54	0.43	0.51
2	0.83	1.00	0.89	0.82	0.71	0.78	0.29	0.40	0.40	0.59
3	0.74	0.89	1.00	0.73	0.64	0.72	0.57	0.55	0.59	0.86
4	0.67	0.82	0.73	1.00	0.72	0.88	0.40	0.44	0.57	0.57
5	0.90	0.71	0.64	0.72	1.00	0.80	0.47	0.71	0.65	0.52
6	0.66	0.78	0.72	0.88	0.80	1.00	0.48	0.54	0.68	0.59
7	0.31	0.29	0.57	0.40	0.47	0.48	1.00	0.83	0.92	0.85
8	0.54	0.40	0.55	0.44	0.71	0.54	0.83	1.00	0.90	0.66
9	0.43	0.40	0.59	0.57	0.65	0.68	0.92	0.90	1.00	0.76
10	0.51	0.59	0.86	0.57	0.52	0.59	0.85	0.66	0.76	1.00

Figure 35: Correlation heatmap of cluster centroids

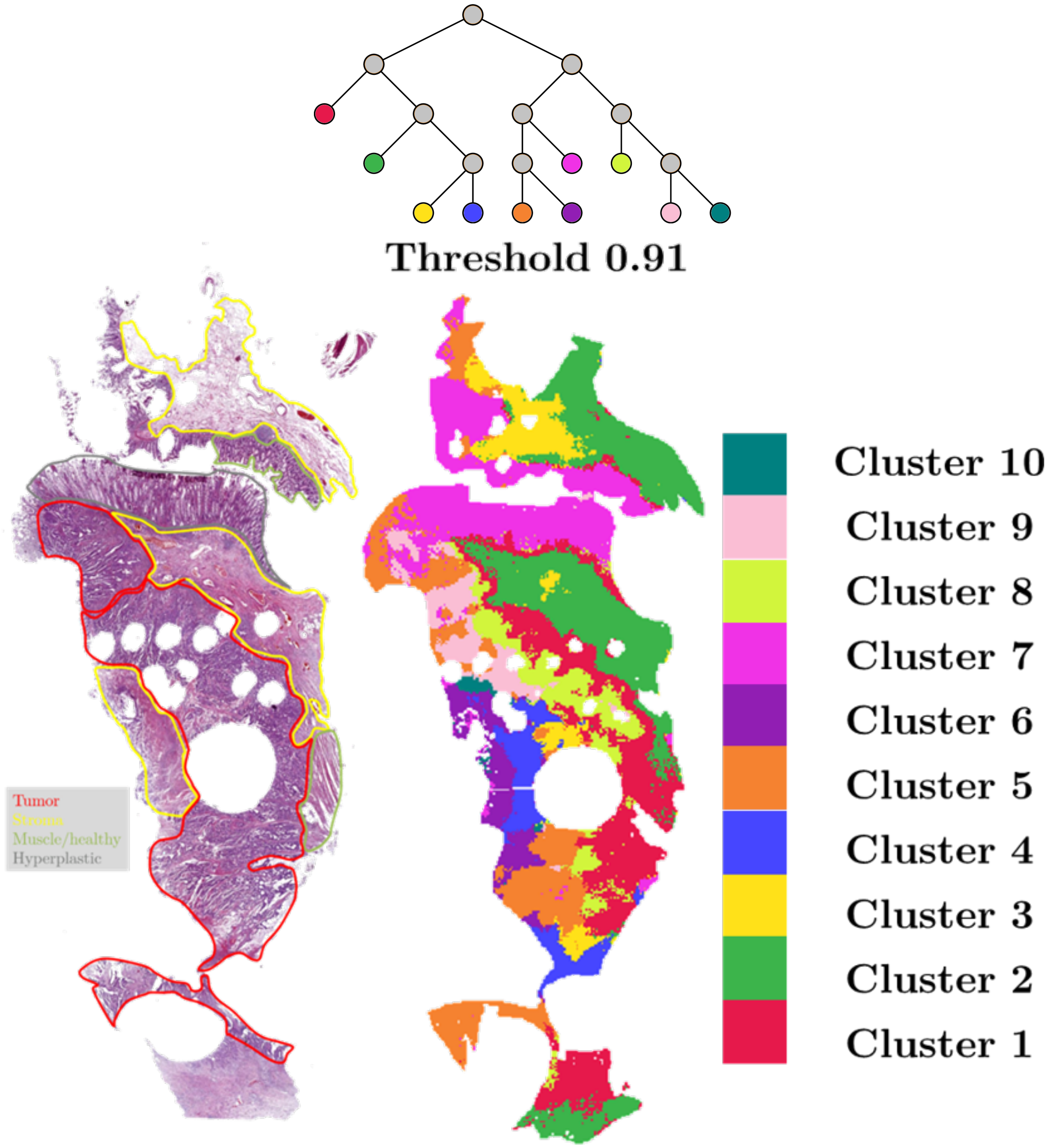


Figure 36: Comparison between final map and histological annotation. On the top of the figure the hierarchical tree generated is represented with the respectively colors

Some correlations of the cluster centroids suggest us that the division process is not the optimal one, so it is not directly linked to the proposed criteria, e.g. clusters 1 and 5 are separated already in the first division, even if they seem to belong to the same part of the tissue, and this is due to the nature of the bisection algorithm, not to the criteria. With an interactive procedure, the results may be the same.

Conclusion and future perspective

Unlike the first section of this chapter, where the work focused on the division (QDD vs bisecting), the focus of this section was to find a set of stop criteria that work properly in the MALDI imaging domain for tissue analysis. However, even though they were approached separately, this does not mean that they are distinct. In fact, at the beginning of this section, it was stated that, for each division, the cases 'not divided' and 'divided' could be considered, where 'divided' could represent the optimal number of clusters, in accordance with the QDD approach. This would improve upon the QDD method, which previously used a predetermined maximum level of tree depth. Using bisecting and QDD with this stopping criterion enables a fairer comparison to be made. This will be the next step in the project. Additionally, the manner in which the similarity between the clusters is considered is relevant. In this case, the clusters were compared using PCA. In our view, the work done with PCA was just the beginning. The main drawback is that retaining only the first principal component discards the smaller variances, which can easily become limiting. It is possible that the differences between the different stages of the tumor lie in very small variations that were not taken into account in this case. As previously explained, the idea was to extract the most important information and establish whether the clusters shared it. Rather than using PCA, a multivariate approach, we considered using a univariate approach: extracting the most interesting variable for each cluster and comparing them. Ideally, this would allow us to identify the most interesting chemical species in each cluster and compare them. There are many feature selection methods in the literature [47], depending on the domain and purpose. An attempt has been made with a feature deflation process among all of them. This idea comes from the Covsel approach [48], but with a relevant simplification. In that approach, the most interesting variables for classification were extracted. Here, however, the idea was to extract the most interesting variables for each cluster independently. The idea was therefore to maintain the deflation precedence in order to extract non-correlated variables, but it was necessary to establish a selection procedure. First, the variance spectrum is computed, and then the variable with the highest variance is selected. This variable is then deflated into an X matrix. This step removes information that is parallel to the selected variable. Successive variables are then extracted until the preselected number is reached. This feature extraction approach could be used as an alternative to the PCA method employed in the study. As mentioned, the main difficulty lies in extracting the key information from the clusters. Using PCA with n PCs or a feature extraction algorithm can alter the results. While PCA or feature extraction can influence the clustering outcome, the primary goal is to obtain a segmentation map that is potentially biologically relevant. Once this has been obtained, the clusters can be properly compared by looking at the centroids, variance of the m/z channels or any other necessary comparisons. Further interpretation of the sample shown in the results of this section will prove that the obtained segmentation can be considered a fair interpretation of the sample.

2.3 Boosting clustering analysis using spatial information

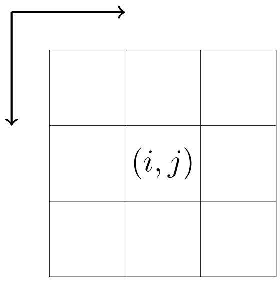
Imaging techniques are used in many fields, from biological studies to remote sensing in agriculture. Too often, however, the advantage of having spatial information is not exploited. As mentioned above, the availability of spatial information is one of the key points of imaging techniques. Nevertheless, one of the main steps is usually to apply an unfolding, e.g. to go from a 3-D cube to a 2-D matix. In doing so, the spatial relationship between pixels is lost. This loss can reduce the information that can help to understand the sample. The goal of this part of the research was to provide a novel approach to improve the analysis by adding spatial information to the spectral information.

There are several ways to extract spatial information from an image, such as using wavelets [49] directly on the image or combining them with the Gray Level Occurrence Matrix (GLCM) [50]. These are not explored in MALDI imaging domain, as our knowledge at the current days. In MALDI imaging domain, an important milestone regarding the use of spatial information was achieved by Alexandrov team [33, 51]. They implemented the total variation minimizing Chambolle algorithm with its Grasmair variant. This had a significant impact, as the algorithm has been implemented in various MSI software tools. The algorithm incorporates spatial information, but applies it after performing data denoising. Alexandrov and his team [52] also proposed a method that integrates spatial information directly into the clustering process, known as spatially aware clustering.

The idea was to try simpler algorithms, given the lack of work in this area, and then use more complex algorithms once we figured out how best to approach the problem. Of all the methods available, it was decided to start by exploring the effects of adding spatial information using the Bharati algorithm [53]. This algorithm was first applied to grayscale images and then implemented for RGB images [54]. An application in HSI domain with Bharati approach has already been done [55], but not in MALDI.

Bharati algorithm

The algorithm works as follow. Assume a generic pixel (i, j) in the image is selected.



Given a spectrum *sp* in a position $p_{i,j}$ the new augmented spectrum concatenation of the 9 spectra of the grid taken into account, as shown in figure 37 below, where a visual representation of Bharati is showed

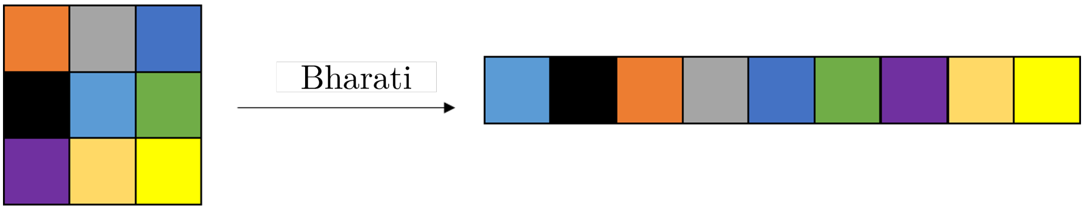


Figure 37: Bharati algorithm representation. Considering the light blue pixel, the neighbordhood pixels are put in a row and became the new information of that pixel (light blue). In this case the color is used as representation for the information. In hyperspectral imaging the color is replaced by the spectrum.

The approach requires that each pixel has 8 neighbors. Since the edge of the image does not meet this requirement, the cube after this preprocessing will be $(n-1)\ x\ (m-1)\ x\ (9*p)$. For cubes that cover the entire sample, this only affects the edge of the image. In the case of MALDI imaging, or whenever the sample is part of the cube, this will mainly affect the edge of the sample where the pixel does not have 8 meaningful pixels. The goal was to improve the Bharati algorithm applied to the hyperspectral images, trying to reduce the augmentation, as it increases the spectral dimension by a factor of 9. Considering a pixel with coordinates i, j, the idea was to find the optimal way to express the information of the 3x3 grid in a single value. In this way, the size problem of this algorithm can be overcome and potentially larger regions can be explored. In fact, considering a possible 5x5 grid, the spectral size of the cube with the Bharati algorithm is multiplied by a factor of 25. For a 7x7 grid, the factor is 49. Both are prohibitive. The problem then is to find a way to reduce the information of the 3x3 grid to just one value? As a starting point, the approach followed was "the simpler the better", so the first attempts were made using a normal statistical descriptor as the mean. Using the spatial mean, a visual representation is shown in figure 38, which compares it with the Bharati approach.

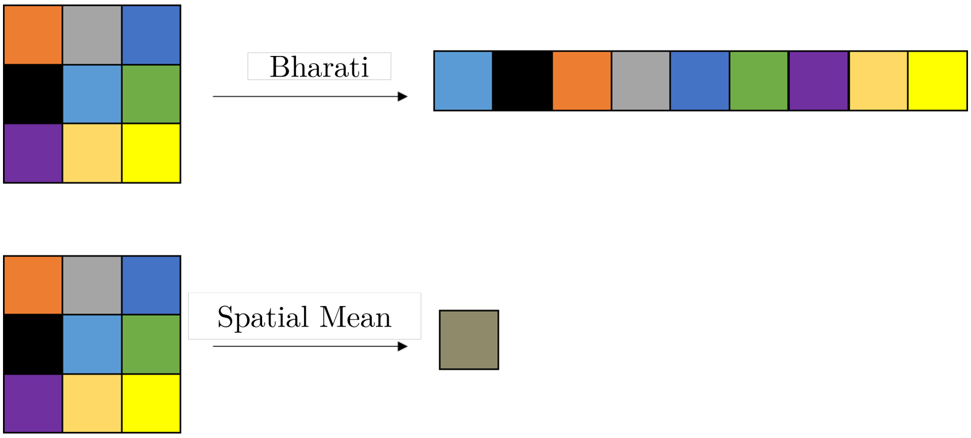


Figure 38: Comparison Bharati and spatial mean. The pixel represented in "spatial mean" is the mean of RGB values of the original grid.

From this point, *spatial mean* will refer to the use of the mean to extract the spatial information. The initial idea was to compare the Bharati approach with the spatial mean. As it will be explained later, the approach for spatial mean can be used for other statistical descriptor of the neighborhood such as standard deviation. These can enhance the spatial information of the neighborhood, as it will be proved with the simulations done during this work.

First Simulation

The first simulation aims to replicate three regions based on two different spectra. The spectra are generated as follow:

Matlab

```
region1  = repmat(1,100,100);
region2 = repmat(5,100,100);
a = repmat([1;5],50,1);
b = repmat([5;1],50,1);
mix = repmat([region1;region2],50,1);
mix = reshape(mix,100,100);
num_points = 1000;
x = linspace(0, 1000, num_points);
n_peaks = 10;
mu = randperm(800,n_peaks)+10;
% Peaks generation
for z=1:2
    spectrum = zeros(size(x));
    for i = 1:n_peaks
        sigma = rand*10;
        amplitude = rand*10;
        spectra{z} = spectrum +amplitude* exp(-((x - mu(i)).^2) / (2 * sigma^2));
    end
    % Normalization
    spectra{z} = spectra{z} / max(spectra{z});
end
data = [region1;mix;region2];
index_1 = find(data==1);
index_10 = find(data==5);
matrix(index_1,:) =repmat(spectra{1},15000,1);
matrix(index_10,:)=repmat(spectra{2},15000,1);
matrix = matrix + rand(300*100,1000);
```

Figure 39: Matlab script for the spectra generation of the first simulation

The simulation is represented in figure 40 below. There are 3 regions: the first region (yellow) is done with spectrum 1 plus noise. The second region (yellow and blue alternating) is done with spectrum 1 and 2 alternating. The third region (blue) is done with spectrum 2. From a spectral point of view, there are 2 clusters. To explain this as a possible real case, consider the yellow region as healthy tissue and the blue region as cancerous tissue. The middle region is a mixture of healthy and diseased cells. From our point of view, the goal is to get three clusters as a better representation of the sample. Figure 41 shows the results of kmeans ($k = 3$) applied to the raw data, the data enriched with Bharati-McGregor, and with the spatial mean.

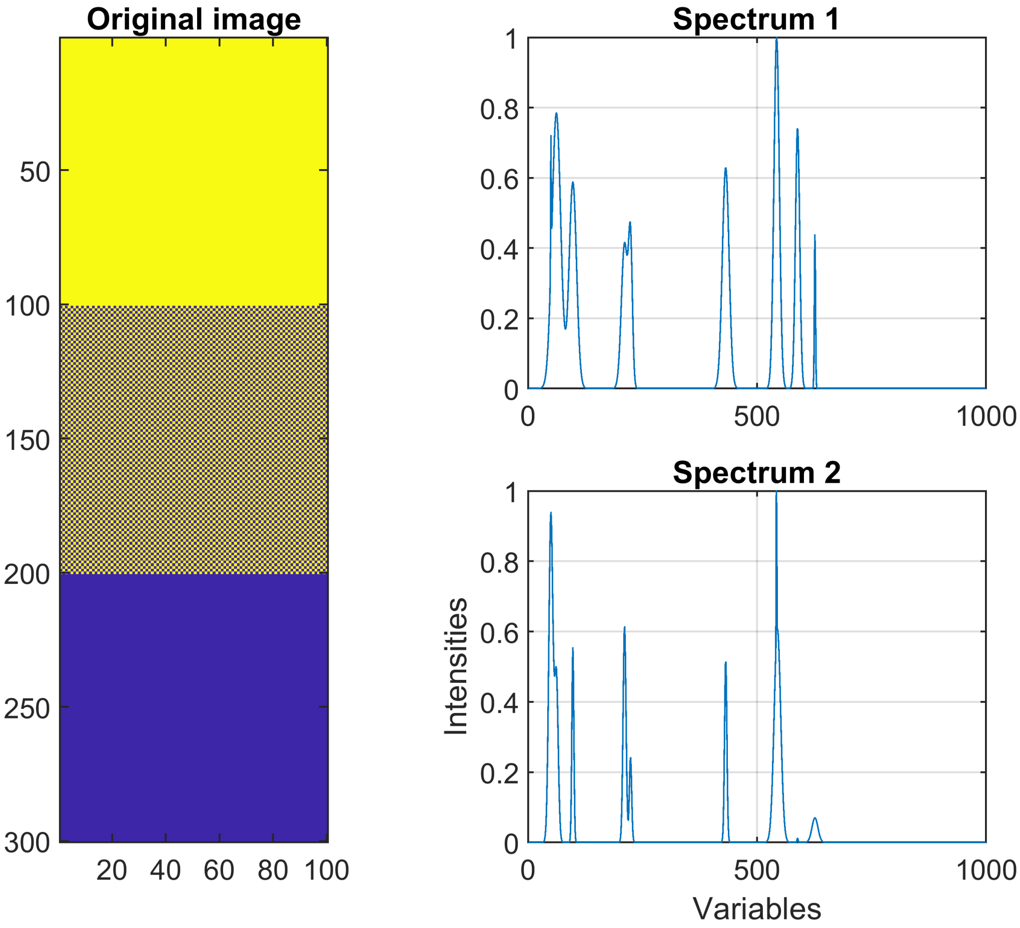


Figure 40: on the left, the yellow pixel are related to the spectrum 1, the blue pixel are related to the spectrum 2

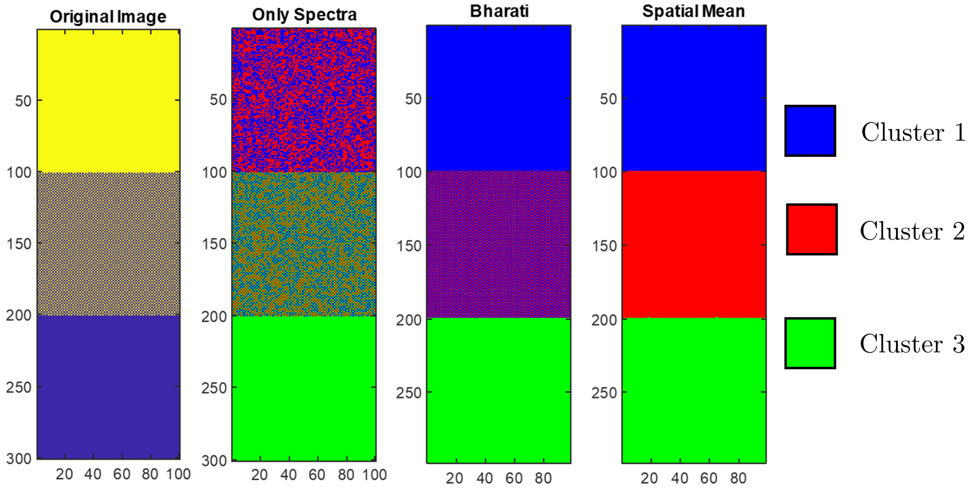


Figure 41: The first image on the left is the original image. "Raw data" corresponds to the segmentation map obtained by kmeans applied to the raw data. "Bharati" corresponds to the segmentation map obtained by kmeans applied to the data treated with the Bharati approach. "Mean Spatial" corresponds to the segmentation map obtained by kmeans applied to the data treated with the spatial mean approach.

The results of kmeans applied to the raw data are complex as expected, because considering only the spectral information, there is no reason why there are 3 different clusters. The situation is different for Bharati and the spatial mean. Even though in this simple toy example Bharati seems to capture the information about the neighboring pixel, the segmentation does not correspond to the expected one. The upper and lower regions are correctly separated, but in the middle region, where there is a mixture with 5 pixels of spectrum 1 plus 4 pixels of spectrum 2 and vice versa, it assigns one of these two situations to the cluster corresponding to the upper region, while the other is a separate cluster. This does not respect our expectation, so we can say that this method fails in this case, even if it is simple. Different are the results of the spatial mean, where the three different regions are better are properly separated, with only a few pixels at the border that are not well assigned, probably due to the noise.

Let's make another example. Looking at the figure 42 below.

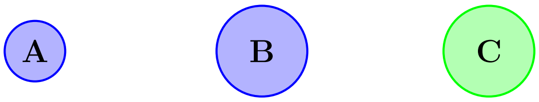


Figure 42: Region A, B, and C. Color represent the main information, radius the dispersion

Within a tissue, there is a region A and a region B with the same main information but different standard deviation shared by the neighboring pixels. This means that for both, the center of the two distributions is similar, but with different dispersion. Applying the spatial mean will group these two regions together. However, in the context of exploring cancerous tissue, perhaps this discrepancy in variance is the factor that the second region is an early stage of cancer, so it will be nice to separate it from the previous one, but this is not possible with this approach. For this region, another way of describing the neighboring pixels is needed. This descriptor should describe the dispersion around these pixels. The standard deviation is the simplest descriptor that can be used. The spatial dispersion is used to approximate the standard deviation applied to the neighboring pixels. However, using only this spatial dispersion will separate A and B, but if there is a region C with different main information but the same dispersion as region B, it will be grouped together with region B. So the segmentation is not optimal in both cases. Considering the fusion of the spatial mean and the spatial dispersion, this should separate these 3 clusters because it groups together regions with not only similar main information but also similar dispersion. Now the expressed case should be simulated. As said before, the idea was to have two types of spectra. The two spectra were simulated and shown in figure 43 below:

Second simulation

The second simulation aims to replicate three regions based on two different spectra, albeit in a different manner to the first simulation.

The spectra were constructed by keeping the position of the peaks but changing the intensities and width of the peaks. This was done on the assumption that the difference between a healthy tissue and a tumor tissue is only a relative composition. Figure 44 shows the map with the original labeling (3 regions) and the relative mean

Matlab

```
num_p = 1000;
x = linspace(0, 1000, num_p);
n_peaks = 10;
mu = randperm(800,n_peaks)+10;
% Peaks generation
for z=1:2
    spectrum = zeros(size(x));
    for i = 1:n_peaks
        sigma = rand*10;
        amplitude = rand*10;
        spectra{z} = spectrum +amplitude* exp(-((x - mu(i))
           .^2) / (2 * sigma^2));
    end
    % Normalization
    spectra{z} = spectra{z} / max(spectra{z});
end
map = [ones(100,100);
        repmat(2,100,100);
        repmat(3,100,100)];
map = reshape(map,300*100,1);
n=10000;
matrix2=zeros(n,num_p);
matrix2(map==1,:)=repmat(spectra{1},num_sp,1)*10+...
   randn(n,num_p)*2;
matrix2(map==2,:)=repmat(spectra{1},num_sp,1)*10+...
   randn(n,num_p)*6;
matrix2(map==3,:)=repmat(spectra{2},num_sp,1)*10+...
   randn(n,num_p)*6;
```

Figure 43: Matlab script for the spectra generation of the second simulation

spectrum and standard deviation spectrum.

Kmeans is applied on the raw data, data processed with spatial mean, data processed with spatial dispersion, and fused data (spatial mean and spatial dispersion).

As expected, kmeans with k=3 applied to raw data, spatial mean and spatial dispersion does not retrieve the three clusters. Regarding the first two, this is due to the fact that from a spectral point of view, there are only two clusters in the original image. There is no linear way to separate B from A. The spatial dispersion does not retrieve the three clusters, as there is no such information in the spatial dispersion. Nevertheless, by combining the spatial mean and the spatial standard deviation, the

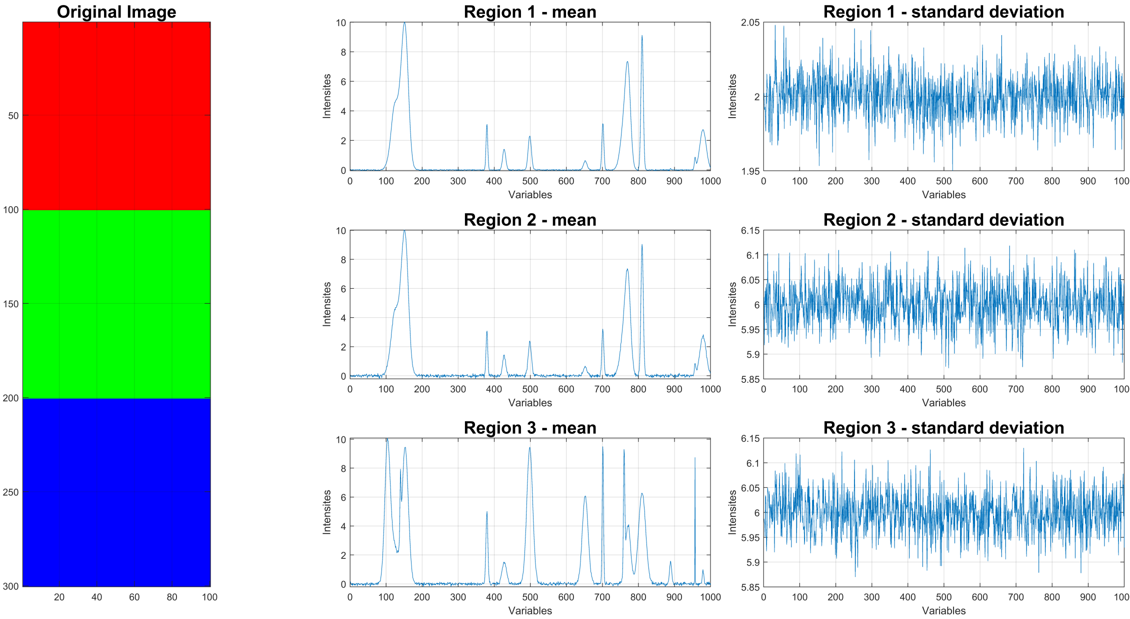


Figure 44: Simulation. 3 different region: red, green, and blue. Red and green have the same mean spectrum but different standard deviation. Blue has the same standard deviation of region green but a different mean spectrum.

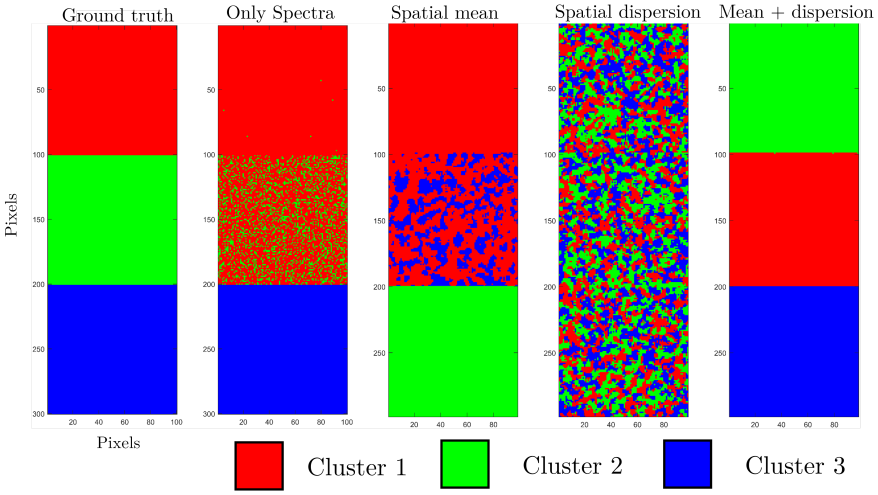


Figure 45: Raw data labelled, spatial segmentation of raw data, spatial mean, spatial dispersion, and fused data.

three data points can be plotted as in figure 42. In this case, a linear separation is possible. In this case, the spatial mean and spatial dispersion should properly retrieve two clusters. Applying kmeans with $k = 2$ to the spatial mean and spatial dispersion, the segmentation maps obtained are shown in figure 46. As shown in figure 46, the segmentation does not work with the standard deviation. This is due to the correlation metric; the scale factor does not matter in this case, so there is no difference between regions 1, 2 and 3. In fact, if the metric used were normal Euclidean distance, the results would be different, as shown in figure 47 below.

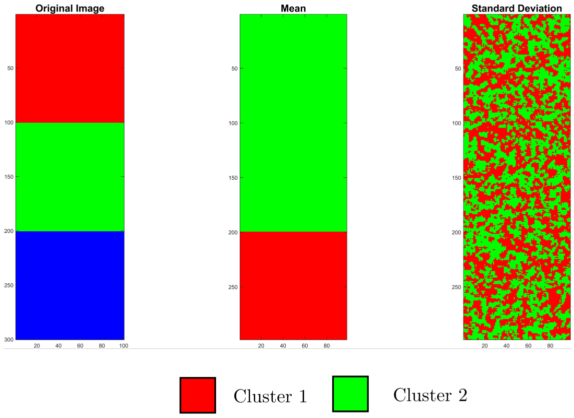


Figure 46: kmeans with k=2 applied to spatial mean and spatial standard deviation

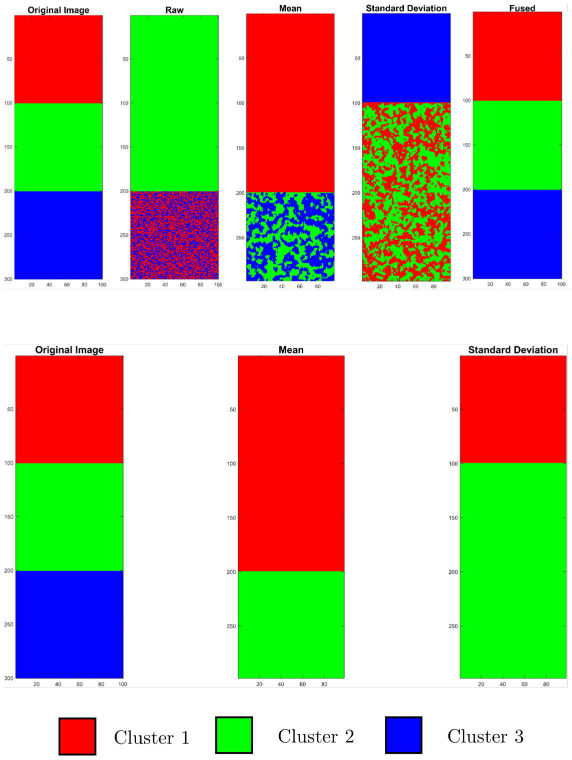


Figure 47: Results of this simulation with euclidean distance

Real Sample

The sample used was one of the 16 analyzed in the article [44]. The collection, preparation and acquisition of the samples were expressed in the previous section, as the sample used if one of those 16. It is now time to apply and compare these approaches to a real case. The idea was to make a comparison between 5 cases: (1) the clustering algorithm (kmeans) applied to the spectra, (2) the spectra enhanced with the Bharati approach, (3) the spatial mean, (4) the spatial dispersion (using the standard deviation) and (5) the fusion between spatial mean and spatial standard deviation. This comparison was done with the aim of discussing the meaning of the clusters obtained, not with the aim of obtaining a "perfect" segmentation map. For this reason, on a sample where it is not possible to know the correct number of clusters, the results of kmeans from k=2 to 10 were calculated and then the Calinski-Harabasz approach was used to determine the optimal number of clusters according to the input spectra used. As it is shown in figure 48 below, the comparison can look quite unfair, as when using the raw spectra, the optimal number of clusters is 3, while when using the fused data, e.g. the mean and standard deviation of the neghiborhood pixels, the optimal number is 8. However, this is also a point for the proposed approach. From a statistical point of view, using only the spectra, there are 3 clusters. But using the information of the mean and the dispersion of the neighborhood pixels, 8 is the optimal number of clusters. So, using a statistical criterion, the fused information gives a higher number of clusters to describe the data. This can prove what was discussed before with the simulated data: it can potentially give more information about the sample and describe it better. As stated before, the goal was not to prove the best clustering algorithm, but to prove that even with a simple clustering algorithm, extracting information about the neighborhood can improve the analysis. Figure 48 below shows the optimal segmentation map based on the CH criteria. A deeper interpretation of the results is needed to validate the results from a biological point of view. However, some initial considerations can be made by looking at the segmentation maps. The spectral segmentation map is used as a reference to show the results obtained using only the spectral information. The Bharati segmentation map, with the optimal number of clusters equal to 3 according to the Calinsky-Harabasz criteria, does not give any improvement (compared to the map obtained with "raw" spectra) in the segmentation map, except for a certain smoothing effect.

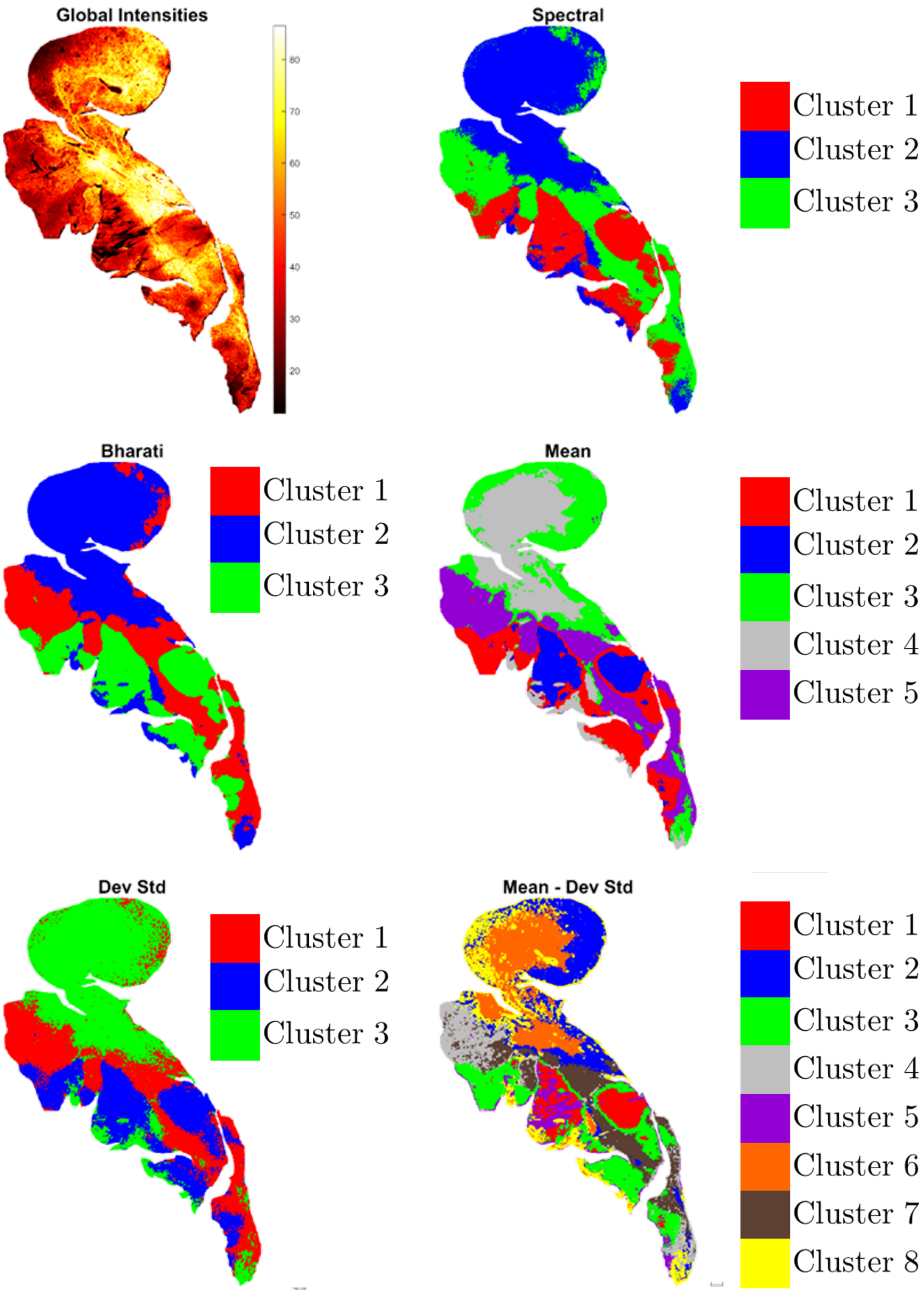


Figure 48: Results of the cluster validation

The only general observation that can be made is that looking at the centroids gives information not only about the "central" spectra, but also about the neighboring pixels. The segmentation map of the spatial mean, with an optimal number of clusters equal to 5 according to the Calinsky-Harabasz criteria, provides a deeper description of the sample. Here, the centroids represent the mean spectrum of the 3x3 grid around each pixel. The centroid is still related to the main information, the concentration. The standard spatial deviation map is quite different. Here, the cluster should not be read as a grouping of similar spectra, but as a grouping of spectra that have a homogeneous spatial dispersion. So for each cluster, the spectra differ from the neighborhood in the same way. This means that really different spectra can be together if they differ from their neighbors in the same way. Finally, the merged map "mean - dev std", with the optimal number of clusters equal to 8 according to the Calinsky-Harabasz criteria, suggests that the real structure of the sample is probably more complex. The centroid here has a part related to the neighborhood mean spectrum and the neighborhood standard deviation spectrum. Each cluster should be read not only as a group of similar spectra, but also as a group of spectra with similar neighborhood and similar dispersion through the neighborhood.

To validate the results, it is important not only to look at the segmentation map, but also to analyze the information contained in the clusters, what the clusters represent. For this reason, it is important to look at the centroids and compare them. There are several ways to compare clusters. Here, the idea was to use a Permutation Features Importance (PFI) [56] to identify the most interesting peaks that characterize the partitioning of the data. This approach is more commonly used in supervised techniques to weight the importance of features in the model, but can also be adapted to clustering algorithms. The idea is to randomly permute the values of each feature, feature by feature, and calculate the difference based on the validation method used. In the field of clustering, the Quality Cluster Criteria can be used. For each m/z, 100 permutations were used. The CH index was calculated for each permutation. At the end of the loop, the mean is calculated. These values are divided by the original CH to get a measure of the increase or decrease of the index. In this way, it is possible to understand which variables have the greatest impact on the segmentation.

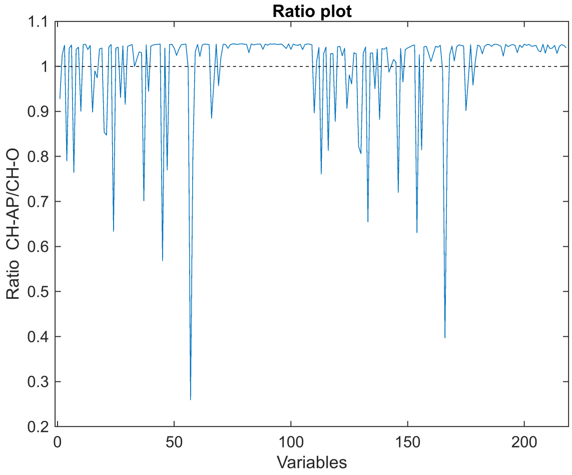


Figure 49: Ratio plot of FPI. The dashed line indicate a ratio equal to 1. Lower values indicates an downgrade of the criteria, so that negative peak indicate how much that variable is important for the clustering process

The first 10 m/z values with the highest impact on the quality criteria were taken. The order is sorted from the m/z with the highest effect (top) to the one with the lowest (bottom).

mz value	type of info
2300.0056	mean
2300.0056	dispersion
2099.8628	mean
2099.8628	dispersion
1809.7550	mean
1809.7550	dispersion
1981.8482	mean
1981.8482	dispersion
1419.5449	dispersion
1485.6070	dispersion

Table 9: The 10 most relevant m/z values for the clusterization of the sample

It can be interesting to see that for the first 8 signals there is an alternation between the signal coming from the spatial mean and the signal coming from the spatial dispersion. The last two are two separate mz values from the spatial dispersion signal.

However, this was just to test a way to get the most meaningful m/z values for the clustering results. These results should also be validated from a biological point of view, with a proper interpretation.

2.3.1 Discussion, conclusions and future perspective

As the simulations above demonstrated, spectral information alone was not always sufficient to answer the research question. In certain cases, spatial information proved essential. While using only spectral data is not inherently incorrect, it is important to consider the context and objectives of the analysis. In both simulated data, there can be two solutions: 2 clusters or 3 clusters. Both are the right answer, but it should be contextualized properly. In case of biological samples, with a spatial resolution of the same order of cells, a pixel can be considered as a cell. Then it can be said that the same cell can have a different meaning, depending on its environment. That was the point of this chapter. Another considerations can be done about the dimension of the neighbourhood explored. The spatial information in this work was extracted from a 3x3 grid. However, it might be interesting to use a larger grid, such as 5x5 or 7x7. Additionally, the mean and standard deviation were calculated over all pixels in the grid, so the central pixel is taken into account. The results of the spatial mean are an image-denoised cube. Excluding the central pixel when calculating the spatial parameters may be a better choice. This allows the spectra to be fused with the different spatial parameters, where the spatial information is calculated using only the neighbouring pixels. In this case, the parameters are the mean and standard deviation. In addition, since the idea was to describe the distribution of neighboring pixels, the extraction of spatial information can be enhanced with kurtosis and skewness, or other parameters that can improve the description. In the end, considering only two parameters for the spatial information will increase the spectral dimension by a factor of 3. This is not the initial goal, but the results obtained justify the increase in dimensionality.

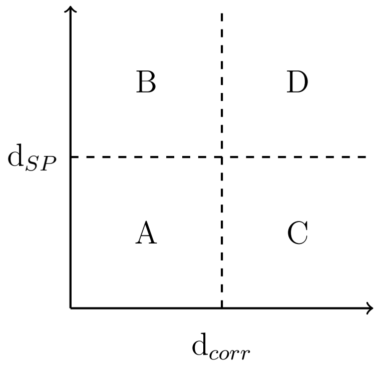
The previous approach gave the same weight to all neighboring pixels. An extension of this version can be to personalize the weighting, so that instead of calculating the normal mean, a weighted mean is calculated using only the neighboring pixels. This information can be added to the original information of the pixel, so that at the end you have a "super-spectrum" composed of the original information of the pixel and the weighted information of the other pixel. An example of weighting is given by the equation 20

$$\omega_i = \frac{1}{d_{corr}} \cdot \exp(-\lambda \cdot d_{sp}) \tag{20}$$

Instead of focusing only on the neighboring pixel, a broader approach can be used. By generating a matrix with the spatial position, it is possible to calculate the spectral similarity and the spatial similarity for each distance pair. This information is used to calculate the previously introduced weight (equation 20).

1. Take a spectrum n.
2. Compute the correlation distance between spectra n and all other spectra
3. Compute all the spatial distances between the positions of the pixels
4. Combine the two distances to calculate the weight of each pixel
5. Compute the weighted average of all spectra (except the n spectra)
6. Compute the weighted standard deviation of all spectra (except spectra n)

In this way, both spatial and spectral distance contribute to the weight. The tricky part here is to properly attenuate the mixture of these two indicators.



There are 4 different situations: A, both spatial and spectral distance are small. B, small spectral distance and large spatial distance. C, small spatial distance and large spectral distance. D, both spatial and spectral distances are large. An approach using a weighted k-medoid clustering algorithm was done. The idea was to try to incorporate the spatial information extraction process into the clustering algorithm. However, if two distant regions have the same neighborhood, there is no need to correlate all pixels. Even the simple 3x3 grid can guarantee the connection between these two regions, so this might complicate the purpose too much without adding significant gain.

2.4 General conclusion and perspective regarding chapter 2

During this chapter, the challenge of clustering in the MALDI imaging domain was addressed by working on three approaches in parallel. These approaches are not mutually exclusive. Fusing the three approaches by using spatial information with an auto-QDD kmeans could potentially produce representative results. The decision to work on them separately was made because it represented a progressive evolution of competence and knowledge of the problem, rather than being based on the One Variable At a Time (OVAT) approach. Now that the first section has been submitted, the goal is to complete the other two parts (auto-bisecting and spatial information), after which all the improvements arising from this thesis work will be fused into one single approach. This will balance the influence of the different parameters.

Chapter 3

Contribution of hyperspectral data fusion in LIBS and MALDI imaging

Using multiple sources to better understand a sample is common in many fields. For example, a physician may use MRI (magnetic resonance imaging), blood tests, urine tests, and other analyses to make a diagnosis. Similarly, in chemistry, it is common to analyze a sample using multiple techniques. The key challenge is to effectively manage and integrate these multiple sources of information. Data fusion can be divided into different categories [57]. The most common divide the data fusion approach into three categories: (1) Measurement level *or* low-level fusion; (2) Feature level fusion *or* medium-level fusion; (3) Decision level fusion *or* high-level fusion. In the first case, *measurement level*, the different blocks are fused as they are and analyze as a single block. In the second case, *feature level*, features are initially extracted from each block and then these features are merged and analyzed, especially for a regression or classification algorithm. The latter case, *decision level*, the predictions of the different blocks are fused together. In hyperspectral imaging (HSI), data fusion involves concatenating different hyper-images. In the LIBS imaging domain, for example, multiple spectrometers can be used simultaneously to acquire different spectral domains [58–60]. This makes data fusion straightforward, as no image registration is required. Conversely, when data are acquired using different systems, an image registration step is necessary prior to analysis. This is due to potential differences in spatial resolution, rotation and translation, which can prevent precise pixel-to-pixel correspondence between different imaging modalities. The image registration process addresses these issues to ensure proper alignment and correspondence between spectra from different sources. In recent years, the literature on HSI data fusion has become increasingly important. Bedia et al. [61] proposed a multimodal chemometric approach based on multivariate resolution to simultaneously analyze spectroscopic images (MALDI-TOF MSI, FT-IR, and RGB images) of tumor tissues revealing distinctive chemical patterns among tumors and demonstrating the effectiveness of fused image analysis to extract high-resolution molecular information. Another study combined MALDI imaging with microscopy through image fusion to improve the spatial detail of molecular distributions [62]. Recently Tuck et al. [63] proposes multimodal imaging as an advanced strategy for integrating vibrational spectroscopies and mass imaging to achieve more comprehensive chemical and morphological characterization of biological tissues in biomedical and multiomics. Desbenoit et al with a mini-review [64] highlights how the integration of MALDI-MSI with other imaging techniques can enrich the morphological and chemical interpretation of tissues, illustrating analytical and computational strategies to address multimodal imaging challenges. LIBS have proven to be useful coupled with other techniques such as Raman [58, 59], PIL[60] , or histological image [65]. LIBS and mass spectrometry it was already explored, but with other purpose and other approaches. A study coupling inductively coupled plasma–mass spectrometry (ICP-MS) and libs was done [66], but it was employed to chemically map and characterize uranium particles. Another study coupling ICP-MS and LIBS [67] showed that trace elements can be detected using the LA-ICP-MS domain of the setup while major components of the samples are analyzed simultaneously using LIBS. The study highlights how the multimodal approach, integrating LA-ICP-MS and LIBS, can overcome the limitations of individual techniques, offering more complete and detailed elemental mapping of biological samples.

This chapter presents two works related to hyperspectral data fusion: (1) The first is the fusion of RGB image and LIBS imaging of a mortar sample; (2) The second is the fusion of MALDI and LIBS imaging applied to mouse brain and kidney tissue. The first fusion was chosen to prove that adding RGB the interpretation of the LIBS data can be improved, and without any additional cost, since we naturally use a capture system that can be seen on virtually all microscopes. At the same time, the interest in fusing the MALDI and LIBS resides in the improvement of biosample interpretation, linking molecular and elemental information.

Image Registration

Image registration is a mandatory step when two or more HSIs are acquired using different systems, or when with the same system two different optical pathway are used for two different spectral domain. Different acquisition conditions result in spatial misalignment. This misalignment prevents perfect pixel-to-pixel correspondence between the modalities. Therefore, an accurate registration process is required to align the images and establish pixel-to-pixel correspondence. The algorithm uses a *moving* image and a $fixed$ one used as a reference. Such algorithm tries to maximize (or minimize) a criterion by applying different transformations to the moving image. The criteria used in our case is the Mutual Information, criteria usually used in medical multimodal image registration [68, 69]. In general, mutual information is defined as the amount of information shared between two variables [70]. In image registration, mutual information quantifies the extent to which knowledge of one image can be used to infer information about the other. It is calculated from the joint histogram using entropy formulas. The aim is to maximise this value, as this indicates that the images share the most information and are best aligned. This evaluation is applied to the moving image after each transformation (e.g. rotation, translation or scaling). The algorithm stops when it finds the maximum value of mutual information. Registration is usually performed on a 2D image. In the case of HSI, where there are more channels, there are many possible combinations for registration. For example, when considering two generic HSI cubes, the global intensity image can be chosen, as well as a specific peak. Selecting the image that is most similar to and contrasts most with the other modality helps the registration and makes it more robust. However, the two selected images cannot simply be used as they are because, as is obvious, they can have a very different range. Therefore, prior to registration, min-max normalization is required to scale the hyperspectral image range to grayscale (range between 0 and 1). This provides consistency in intensity ranges across different modalities and improves the robustness of the registration. The scaling is done using the equation 21

$$im_{(i,j)} = \frac{im_{(i,j)} - \min(\mathrm{im})}{\max(\mathrm{im}) - \min(\mathrm{im})} * 255 \tag{21}$$

Once the optimal transformation has been obtained, it is applied to all the spectral channels of the HSI cube that correspond to the moving image. This establishes a pixel-to-pixel correspondence with the reference image. Once this has been achieved, the data fusion strategies and all the processing algorithms can be applied to the data. Depending on the data strategy chosen for the problem, different preprocessing methods are required for each block prior to concatenation [42]. In our case *low-level* fusion was used. Prior to the concatenation, each cube was normalized by its Frobenium norm (equation 22)

$$||X||_F = \sqrt{\sum_{i,j} |X_{ij}|^2} \tag{22}$$

where $x_{i,j}$ is the value of the pixel at i, j position. The Frobenius norm is the extension of the Euclidean norm for the vector
In table 50 is showed the matlab code per image registration.

Matlab

```matlab
[optimizer,metric] = imregconfig("multimodal");
tform = imregtform(MOVING,FIXED,"similarity",optimizer,metric);
imreg = imwarp(MOVING,tform,"OutputView",imref2d(size(FIXED)));
cube_reg = zeros(size(cube_to_reg));
for i=1:size(cube_to_reg)
cube_reg(:,:,i) = imwarp(cube_to_reg(:,:,i),tform,"OutputView",imref2d(size(FIXED)));
end
```

Figure 50: Matlab script for image registration. *imregtform* is a function included in **image processing toolbox**

3.1 LIBS and RGB

Mineral samples are difficult to analyze and understand. The differences between them are more related to the ratio of the different elements and these differences can be really small. The idea was to show that even though PCA is such a sensitive technique, and LIBS is also sensitive, capable of detecting down to ppm concentration, sometimes the complexity of these samples is so great that the normal analysis, such as PCA, fails to give a comprehensive view of the samples. Adding color information can help to get a deeper understanding of the sample. The idea was to combine visual information with the LIBS one. RGB is the simplest type of multivariate image Red $\lambda \approx 630$ nm, Green $\lambda \approx 545$ nm, Blue $\lambda \approx 435$ nm, these wavelengths being chosen to match the spectral response of the human eye [71]. Each channel has a range of 0 to 255. Distances between colors in RGB space do not necessarily correspond to perceived color differences. RGB images, due to their simplicity and availability of low-cost acquisition, are widely used in numerous application areas, from the food industry [72, 73] to quality control to environmental and biomedical analysis [61]. The RGB analysis allows the extraction [74] of morphological and color information useful for classification, segmentation [75]. and process monitoring tasks [76].

Experiment

This study focuses on aligning and registering the RGB image with the LIBS image. The primary goal is to enhance the interpretation and understanding of the data by merging these two imaging techniques. For instance, as will be demonstrated later using quartz crystals as an example, members of the same crystal family may appear different colours in the RGB image. However, LIBS spectra alone do not provide sufficient information to determine whether these colour differences are due to slight variations in composition, trace elements or other factors. This limitation highlights the importance of integrating RGB imaging to complement the LIBS data. The main objective of this approach is to use information from the RGB image to enhance the analysis of the LIBS image. In particular, the RGB data facilitated the creation of masks to identify specific crystals as it will be showed later, which was a non possible task using LIBS data alone. These masks enabled the differences between the crystals to be studied in detail. In this case study, the LIBS system was configured to acquire data in two spectral regions simultaneously, thereby increasing the potential for detailed compositional analysis. The RGB image and the two LIBS images were imported into MATLAB (version 2023a, MathWorks®, Natick, MA, USA). The PCA approach was chosen for the unsupervised analysis because of its simplicity and ability to extract the most important information, as well as the minor compounds, from the sample. To compare the information obtained by applying PCA to LIBS data only with the information obtained by combining LIBS and colour data, the process was run in parallel.

Sample

The analyzed sample is a piece of mortar taken from the internal masonry of the apse of the ancient part of the church *Saint-Irenee*, on the hill of Fourviere in Lyon (France). It is a lime mortar mixed with a quartz and siliceous coarse sand with some limestone grains. It has been prepared as a thin section to be observed with transmitted light on a petrographic microscope. Figure 51 shows the original church and the polarized image of the mortar sample.

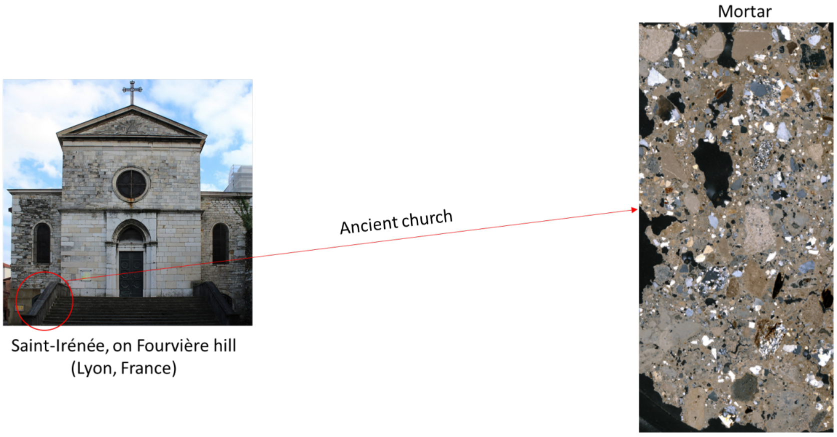


Figure 51: Origin of the mortar sample

Data Acquisition

The μ-LIBS system used consists of a 1 kHz Cobolt Tor XE pulse laser (Q-switched, emitting at 1064 nm with 0.5 mJ/pulse). Plasma emission is measured using two spectrometers that can observe several wavelength ranges simultaneously, allowing to focus on the lines of all the elements of interest. The first one is a compact Avantes spectrometer (Evo Sensline XL) configured from 238 to 357 nm, for main major and minor elements: Mg, Al, Si, Ca, Ti, Mn, Fe, Cu mainly. The second spectrometer is a Kymera (equipped with an sCMOS camera, Andor), recording spectral ranges between 672 and 833 nm for completing the lack elements, as Na and K. Acquisition was performed at room temperature under ambient pressure conditions. An argon flow blowing through the plasma was used both to enhance emission and to prevent surface contamination due to deposition of material ablated by previous laser shots. Finally, the sample was placed on an XYZ stage and scanned pixel by pixel with a lateral resolution of 15 μm, resulting in a 2000 by 1000 spectra for a 2 million pixels image. Due to the amount of information, attention was focused on the section highlighted in the figure 52. This region (composed of 220x240 pixels, 52,800 spectra) was initially selected because of the heterogeneity it showed in the RGB image, while during the LIBS analysis this heterogeneity was not observed. The original idea was to try to find out the color differences from an elemental point of view.

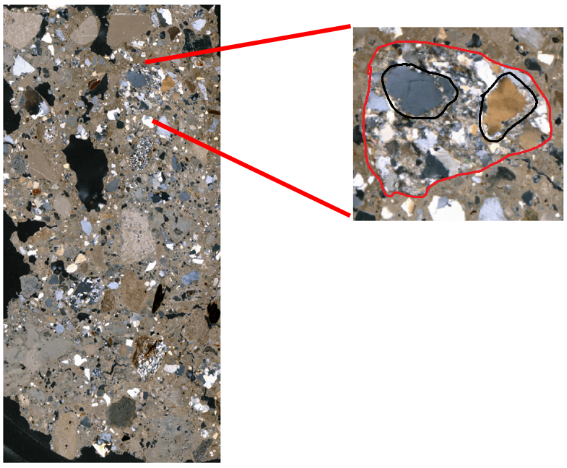


Figure 52: From the original image, a section was extracted and analyzed. Red indicates the shape of the mask used in the following figures. The blacks show the two crystals under consideration.

3.1.1 Results and discussion

The first approach was to apply PCA to the LIBS data, once a mask was generated. To generate the mask, the silicium signal was used. In the results shown in figure 53 below, it is not possible to distinguish all the crystals that appear in the RGB image.

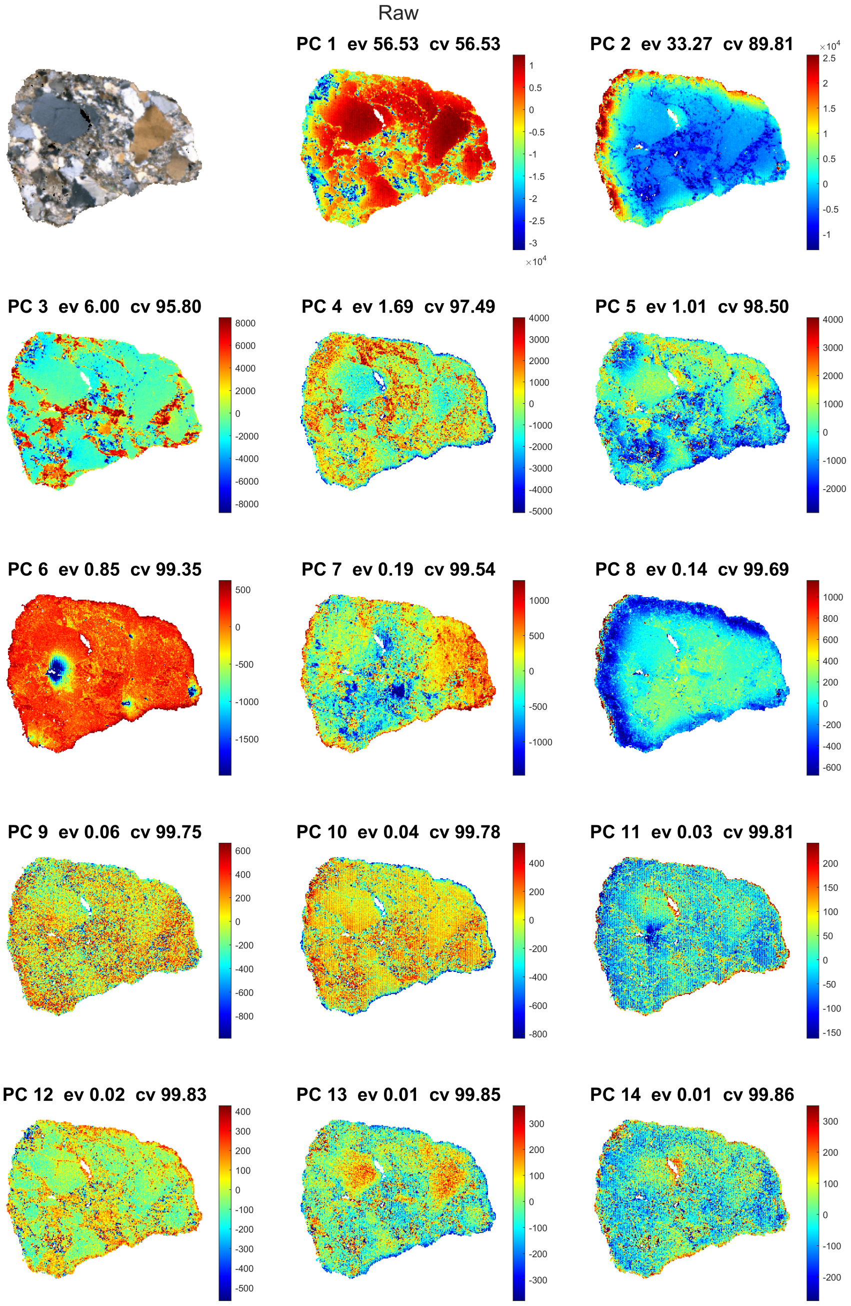


Figure 53: Score maps from PC-1 to PC-15. PCA applied to the LIBS data

Fusion

For this reason, the LIBS data were fused with the RGB data. For the moving image, the grayscale RGB image was used. The grayscale image of the global intensities image was used as the fixed image. The obtained registration parameters are: scale

0.8734; rotation angle 7.7697°; translation [–287, 248]. Once the transformation has been obtained, it is applied to all channels of the RGB image. Before concatenating the data, each block was normalised using the Frobenius norm. Then, the PCA is applied to the fused data.

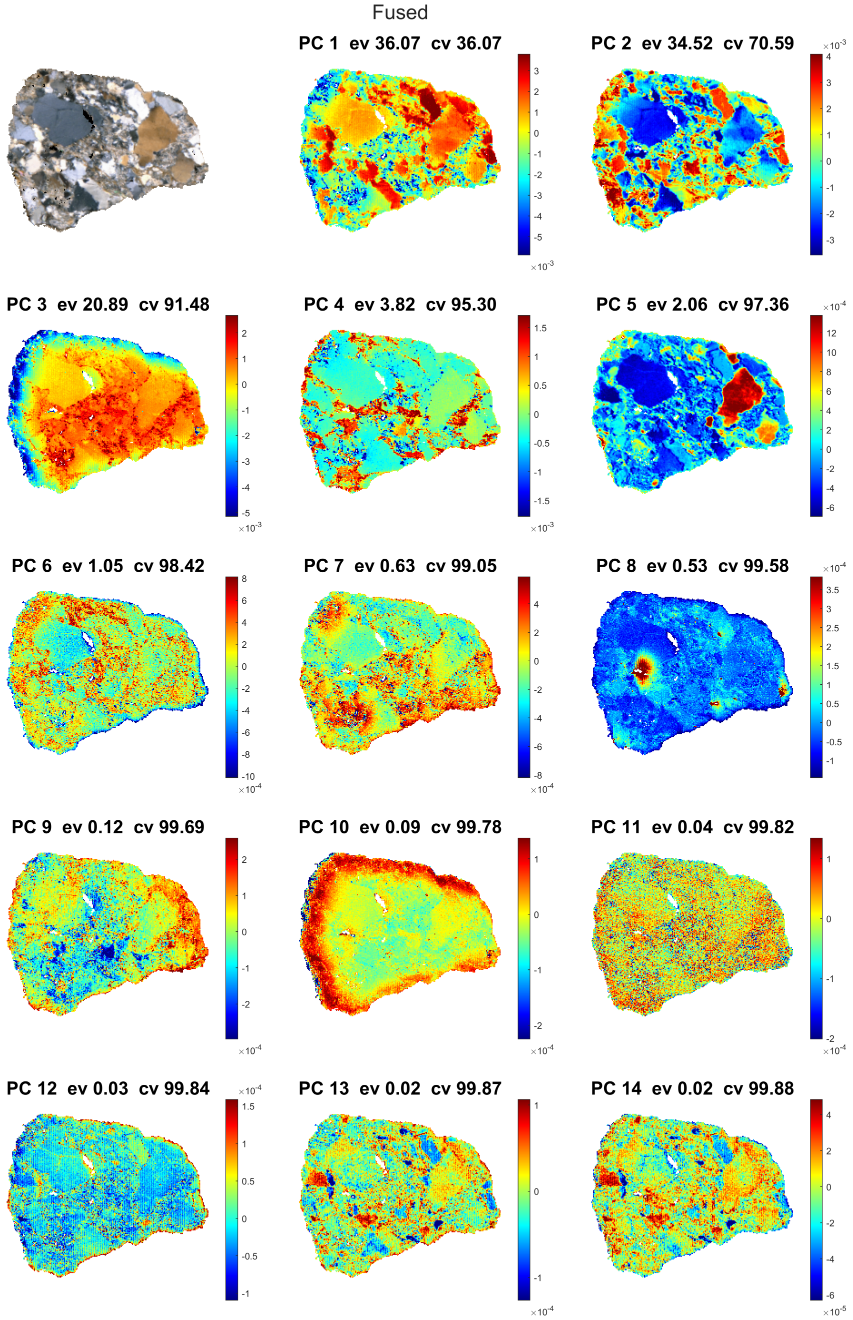


Figure 54: Score maps from PC-1 to PC-15. PCA applied to the fused data.

Compared to the results obtained using only LIBS data, noticeable differences can be observed. Starting from the PC-1 and PC-2, the core maps seems to highlight different structures inside the main crystal. However, the most interesting point from our perspective is showed in PC-5. Here, it is possible to discriminate between the two crystals that showed opposite values. Looking at figure 55, the LIBS data loadings have low intensities, suggesting that most of the variance is due to color.

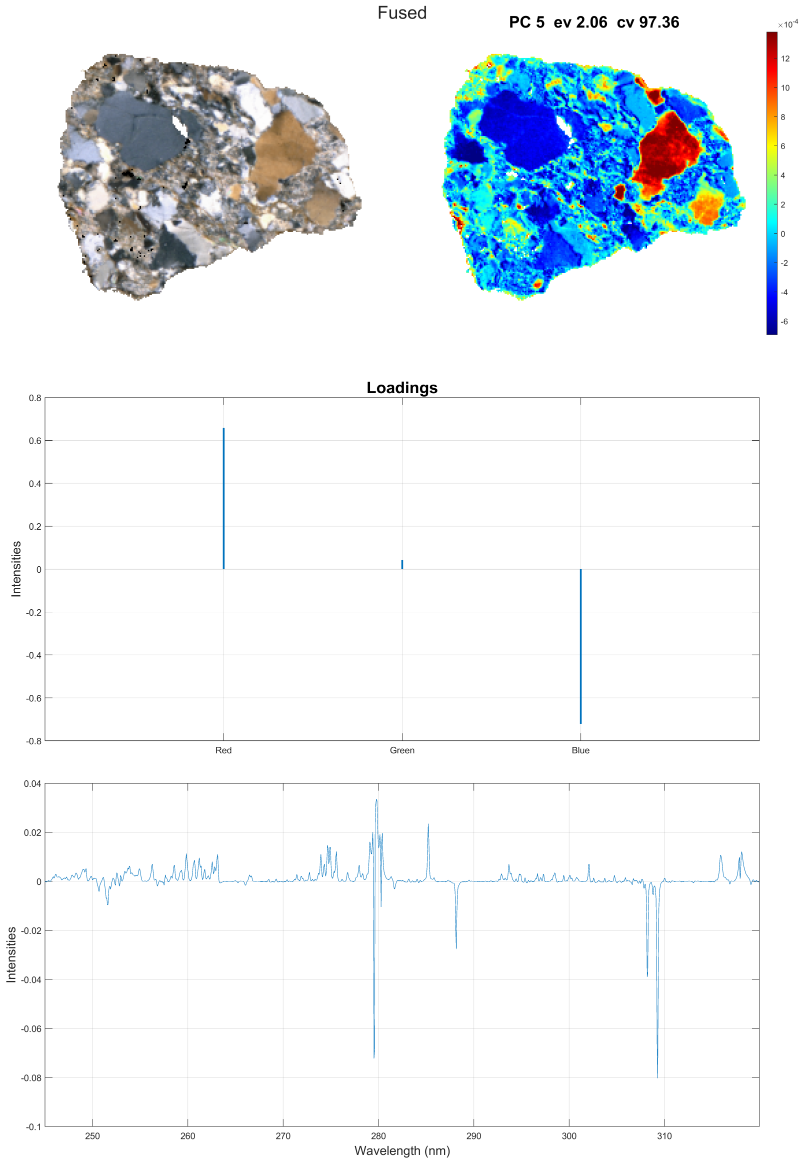


Figure 55: RGB image, scores map and loadings PC-5 of PCA applied to the fused data

In order to explore more the two crystals, a mask was generated using the scores of the PC-5. Using the mask it was possible to isolate the pixels of the two crystals. PCA was calculated on the LIBS data and the fused data using this mask. Even in this case PCA applied to only LIBS data did not discriminate clearly the two crystals (see figure 56 below). The score maps shown are until the PC-15, with a cumulative variance of 99.02%.

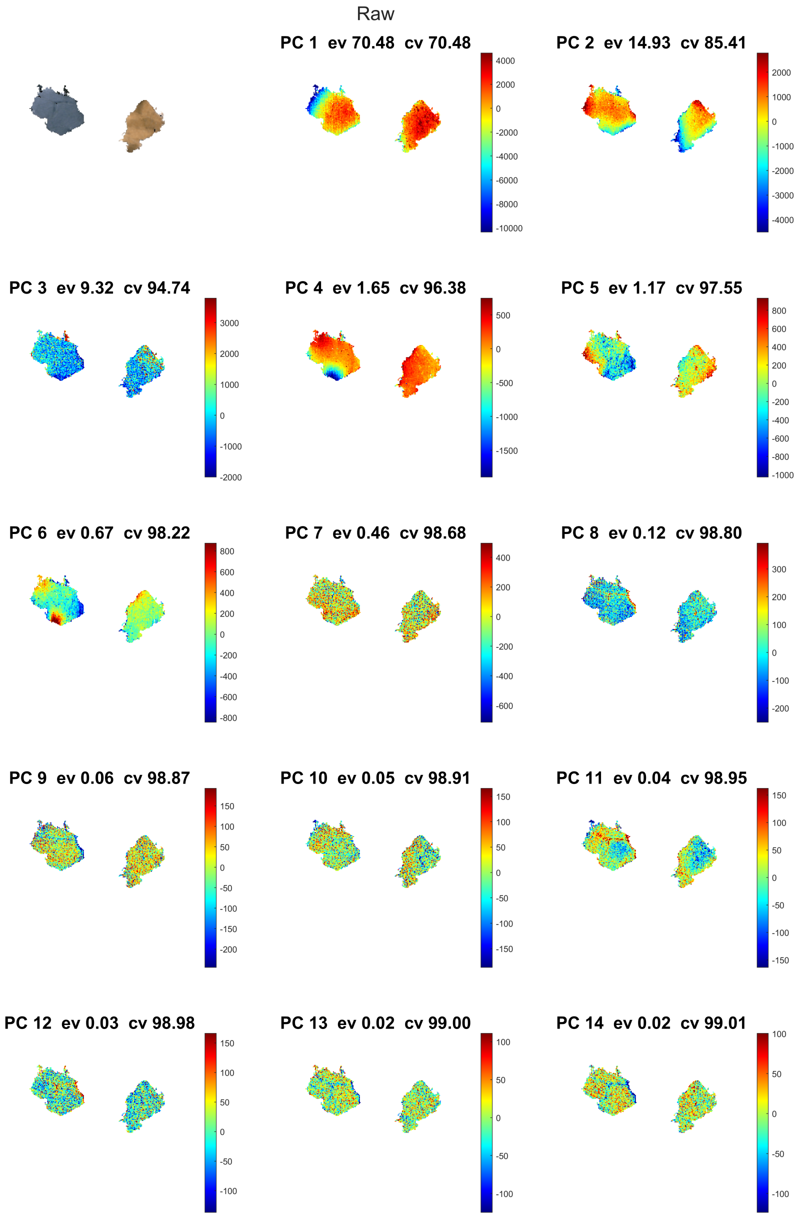


Figure 56: RGB image of the quartz crystals and score maps from PC-1 to PC-14. PCA applied to the LIBS data.

Applying the PCA to the fused data allowed this discrimination with the PC-1 as shown in figure 57. Here, the discrimination was possible with an explained variance of 61.13%. Even in this case the score maps shown are until the PC-15, with a cumulative variance of 99.74%.

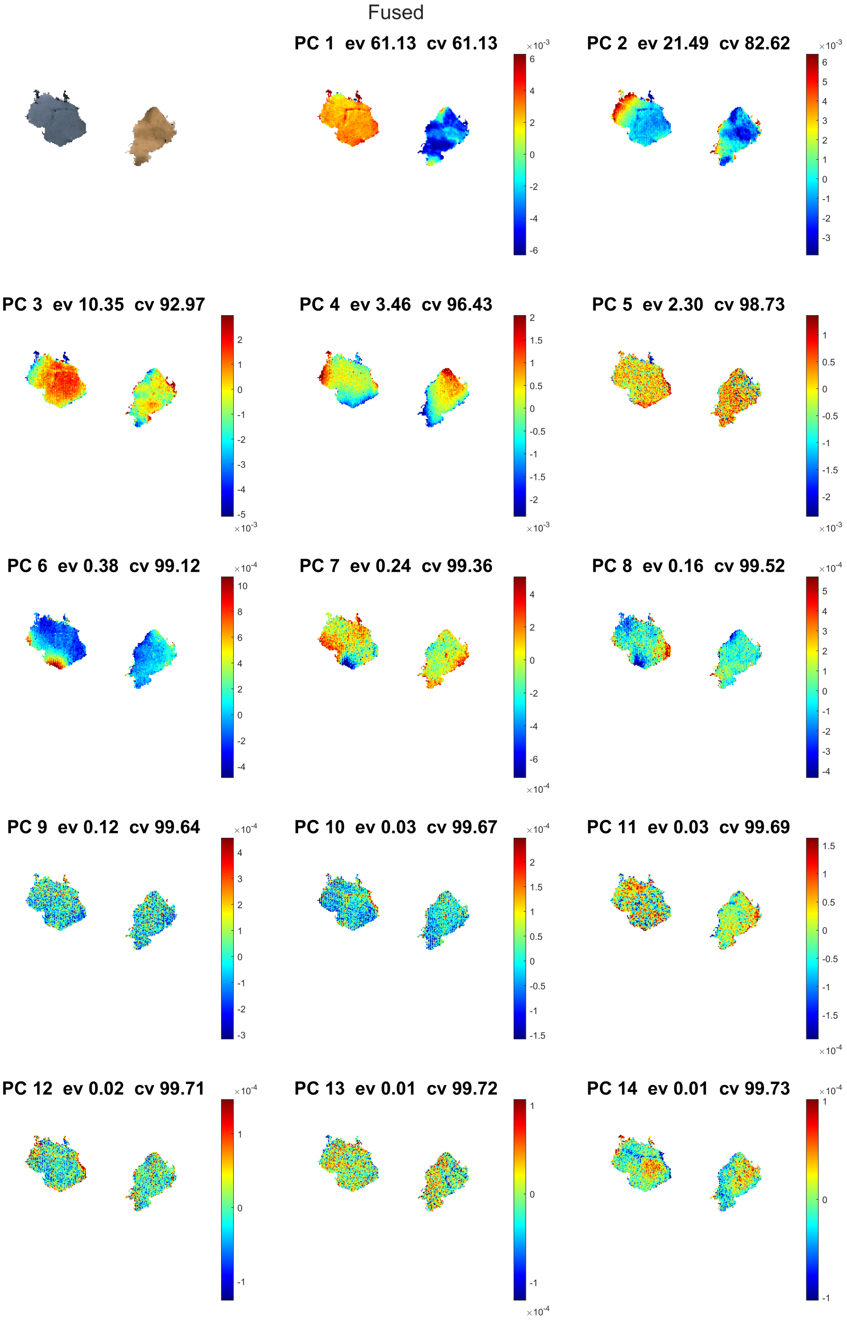


Figure 57: RGB image of the quartz crystals and score maps from PC-1 to PC-14. PCA applied to the fused data.

As expressed previously, PCA applied to the fused data allowed for a clear separation between the two crystals with the PC-1, which explains 61.13% of the total variance. To investigate more the reason behind this discrimination, in figure 58, the score map and the loadings are reported.

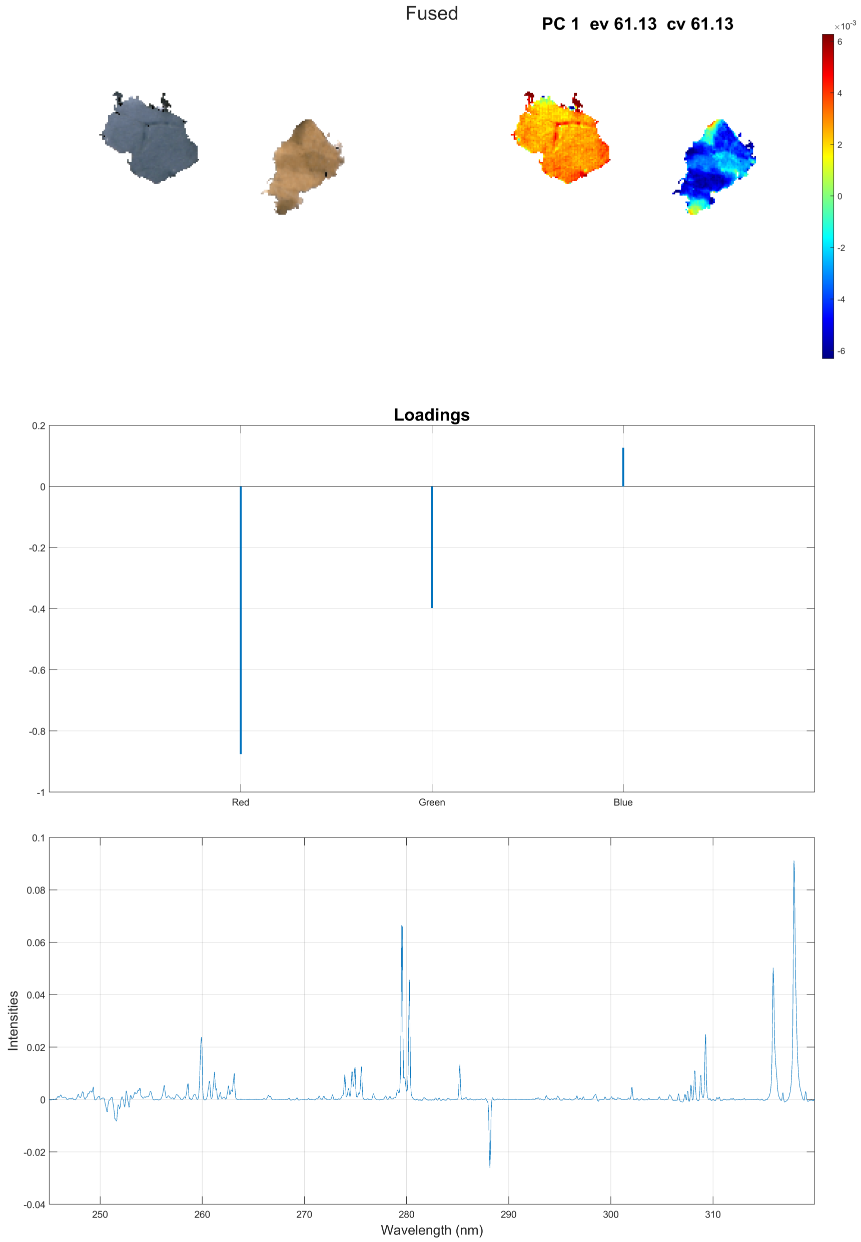


Figure 58: RGB image of the quartz crystals, score maps and loadings of PC-1. PCA applied on the fused data of the 2 crystals.

As it was showed before, this separation could not be achieved using LIBS data alone. However, the integration of RGB information enables us to distinguish the two regions, suggesting that color carries relevant information. The loading plot shows that both the LIBS signal and color channels contribute to PC-1. The crystal with a negative PC-1 score appears redder and is characterized by a higher silicon (Si) signal in the LIBS spectrum. The crystal with a positive PC-1 score appears bluish and shows stronger contributions from Mn, Mg ionic, Mg, Al, and Ca. This indicates that, even though RGB alone may appear simplistic, it actually guides us toward meaningful chemical distinctions that are otherwise harder to capture. This relationship will be further demonstrated in figure 59, through the analysis of the mean spectra, variance spectra, and the complete spectral sets for each crystal. It is possible to directly compare the spectra of the two crystals using the mask. Figure 59 shows the mean spectrum, the variance spectrum, and the spectra of the two crystals. Looking simultaneously at the spectra shown in Figure 59 and at PC-1 shown in Figure 58, some observations can be reported. Peaks around 280, 288, 310, and 316 nm appear to be the main differences between the crystals. These peaks, in fact, are highlighted in PC-1, showing a positive score for peaks 280, 310, and 316 nm, which are therefore related to crystal 1, and a negative score for peak 288 nm, related to crystal 2. In fact, looking at Figure 59, the peaks at 280, 310, and 316 nm show higher variance in crystal 1 compared to crystal 2. Additionally, close to 310 nm, multiple peaks seem to be present in crystal 1. These peaks are not clearly visible in the mean spectrum, but are quite evident when looking at all the plotted spectra and the variance spectrum.

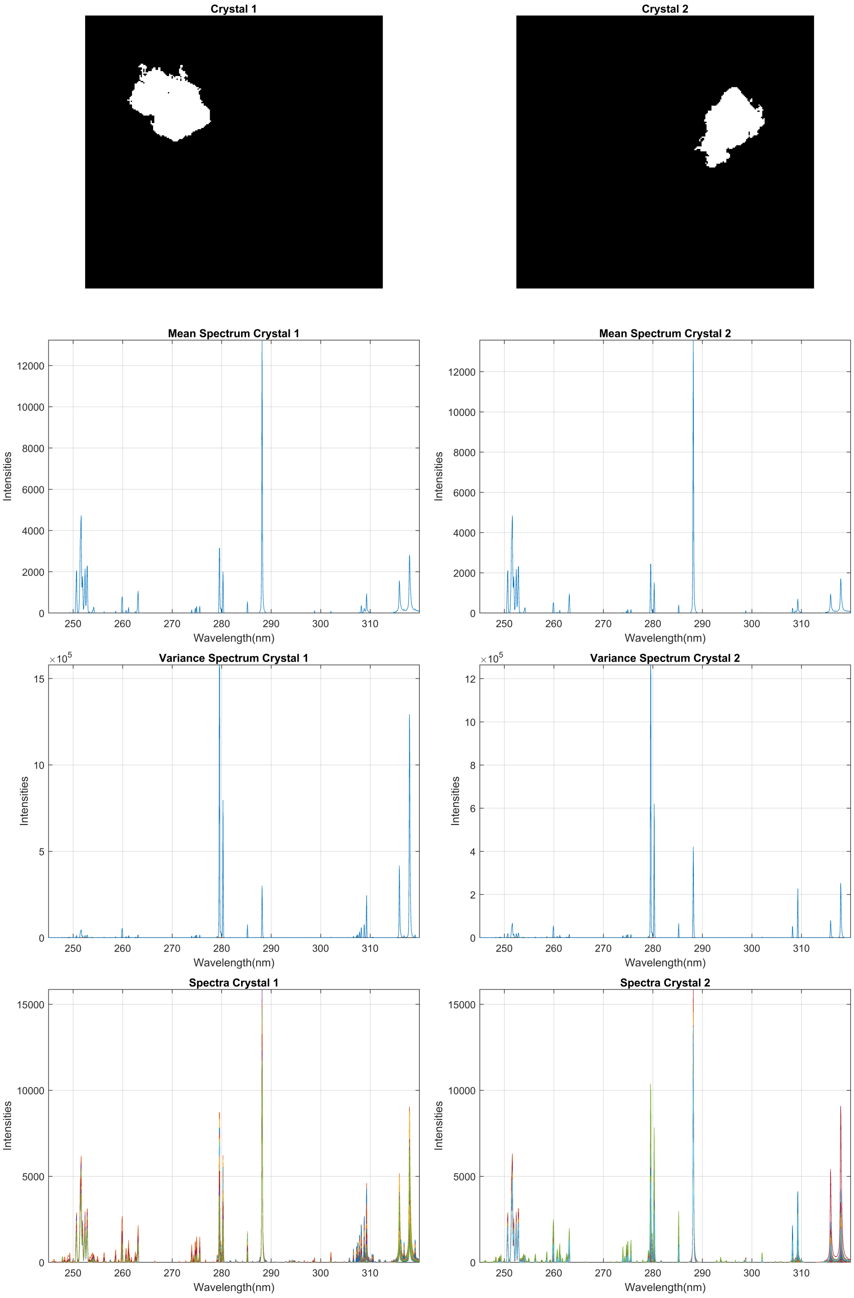


Figure 59: Mask of the crystals: Mean spectrum, Variance spectrum, and spectra

Conclusions and future perspectives

The work demonstrated how combining visible image and LIBS information can provide more detailed insights and this without any additional cost, since we naturally have a capture system that can be seen on almost all light microscopes. Although LIBS can capture trace elements, differences in the compositions of different crystals are sometimes so small that standard exploratory analyses, such as PCA, cannot detect them. Adding colour information where the differences are more pronounced makes it possible to distinguish more zones easily. In this case, coupling PCA with an examination of the mean and variance spectra of the two crystals revealed the differences between them. It is also important to note that the image was taken with a polarised light microscope. Changing the polarised light causes a change in the colour of the recording due to the light's different refraction. In the future, it may be interesting to have more than one polarised image with different degrees to see if the distinction can be improved by a simultaneous use of several visible images fused with LIBS data. This approach could be tested for other purposes, notably in the preparation of radiocarbon dating of mortars. Indeed, to do so, it is necessary to discriminate between the various carbonates in the mortar, in particular the binder carbonate, which must be separated from the carbonate grains (limestone, shells, slaked lime, etc.) that contaminate the dated samples and bias the ages. A difference in LIBS signal, mainly due to a difference in density of these materials, helps discrimination, but it does not work completely, and identification errors may remain in the segmentation. This is particularly the case when carbonate grains come from porous limestones which provide a signal similar to that of the binder. The combination of μLIBS imaging and optical microphotography colors could ensure better identification in this framework. At the beginning of this work there was the will to enhance the color information with an augmentation, using an approach similar to Cocchi et al. [77], but a different approach was pursued. There are other ways to represent the color, that can be easily obtained from the RGB. A color space [78] is a way of representing color in a numerical format that a computer, or more generally an electrical device, can interpret. There are different color spaces, and each emphasising different aspects of color perception or reproduction, making them useful for different applications. Color spaces can be divided into additive spaces, which are based on light combinations (e.g., RGB), and perceptual or decorrelated spaces, which separate brightness and chromaticity (e.g., LAB, HSV). As will be demonstrated later, the color space plays a key role in enhancing the information that can be extracted from the data. The color spaces used in this work are listed in the table 10.

RGB	R	Red
	G	Green
	B	Blue
HSV	H	Hue
	S	Saturation
	V	Value/Brightness
lab	L	Luminance
	a	Green to Red
	b	Blue to Yellow
XYZ	X	Approximate Red-Green perception
	Y	Brightness perception
	Z	Approximate Blue perception
NTSC	N	Luminance
	I	Orange-Cyan Chrominance
	Q	Green-Magenta Chrominance
YCbCr	Y	Luminance
	Cb	Blue-Yellow Chrominance
	Cr	Red-Green Chrominance

Table 10: Colorspaces

HSV (Hue, Saturation, Value) provides a more intuitive representation of color. It is not perceptually uniform, saturation behaves differently depending on brightness. LAB separates lightness (L) from chromaticity (a and b). LAB is a perceptually uniform color space, meaning that the Euclidean distance between two colors in this space correlates well with how we perceive their difference. The XYZ color space, based on the response of the human eye, is device independent and is widely used in color space conversions. NTSC (National Television System Committee) has historically been used for television color encoding in the U.S., based on luminance and chrominance. YCbCr is a color space used primarily for video compression and broadcasting. It separates brightness (Y) from color information (Cb and Cr), allowing for more efficient compression by downsampling the chrominance components without significantly affecting visual quality. This separation mimics the human visual system, which is more sensitive to brightness than to color.

3.2 MALDI and LIBS: coupling molecular and elemental information to enhance interpretation of biological samples

LIBS has proven to be an effective technique to detect both endogenous and exogenous elements in brain tissue, with the advantage of not requiring complex sample preparation [79–83]. Alongside it, more specialized methodologies such as LMD-ICP-MS (Laser microdissection inductively coupled plasma mass spectrometer) have been employed to obtain very high resolution elemental images, for example in the hippocampus of rat brain, allowing the distribution of trace metals to be analyzed with greater precision [84]. Additionally, LA-ICP-MS (Laser ablation inductively coupled plasma mass spectrometer) has been widely applied in neurodegenerative disease models, with applications ranging from Parkinson's disease to stroke, and can be combined with MALDI-MS to identify metalloproteins [85]. A comprehensive review further summarizes ten years of advancements in LA-ICP-MS bioimaging for trace element analysis across various biological systems, highlighting its role in biomedical and environmental research [86]. Finally, a dedicated review of LIBS instrumentation and methodology outlines recent developments, including tandem approaches and chemometric strategies, that extend LIBS applications to a broad range of analytical fields [87]. These complementary techniques underscore the growing importance of multi-modal imaging in neurobiological research and support the rationale for integrating molecular and elemental data despite technical challenges. However, to our knowledge, no existing instrument allows the simultaneous acquisition of MALDI and LIBS images. Since both techniques are destructive, it is not feasible to analyze the exact same tissue section with both modalities. Therefore, this study uses two consecutive tissue sections, assuming their differences are negligible. Image registration procedures are applied to ensure proper spatial alignment between modalities. In this case two consecutive sections were analyzed, one section for MALDI and the following one for LIBS. The MALDI acquisition was done at the Leiden Medical University, while the LIBS acquisition was done in Lyon, at the ILM (Institut Lumiere Matiere). The main challenge of hyperspectral data fusion lies in the way images are acquired. There is a problem, though: it is not possible to obtain the images with the same instrument, meaning the data does not come from the same exact sample. The solution to the this problem could be, as adopted in this thesis, to consider tiny consecutive tissues and consider them as equal. Even if the two consecutive sections are considered equal, there is a general problem of image registration. An image registration step is mandatory.

Sample

The sample analyzed was a sagittal slice of rat brain. Both MALDI and LIBS images were acquired with a spatial resolution of 30 μm.

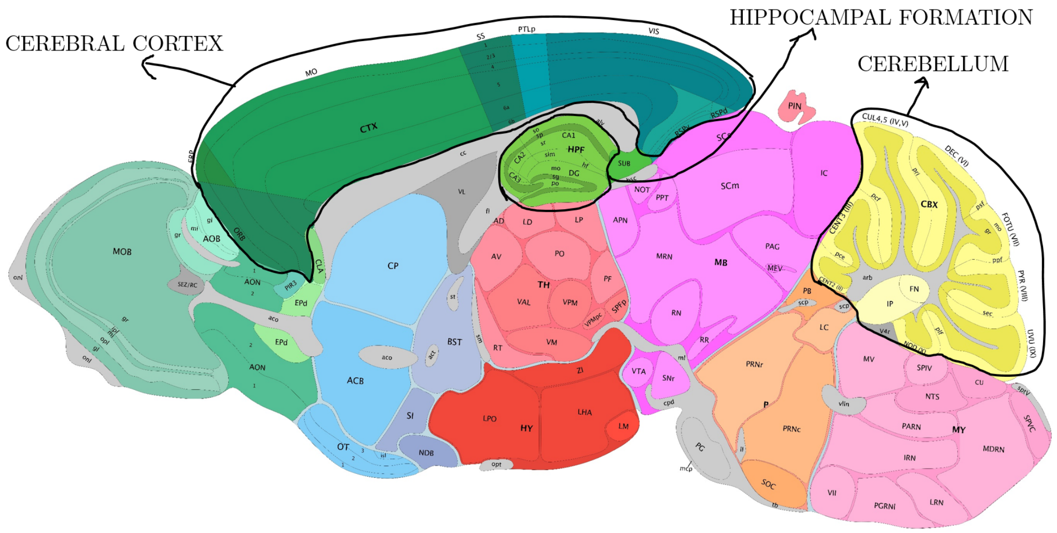


Figure 60: Atlas Sagittal Rat Brain

The main areas were annotated based on ATLAS (https://atlas.brain-map.org).

LIBS

The experimental LIBS setup utilized in this study is located at the ILMTech - OPTOLYSE platform (ILM-UMR 5306, CNRS Université Lyon1, Villeurbanne, France). Its main characteristics have been detailed in our previous work [88]. Briefly, a Nd:YAG laser (1064 nm) operating at 1.25 mJ was employed. The laser-induced plasma emission was collected using two spectrographs to simultaneously record two distinct spectral regions (201–233 nm and 265–349 nm). Specifically the

2 spectrograph are Czerny-Turner (Shamrock 303i and Shamrock 500 Andor technology) equipped with 2 ICCD cameras (iStar DH340T-18F-E3, Andor Technology) . Optimal delay and integration times were determined to be 0.1/5 μs and 1/5 μs, respectively, with a gain of 1750 (arbitrary units). The plasma was generated under a 0.8 liter/min Ar flux.

Looking at figure 61 Cerebral region and hippocampus are clearly visible with Ca and Mg signal. The Cu signal is slightly different. Looking at Mg signal there are saturations, and this can affect analyses such as PCA [89]. This will cause non-sample related variance to be captured by analyses such as PCA, obscuring the variance of interest and making it difficult to interpret the results.

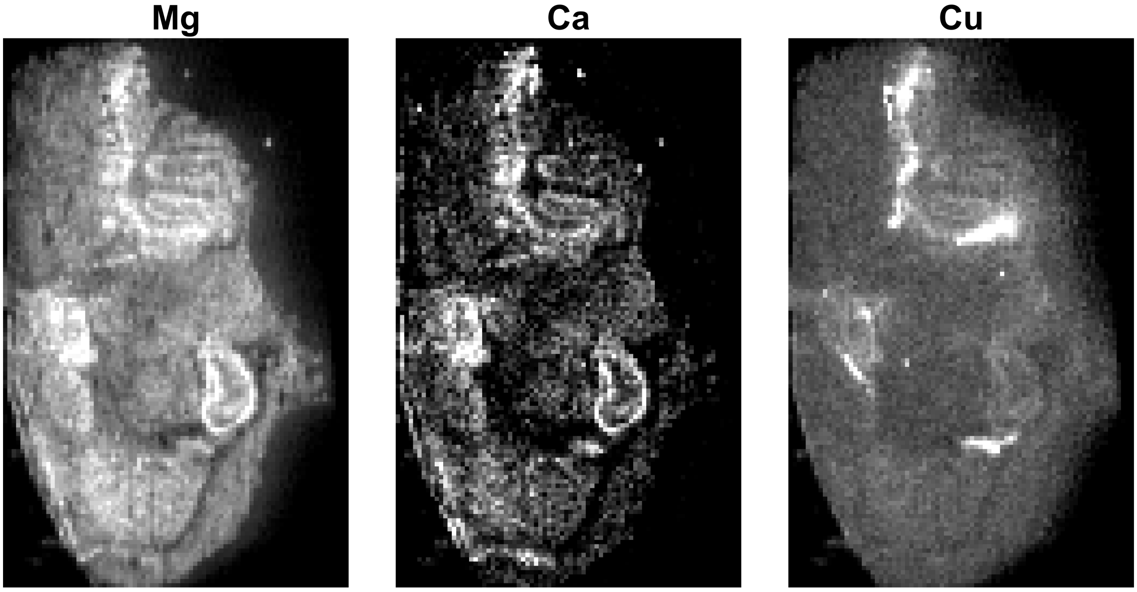


Figure 61: Chemical map of different elements. The signal were chosen by the expert in Lyon. The peaks are: Mg 285 nm, Ca 318 nm, Cu 325 nm.

Therefore, a proper preprocessing step is required. The first step was to create a mask to reduce the dimensionality of the data and speed up the analysis. To create a mask, the signal of Mg was used. Then the baseline correction is applied on LIBS spectra. The Local Asymmetric Least Squares (LALS) algorithm, available at Lovelace's Square (https://lovelacesquare.org/), was used to apply the baseline correction. The parameters were estimated on a small fraction of the spectra to speed up the analysis. Once the results were deemed acceptable, these parameters were applied to the entire data set. At this point, we need to deal with saturation. To do this, we need to look at the spectrum. All the signals that have intensities higher than 55000 are close to the limit of the detector, so we can consider them as saturated or distorted. These values have been replaced by NaN (Not a Number) values. At this point, the matrix contains missing values. The approach to face this challenge was recently proposed by Gómez-Sánchez et al. [90]. The algorithm was applied to the data set containing the missing values with 200 iterations and 6 principal components, initially estimated as the first 6 PCs obtained from the data without considering the pixels containing missing values. The 6 PCs were chosen after an evaluation combining the visual evaluation of the scores, loadings and eigenvalues. This process gives an "approximation" of all the data, not just the missing values. The original NaN values (where there were sauturations) are changed with the values obtained with this method, in order to do not change values with lower intensities.

MALDI

Ten micron thick tissue sections were cut from each of the frozen tissue samples using a Leica CM3050S cryostat (Leica Biosystems), thaw-mounted onto conductive indium-tin-oxide (ITO)-coated glass slides (VisionTek Systems), and stored at -80°C until use. Before use, slides were placed in a vacuum freeze-drier for a minimum of 20 minutes. The tissue sections were sprayed withthe 7 mg/mL Norharmane matrix in 90%MeOH using the HTX-Sprayer. The tissue section was subjected to high-resolution, accurate-mass (HRAM) imaging using a 12·T solariX MALDI-FTICR mass spectrometer (Bruker Daltonics). The MALDI-FTICR was operated in negative-ion mode, using a 1M or 2M transient length, and set to detect a mass range of 300 to 2000 m/z. Each HRAM lipid profile was collected from 150 laser shots, using a laser repetition rate of 500 Hz and with the laser focus set to "small."

Fusion

The fixed image is the global intensity image of the MALDI, the moving image is the global intensity image. Once the registration is obtained, it is applied to the entire LIBS cube. Then a spatial binning was applied to reduce the small discrepancy that comes from the two images. In this way, even if the spatial resolution is lost, the relationship between the pixel information from the two modalities is stronger. The registration parameters are: Scale 0.9863, Rotation angle -4.8117, Translation

-38.0722, -7.0967. The data are now merged.

3.2.1 Exploratory Analysis: Results and Discussion

The first approach was to apply PCA to the MALDI and LIBS data individually and then to the fused data. This was done to see if fusion would improve understanding of the sample and reveal patterns not shown by the individual PCA analyses and to obtain information about the correlation between elements and molecules. In the following two pages the results of the PCA applied individually will be showed. The score maps of the PCA applied to only LIBS data are not very informative. The PCA is applied to only LIBS data after registration and using the mask that comes from MALDI; in this way, the outside region is not taken into account. As can be seen from the score maps of the first 9 PCs in figure 62, it is quite difficult to appreciate the overall structure of the brain. Only PC-1, which accounts for 98.82% of the explained variance, highlights the Mg signal and the tissue edges. PC-7 notably highlights the Cu signal. These images do not provide much additional relevant information. The score maps of the PCA applied to only MALDI data (figure 63) better highlight the different internal structures of the brain. The main exception is PC-2, which highlights the differences between the edges of the tissue and its internal regions. A deeper interpretation of these images, as the link between the different molecules and their spatial distribution, can be made, but it is not the purpose of this study.

Moving to the results of the fused data, it can be seen that different primarily structure of the brain such as cerebellum, fiber tracks and hippocampal formation can be seen among all the PCs shown in figure 64. However, even with the mask, coming from MALDI cube, used, more than one PCs seem to highlight mainly the variability between the tissue and its edges, as it can be seen mainly in PC-2 and PC-3, lower in PC-4. As can be seen from the loadings (figure 65), the contribution of each component is unequal between MALDI and LIBS. For example, the maximum range for the MALDI loading is from -0.5 to 0.4, whereas for the LIBS loading it is from 0 to 0.17. This disparity is also evident among the other components. Additionally, the LIBS loadings are consistently positive or negative until PC-4. In PCs 5 and 7, a peak shift can be seen for the two peaks at 280 and 285 nm. In PCs 6, 8 and 9, these peaks are reversed, indicating a difference between the ionic (280 nm) and elemental (285 nm) forms of magnesium. Additionally, for all the components, the impact of the other LIBS peaks are really low, compared even to the magnesium peaks.

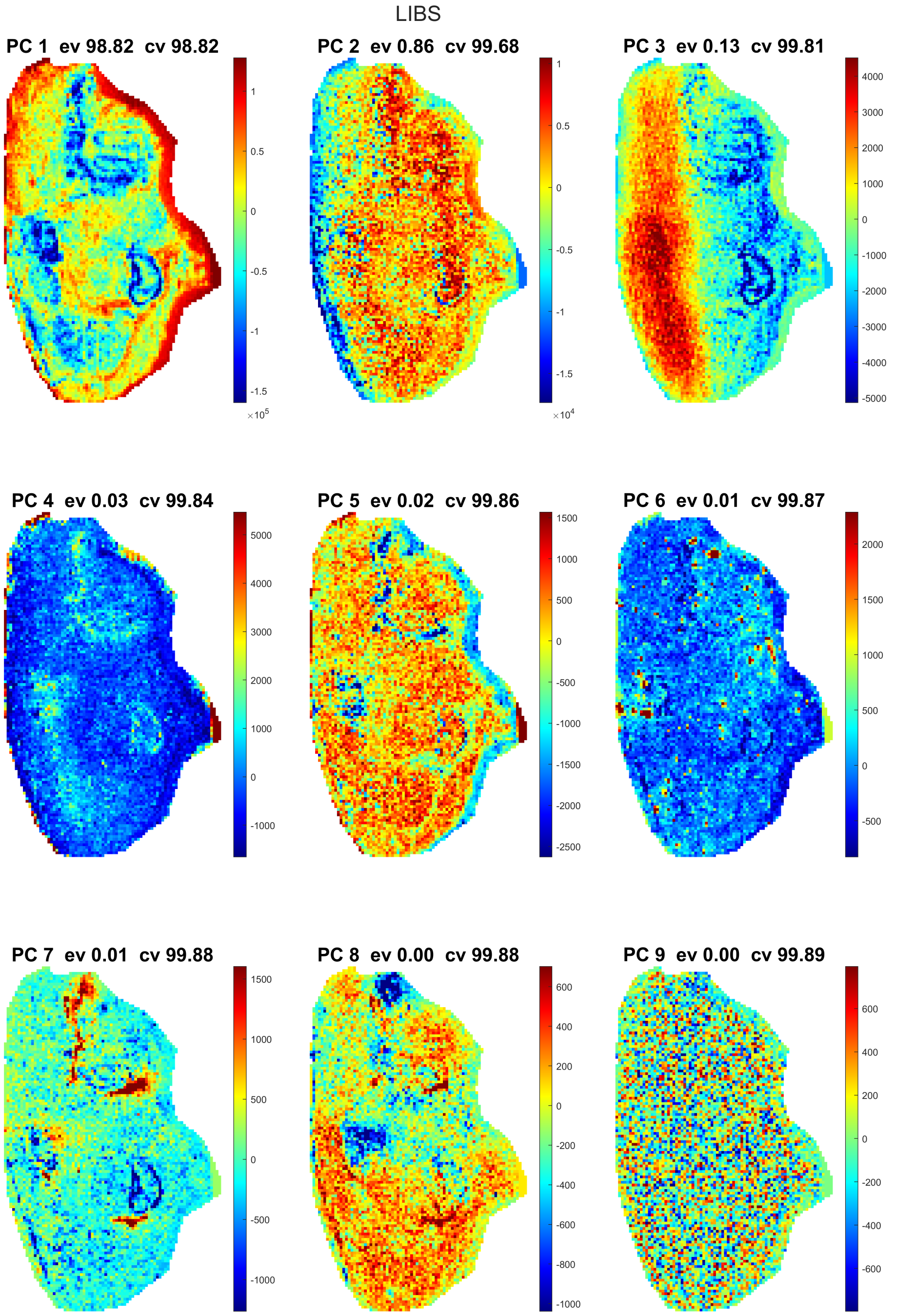


Figure 62: Score maps from PC-1 to PC-9. PCA applied to the LIBS data.

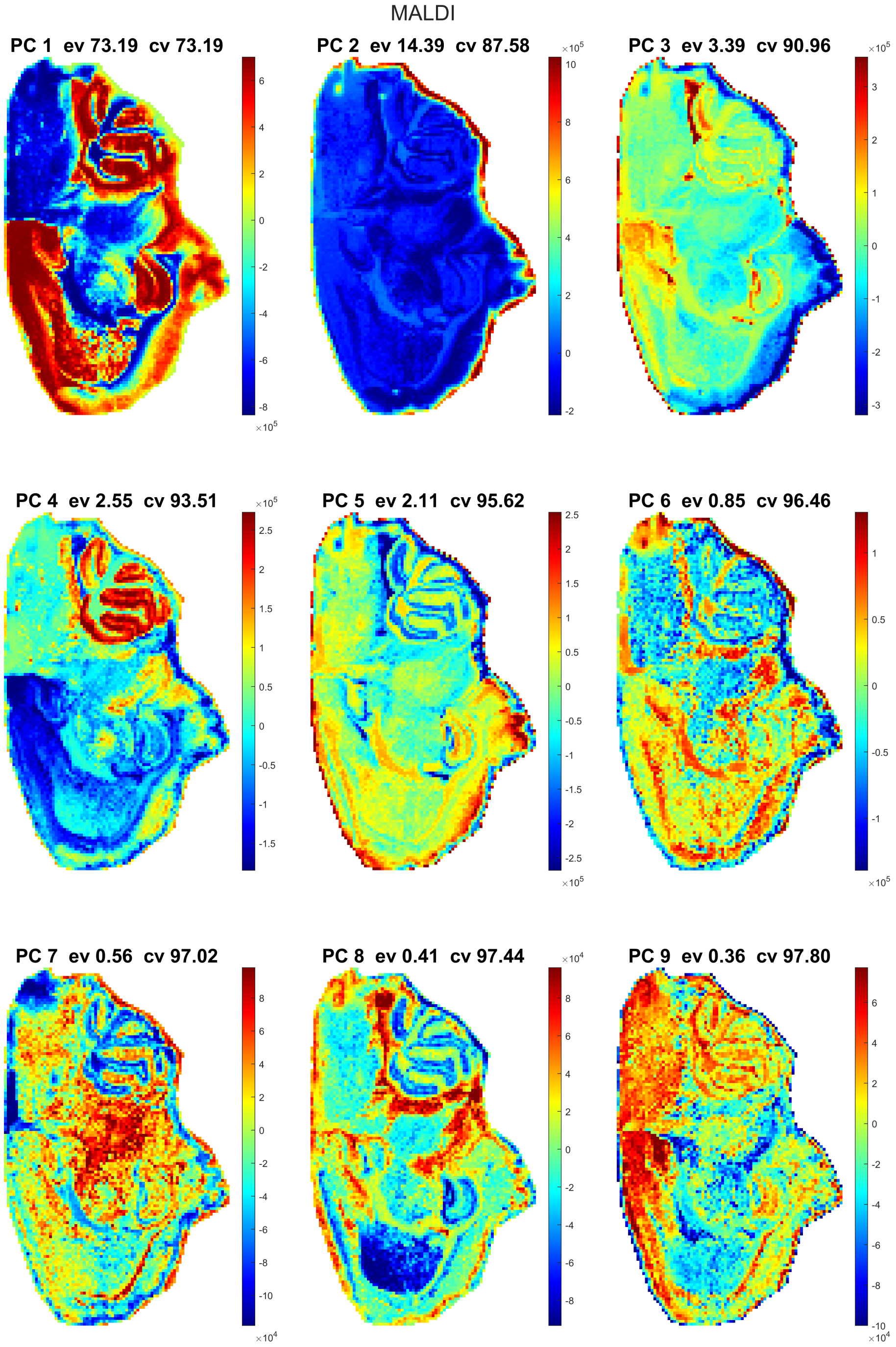


Figure 63: Score maps from PC-1 to PC-9. PCA applied to the MALDI data.

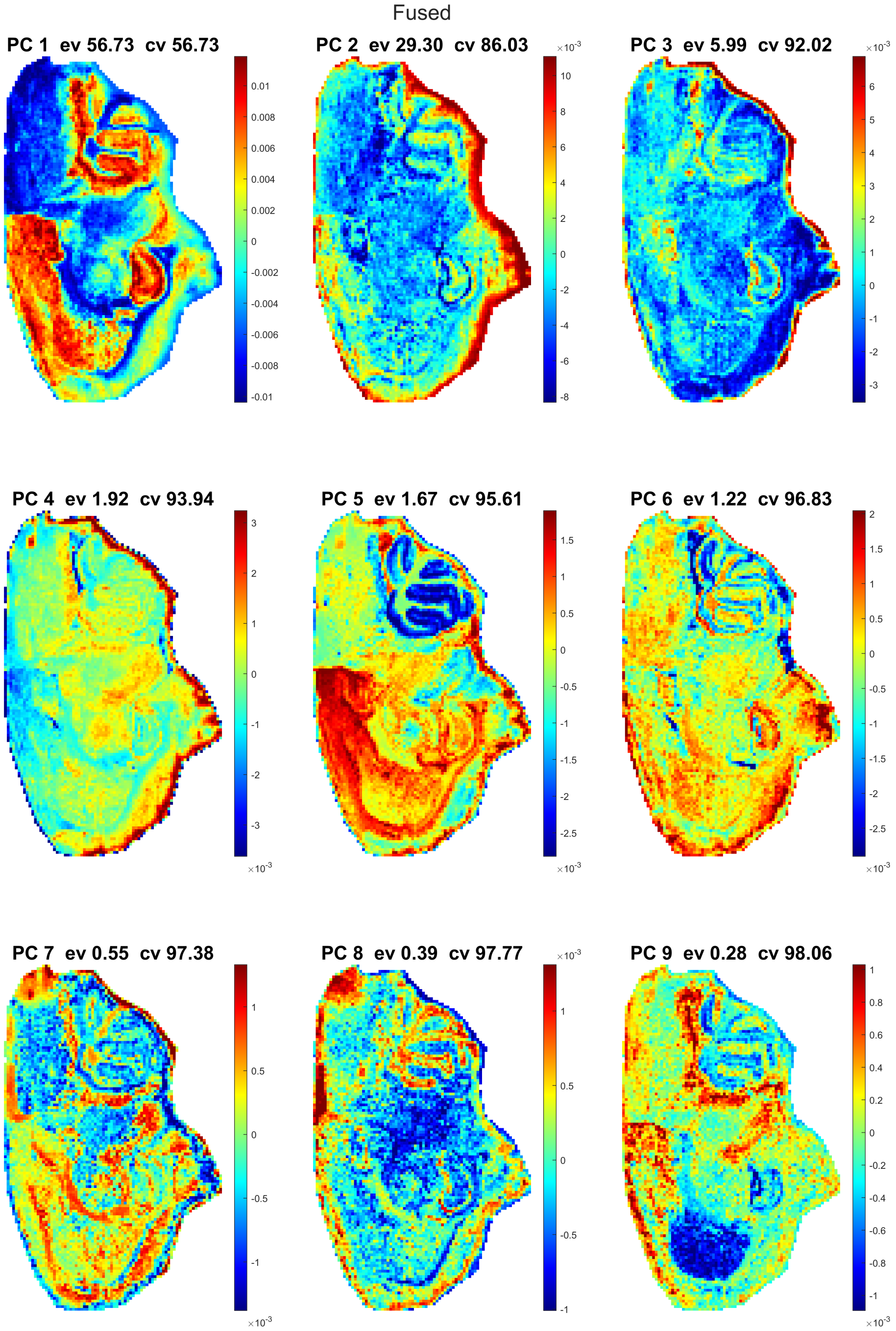


Figure 64: Score maps from PC-1 to PC-9. PCA applied to the fused data.

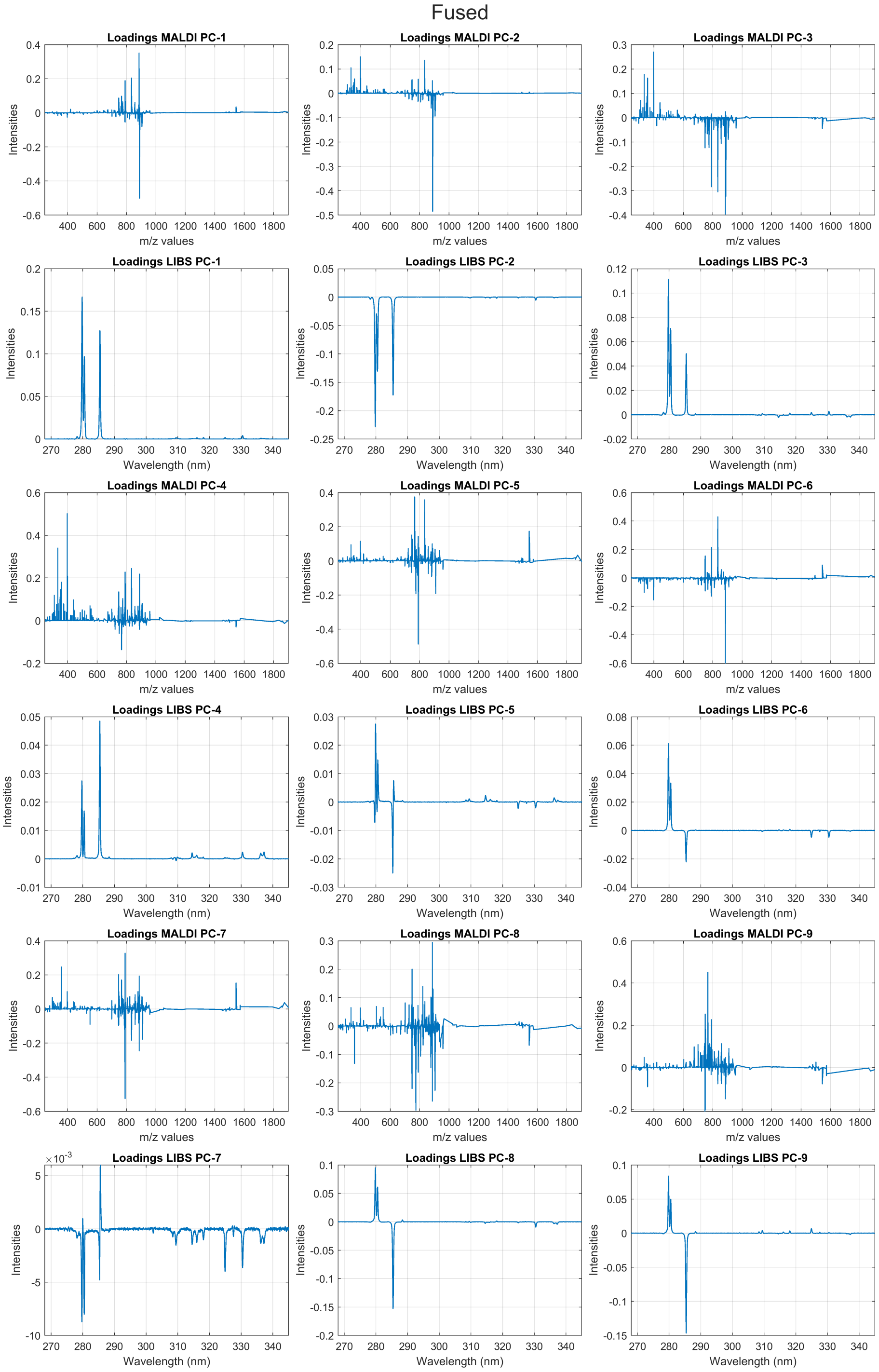


Figure 65: Loadings from PC-1 to PC-9. PCA applied to the fused data.

Interpreting this case can be difficult. However, the results can be improved. Taking into account the effect of the tissue edge, another mask can be generated using PC-2, for example, to exclude the edge pixels, and another PCA can be applied. However, when PCA is applied to only the LIBS data, the disproportion of Mg and the other elements is still noticeable. For this reason, it was decided to pursue another approach.

Correlation and Structural Similarity

The results of the PCA were not straightforward to interpret. As the purpose of this work was to understand the relationship between elemental and molecular information, another approach was used. As the main difficulties lay in the LIBS data, it was necessary to simplify them. Unlike MALDI, where there is relevant specificity, not all peaks in LIBS data are specific. Therefore, only a subset of the peaks can be used for species identification, since overlapping can make some of them non-specific. Then, the most interesting elements can be selected using the more specific peaks. The new variable can be labeled as the element that it represents. This variable corresponds to the element's chemical map vectorized. In this case, magnesium, calcium, and copper were selected as the most interesting elements. This choice can be justified because the other elements that can be extracted from the LIBS data do not highlight interesting structures of the brain. Once the variables representing the elements are extracted, they need to be compared with the MALDI data. Two different approaches were used: (1) calculating the correlation between the element's vector and each m/z channel, and selecting the top n most correlated channels; and (2) comparing the chemical map of the element with each m/z image based on the *structural similarity* (SSIM) index, and selecting the top n channels with higher SSIM. The idea was to obtain a list of the most related m/z channels for each element using different approaches, and to assess which method appeared more robust, subsequently evaluating whether the observed correlations were meaningful from biological point of view. Figure 66, figure 67, and figure 68 show the chemical map and the most correlated m/z channel with Mg, Ca, and Cu respectively.

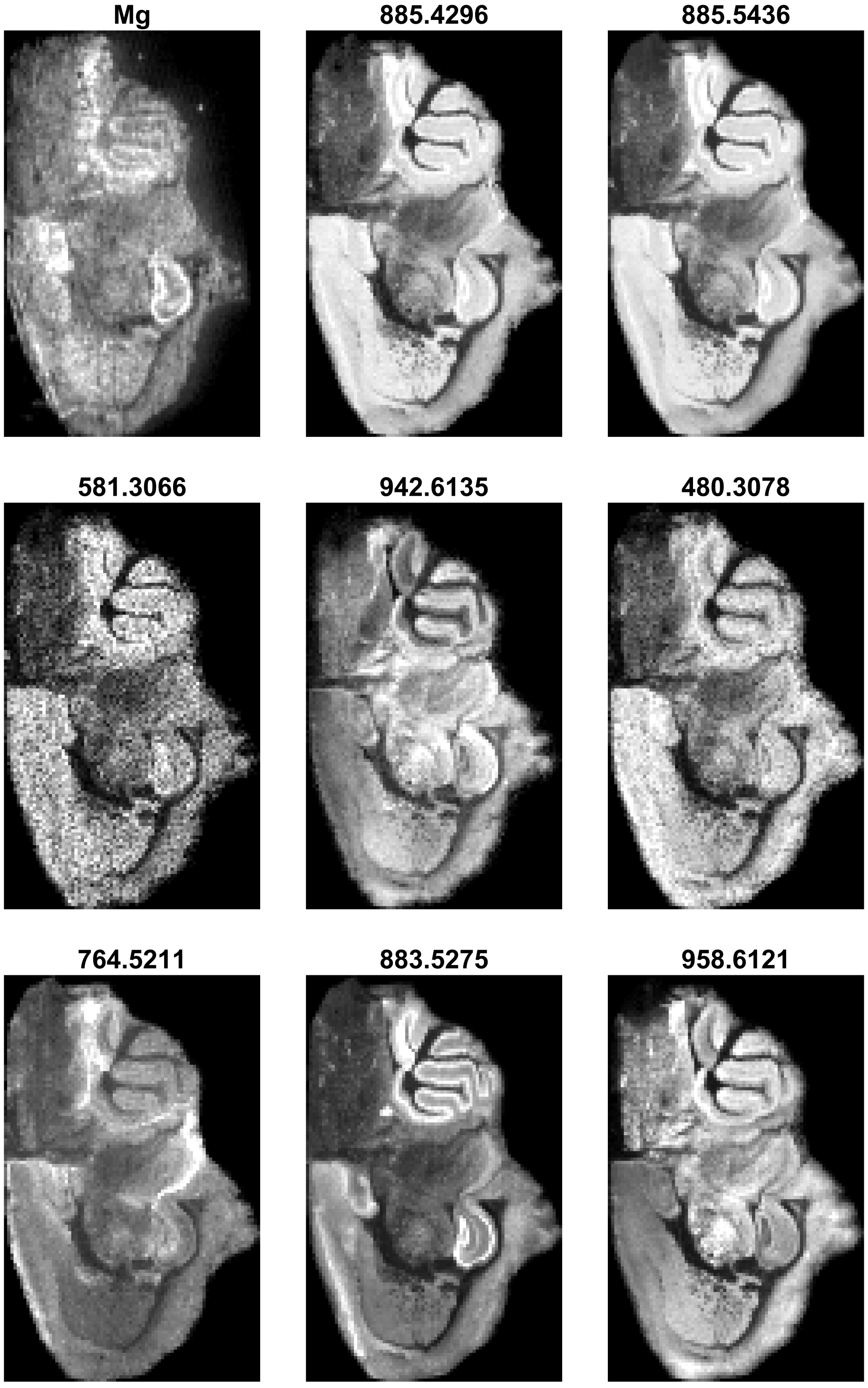


Figure 66: Mg chemical map and the most m/z correlated channel map

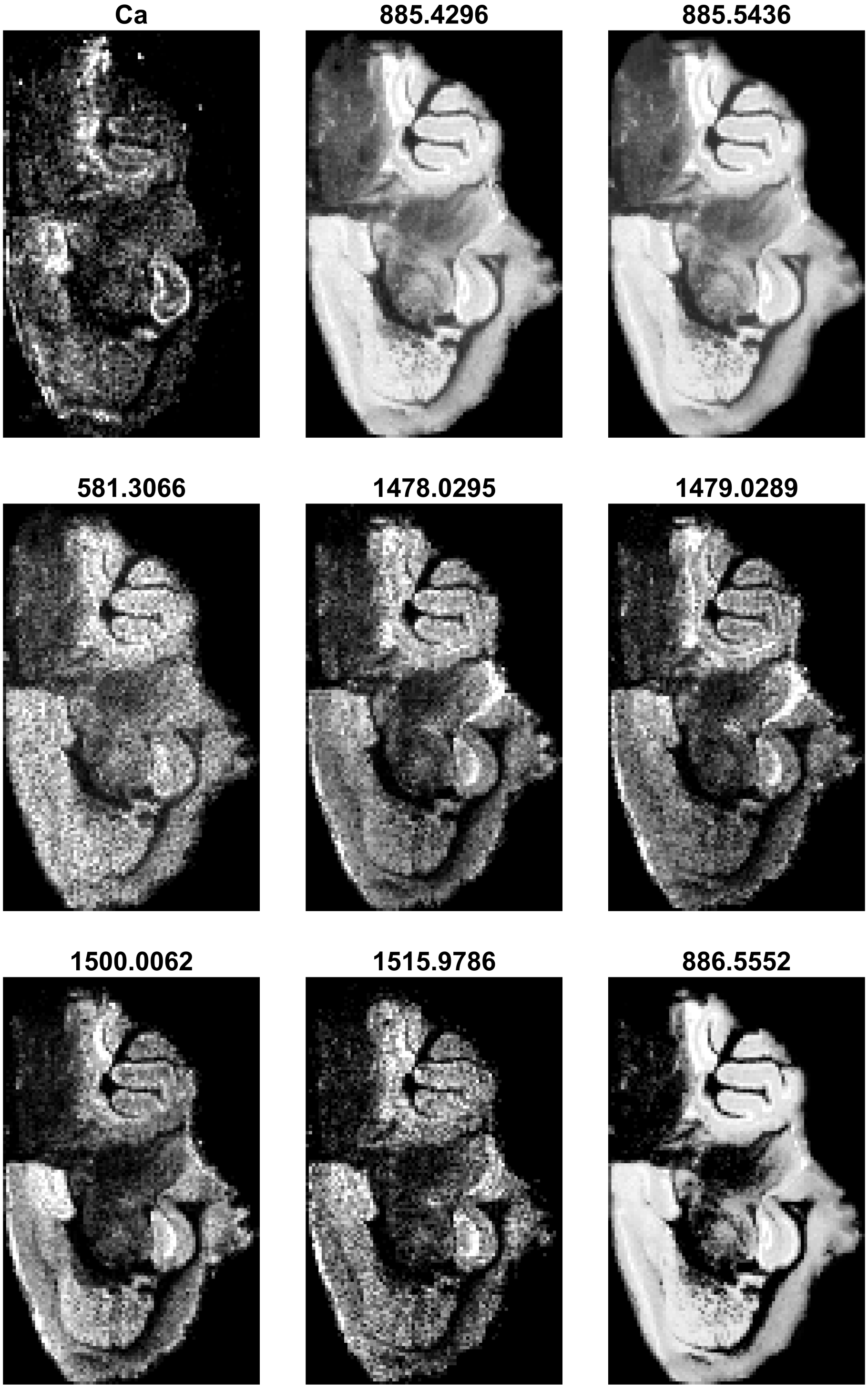


Figure 67: Ca chemical map and the most m/z correlated channel map

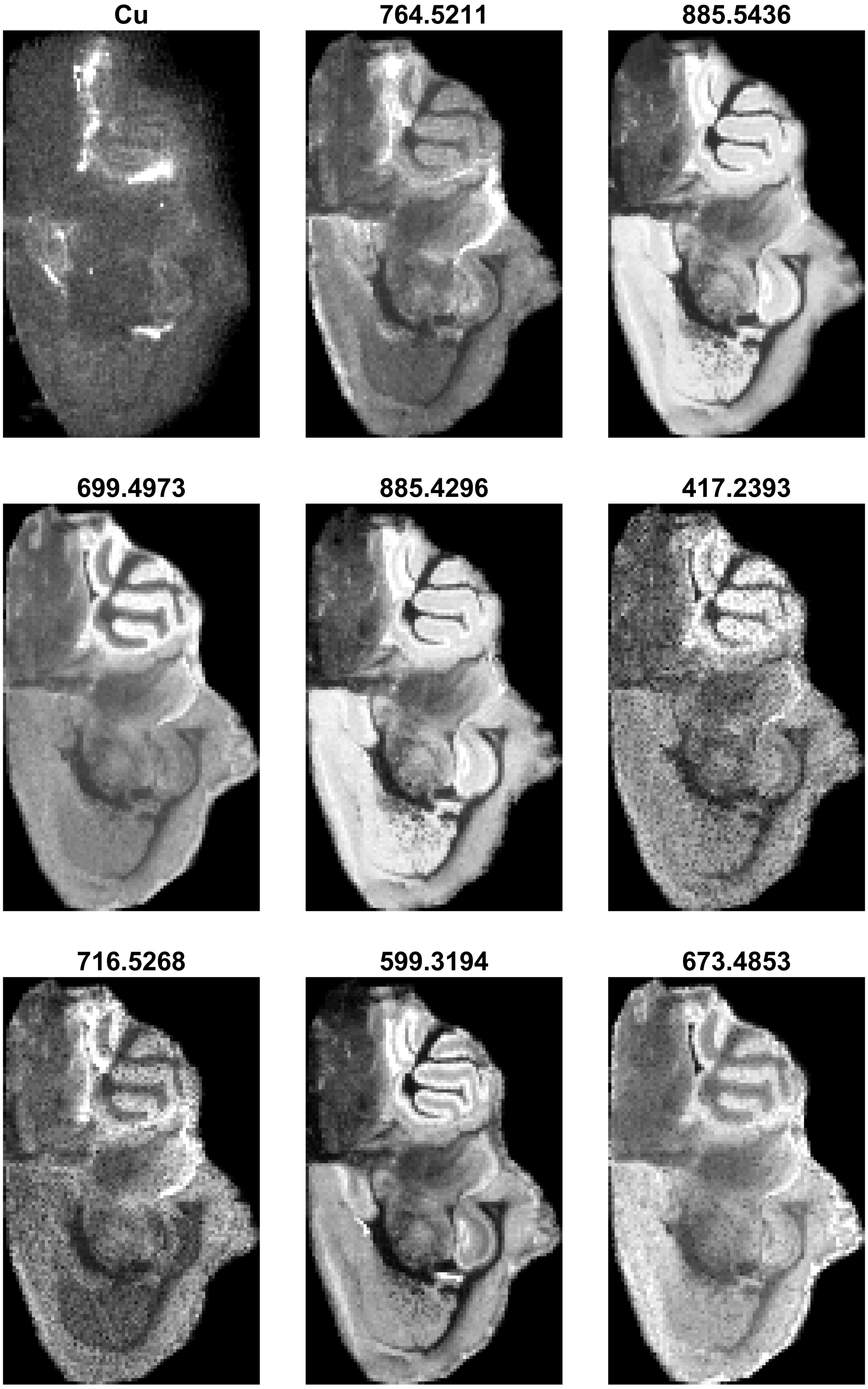


Figure 68: Cu chemical map and the most m/z correlated channel map

The most interesting results among the previous figures are related to those obtained with copper. As shown in its chemical map, copper is strongly concentrated at the edges of the cerebellum and near the hippocampus. However, except for m/z 764.5211, which showed high pixel intensities at the edge of the cerebellum, the other regions seem not to be so close to the chemical map of copper. This may be due to the dimensions of the vector that was compared. The image is 127 x 80 pixels, for a total of 10,160 values. Considering that copper covers less than 10% of the image, the rest of the image probably covers the relationship of interest. Additionally, as most of the pixels are located outside the region or tissue, this may have influenced the results. For this reason, rather than calculating the correlation across the entire image, a mask is generated for each element, identifying the pixels where the element is present. The correlation is then calculated only on these pixels. This should assure that we are looking for the relationship for the region where the element is present. The masks are shown in figure 69. Once the mask had been generated, pixels were extracted from both the LIBS and the MALDI matrices, and the correlation between the variables was calculated. Right after, the chemical map of Mg, Ca, and Cu, along with the map of their most correlated m/z channel, are shown, respectively, in figure 70, figure 71, and figure 72. Compared to the results obtained without the mask, those obtained using the mask seems to be more similar compared to the case without the mask. As it can be seen looking at the previous figures, particularly for copper, the m/z channel with the highest correlation obtained are not necessarily the most biologically significant ones, the ones related more to the respective elements. Some m/z as 687.5342 seems to be more related to copper than 716.5268, even if the latter one showed an higher correlation. For this reason, this approach can be viewed as a way to drastically reduce the number of m/z channels related to the elements. Further analysis and interpretation can be done among this small group. This does not necessarily require chemometric analysis, but rather dialogue with bio experts to determine whether these correlations are spurious or if there is a biological link. However, this was not possible due to a lack of time.

Correlation is not the only way to compare variables. Until now, references to the image meaning of the variables were always present. For this reason, another criterion based on image comparison was tested and the results were compared with those of the 1-D correlation approach.

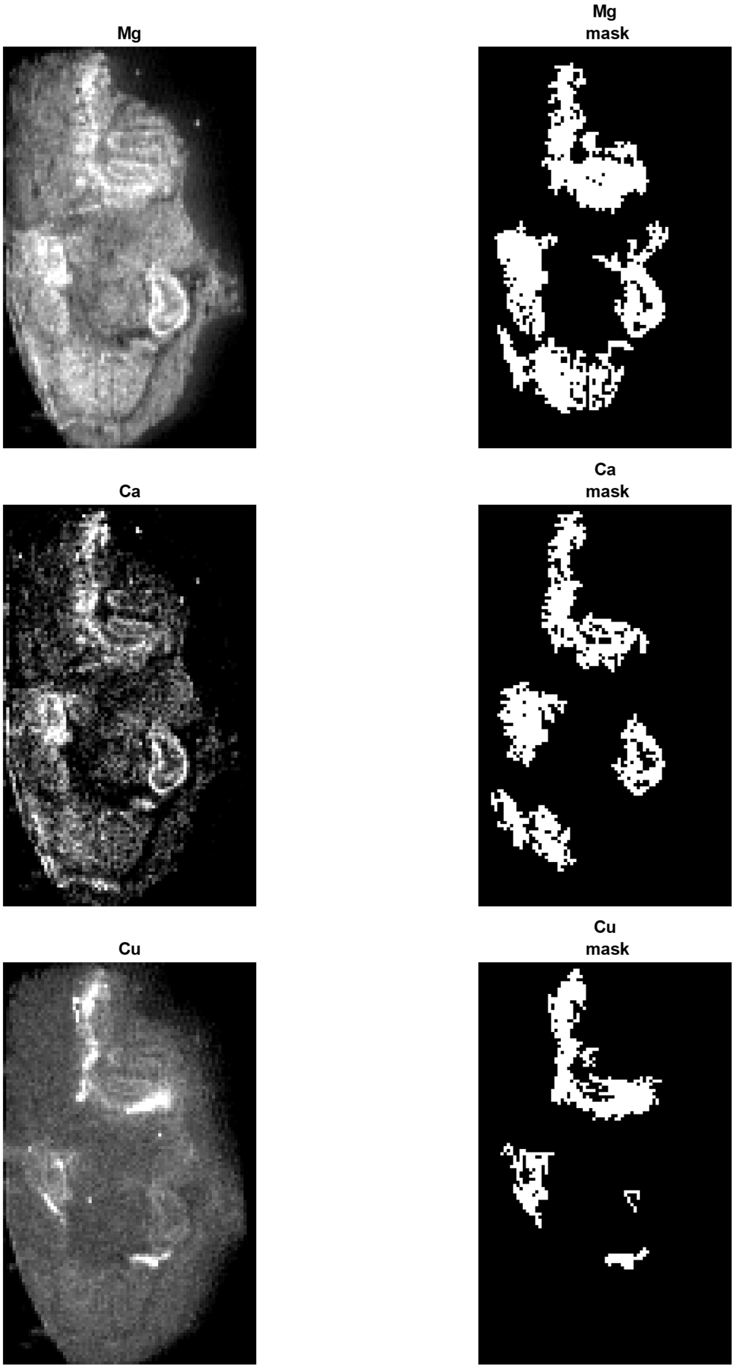


Figure 69: Chemical map of Mg, Ca, and Cu (left) and relative mask (right).

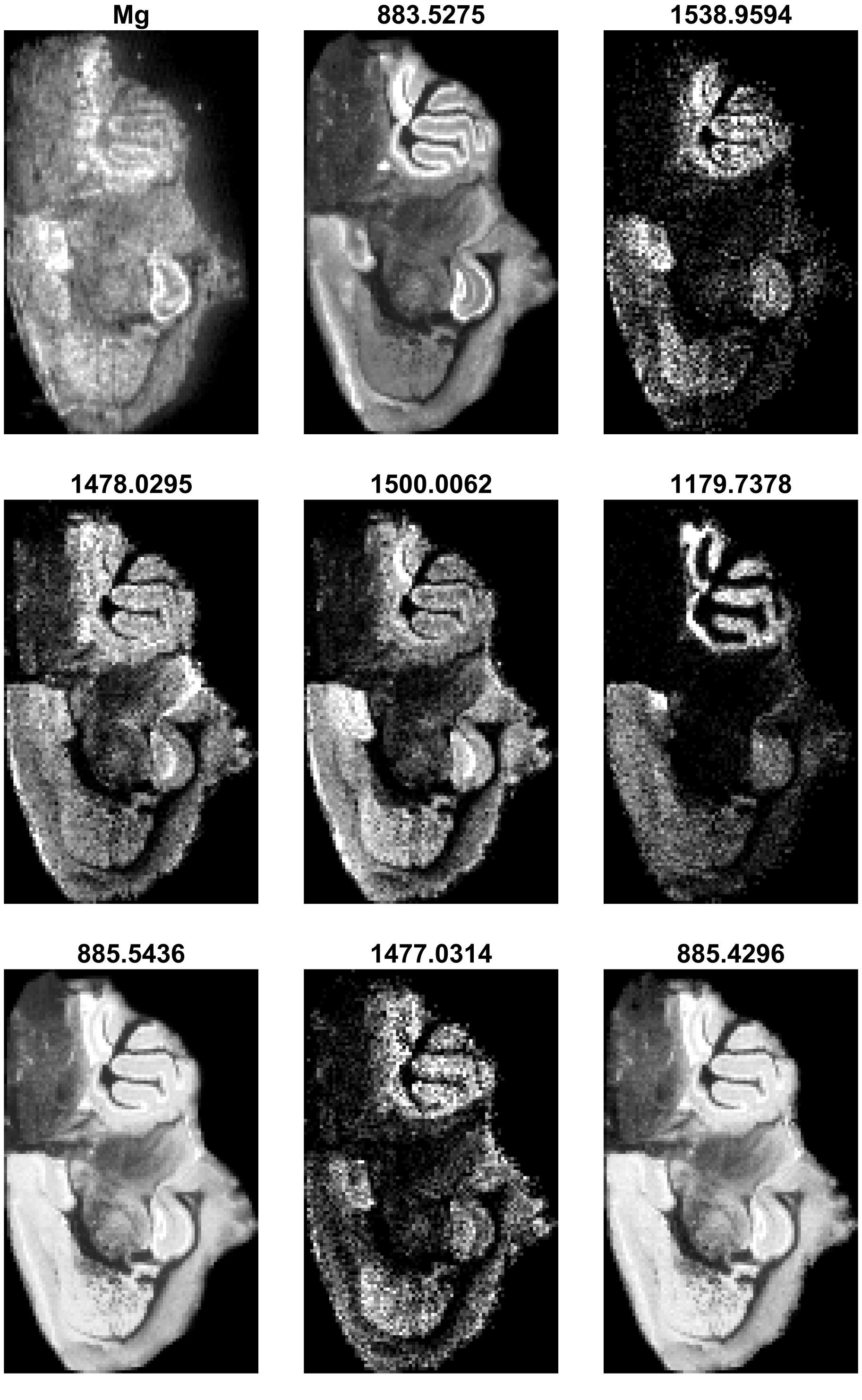


Figure 70: Mg chemical map and the most m/z correlated channel map

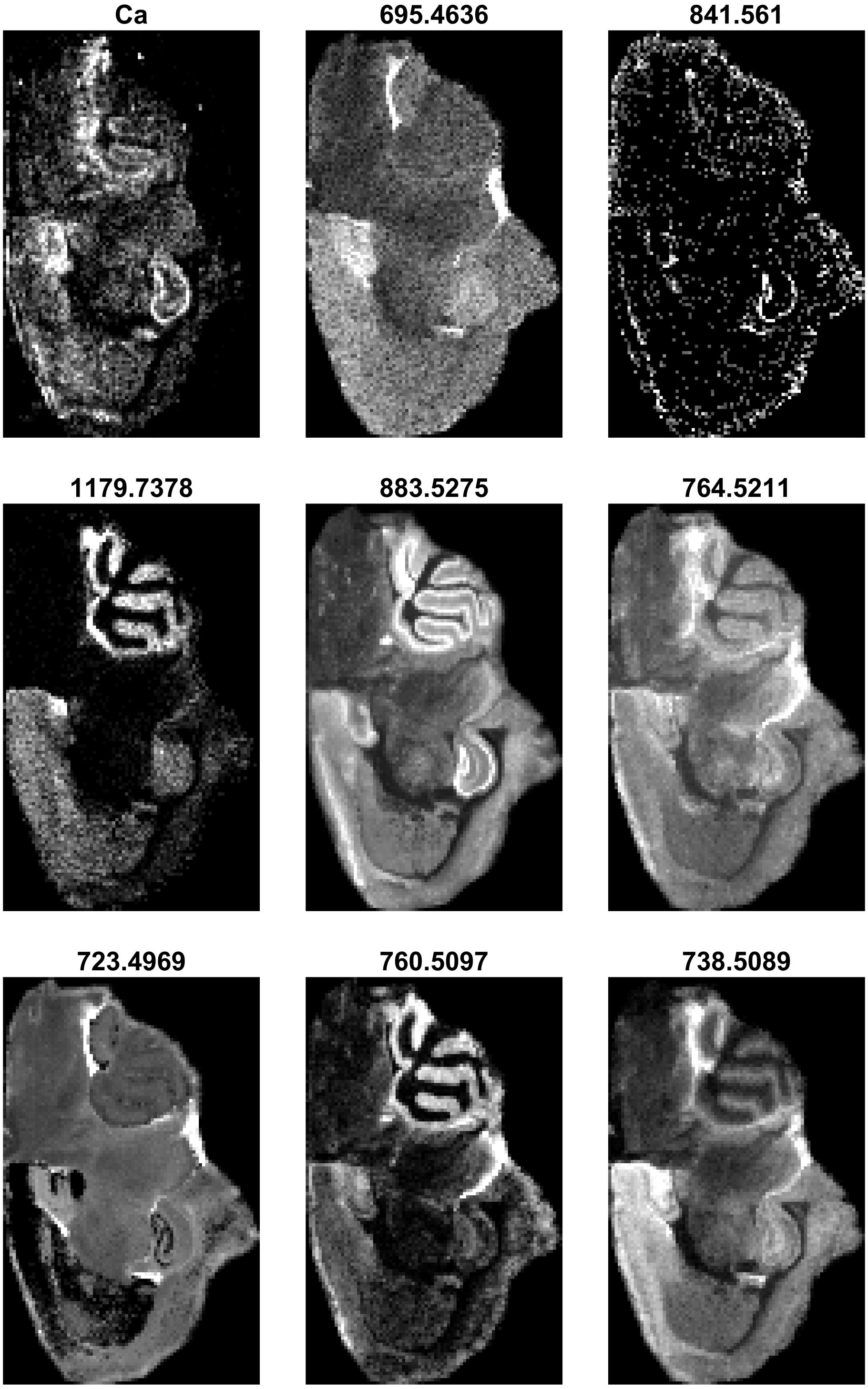


Figure 71: Ca chemical map and the most m/z correlated channel map

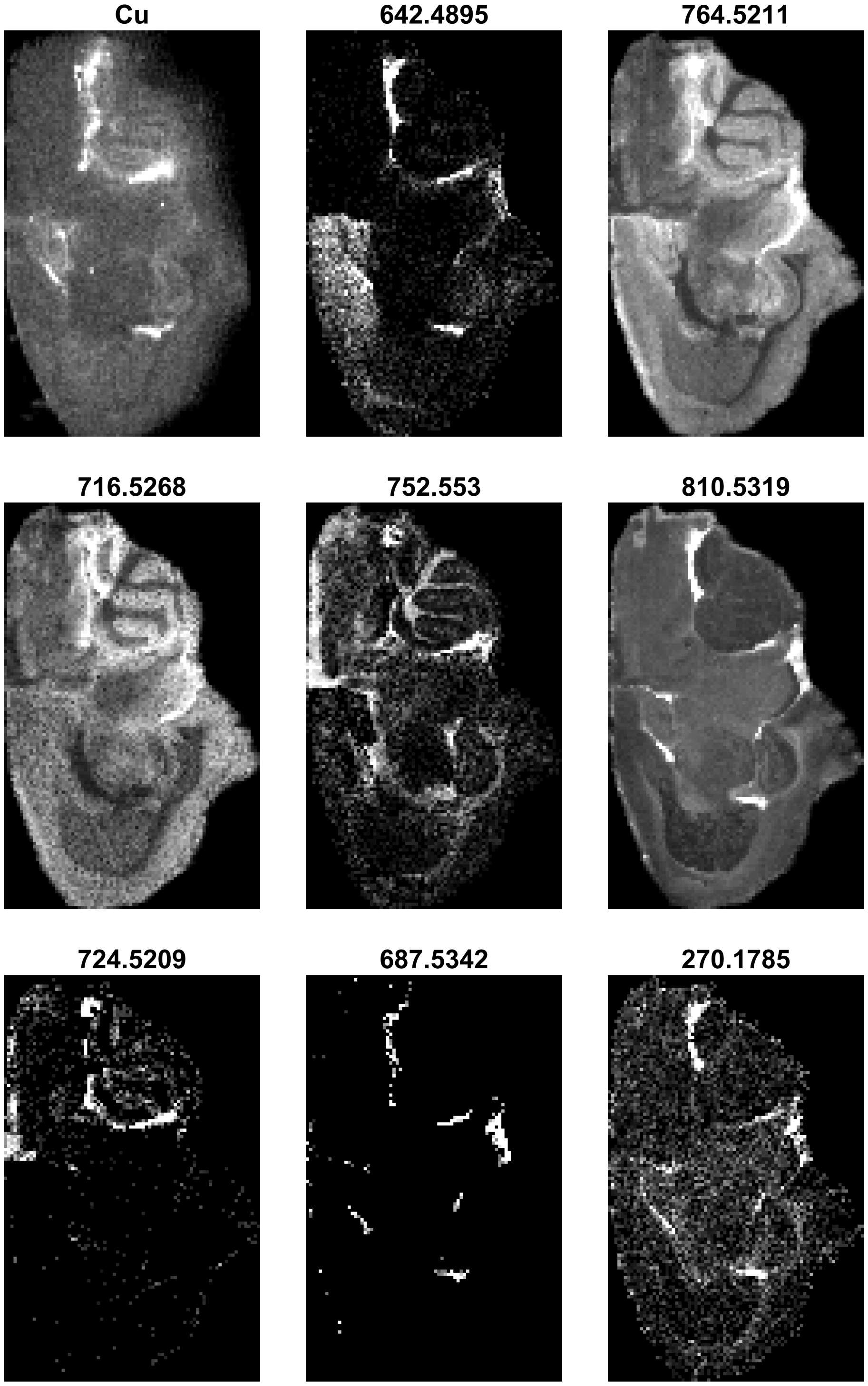


Figure 72: Cu chemical map and the most m/z correlated channel map

Structural Similarity

Structural similarity (SSIM) index for measuring image quality [91] measures the structural similarity between two images to assess how similar they are from a perceptual point of view. Given two image signals, im_a and im_b, the SSIM index is built from three components: (1) luminance comparison, (2) contrast comparison and (3) structure comparison. The general SSIM formula is:

$$\mathrm{SSIM}(im_a, im_b) = [l(im_a, im_b)]^{\alpha} \cdot [c(im_a, im_b)]^{\beta} \cdot [s(im_a, im_b)]^{\gamma} \tag{23}$$

Typically, the exponents are set to $\alpha = \beta = \gamma = 1$, so:

$$\mathrm{SSIM}(im_a, im_b) = l(im_a, im_b) \cdot c(im_a, im_b) \cdot s(im_a, im_b) \tag{24}$$

Luminance comparison is based on the means μ_{im_a} and μ_{im_b}:

$$l(im_a, im_b) = \frac{2\mu_{im_a}\mu_{im_b} + C_1}{\mu_{im_a}^2 + \mu_{im_b}^2 + C_1} \tag{25}$$

Contrast comparison is based on the standard deviations σ_{im_a} and σ_{im_b}:

$$c(im_a, im_b) = \frac{2\sigma_{im_a}\sigma_{im_b} + C_2}{\sigma_{im_a}^2 + \sigma_{im_b}^2 + C_2} \tag{26}$$

Structure comparison is based on the covariance $\sigma_{im_a im_b}$:

$$s(im_a, im_b) = \frac{\sigma_{im_a im_b} + C_3}{\sigma_{im_a}\sigma_{im_b} + C_3} \tag{27}$$

Where $C_3 = \frac{C_2}{2}$. Combining the above, assuming $C_3 = \frac{C_2}{2}$, we get:

$$\mathrm{SSIM}(im_a, im_b) = \frac{(2\mu_{im_a}\mu_{im_b} + C_1)(2\sigma_{im_a im_b} + C_2)}{(\mu_{im_a}^2 + \mu_{im_b}^2 + C_1)(\sigma_{im_a}^2 + \sigma_{im_b}^2 + C_2)} \tag{28}$$

The typical Parameters are:(A) $C_1 = (K_1 \cdot L)^2$; (B) $C_2 = (K_2 \cdot L)^2$; (C) L is the dynamic range of pixel values (255 for 8-bit grayscale images) (D) $K_1 = 0.01$, $K_2 = 0.03$. It is computed locally within a small, moving window with Gaussian weighting, and then the mean is calculated to obtain a global value. For each chemical map (Mg, Ca, and Cu) SSIM is calculate for all the m/z channel map.

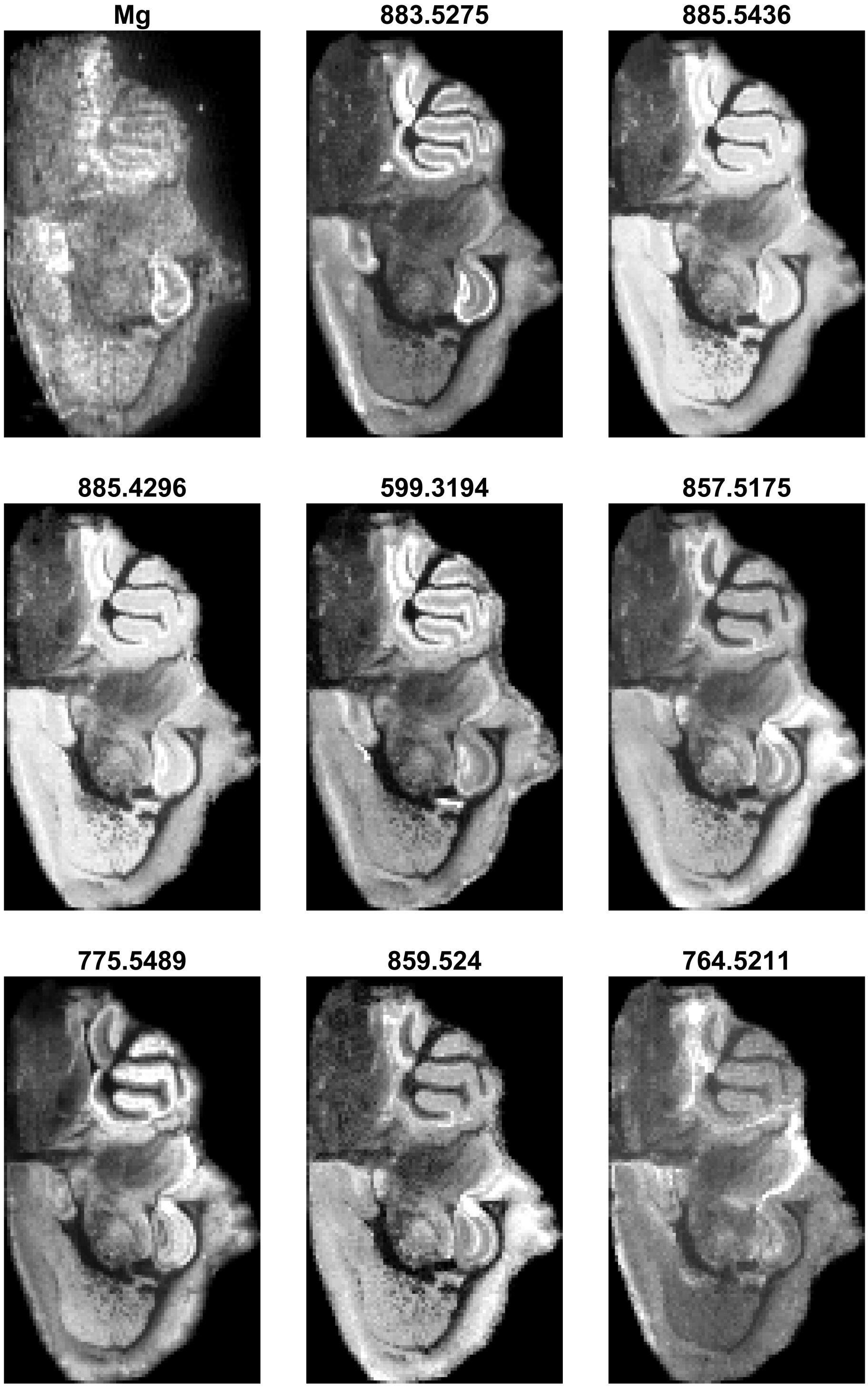


Figure 73: Mg chemical map and the m/z channel map with the highest SSIM

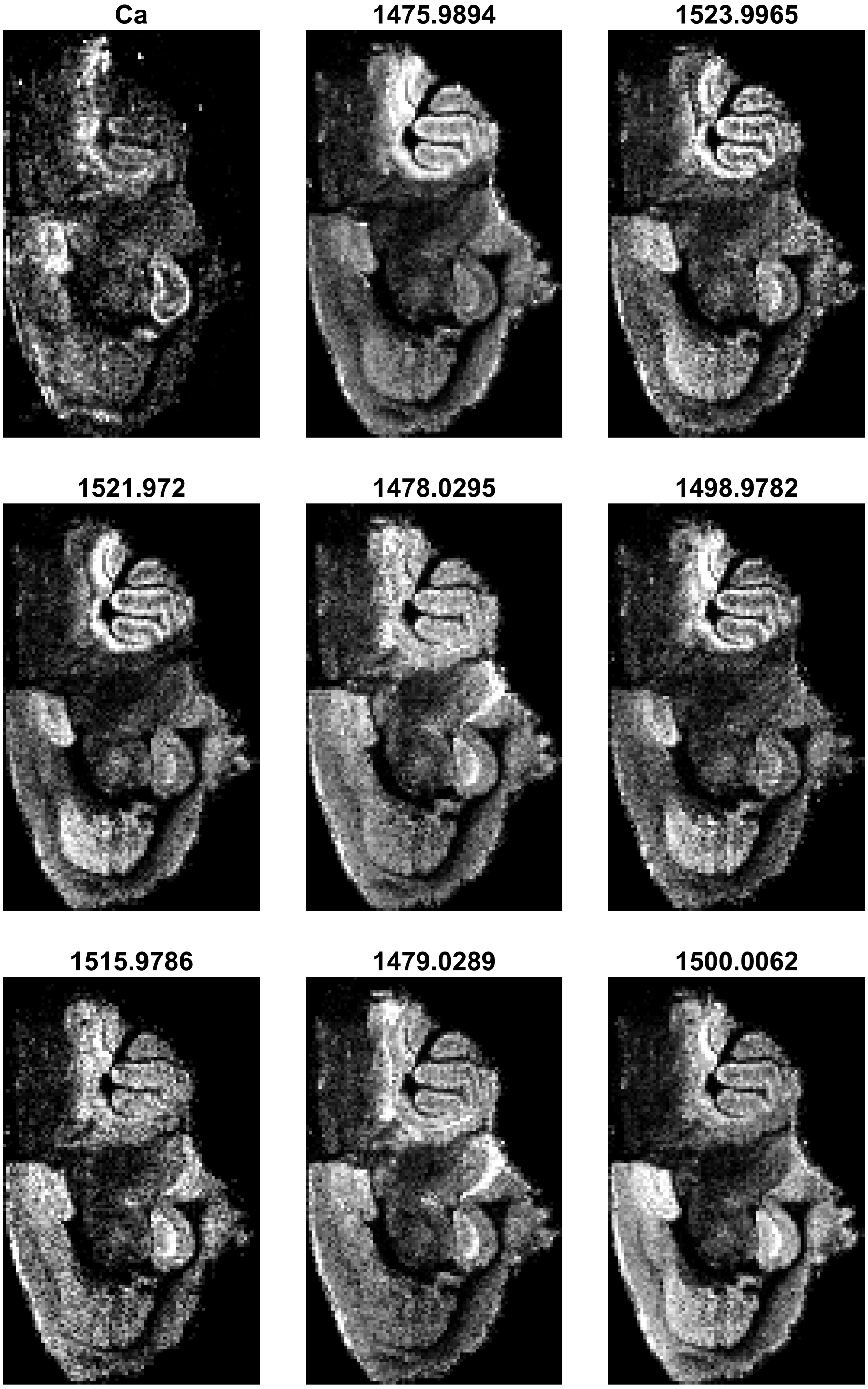


Figure 74: Ca chemical map and the m/z channel map with the highest SSIM

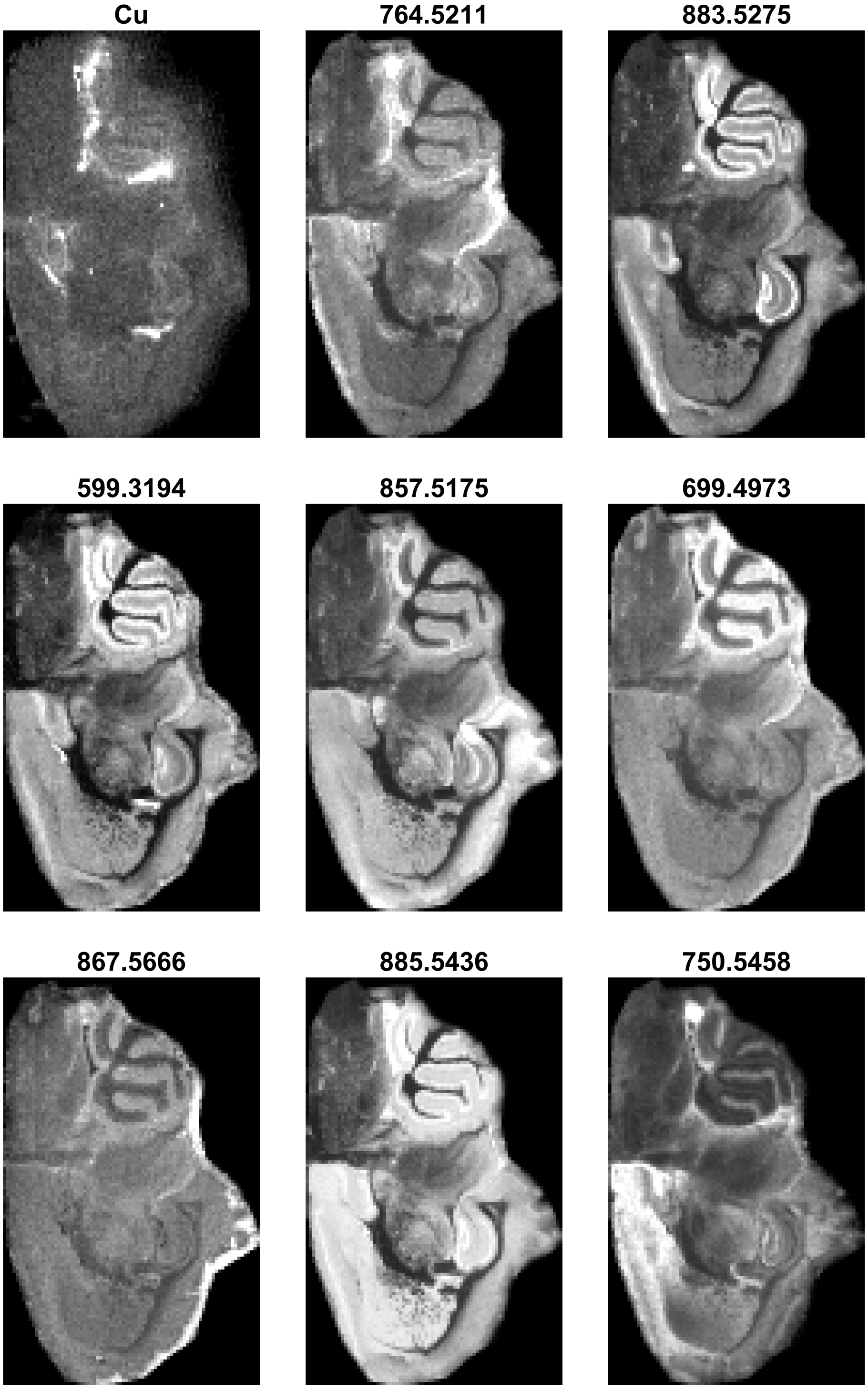


Figure 75: Cu chemical map and the m/z channel map with the highest SSIM

Figure 73, 74, 75 show the chemical map with the m/z channel map with the highest SSIM. The previous m/z channel map with the highest SSIM does not appear to be particularly similar to the respective elemental chemical map, especially in the case of Cu. This may be due to the significant influence of the external part of the image (outside the tissue). For this reason, rather than calculating a global SSIM for the image, SSIM was computed using the tissue mask obtained from the NaN values of the MALDI cube. However, even in this case, the m/z channel maps were not very similar in the case of Cu. Since Cu is a minor element (the mask used covered less than 8%), the value may be driven more by other parts of the tissue, with the region where Cu is present contributing less and not being sufficient to highlight the correct m/z channel. For this reason, SSIM was also calculated using the previously generated mask (see Figure 69). As expected, in this case the criteria should be more representative of the tissue region where the element is present.

Later, Figure 79 will show the comparison of the criteria used (correlation and SSIM) across three cases: the image without a mask, using a tissue mask, and using an element-specific mask. As done previously, the first 9 m/z channels with the highest SSIM will be shown.

Looking at the results, the m/z channel map appears more similar when the mask is applied, compared to the case without the mask. For example, m/z channel 738.51 is presented as one of those with the highest SSIM with Cu. Here, the region close to the hippocampus shows a higher concentration of this m/z. The m/z channels with the highest SSIM (or correlation) are not necessarily the most biologically significant ones, i.e., those more strongly related to the respective elements. These two approaches were adopted to explore the relationship between elemental and molecular imaging. Further studies focusing on the relationship between elements and related molecules need to be pursued.

Previously, it was hypothesized that the results obtained without the mask were driven by the region outside the tissue. To demonstrate this, the values of the criteria without the mask, with the tissue mask, and with the element mask are shown in Figure 79. Except for a few m/z, the criteria calculated using the masks are lower than those obtained without the mask.

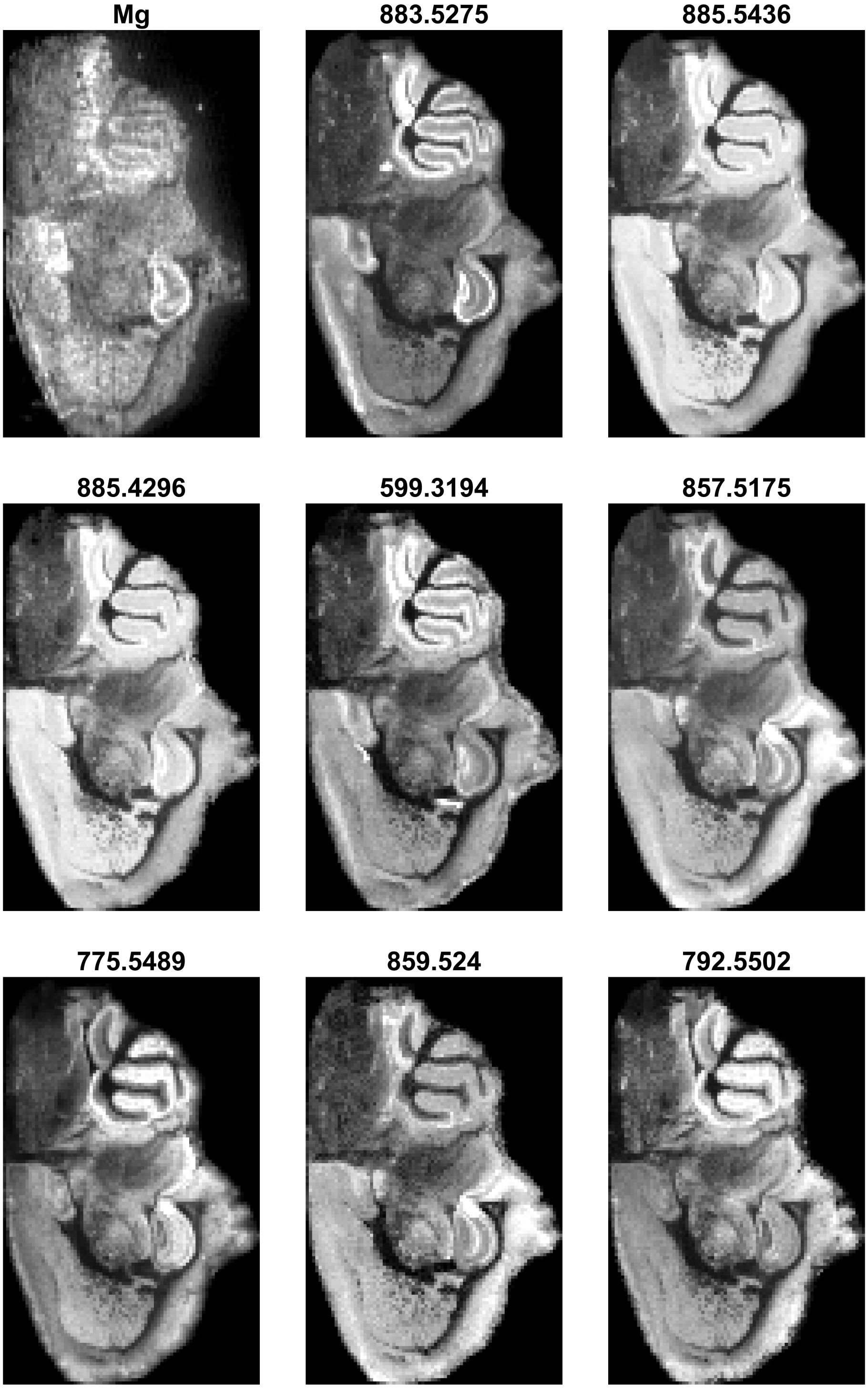


Figure 76: Mg chemical map and the m/z channel map with the highest SSIM

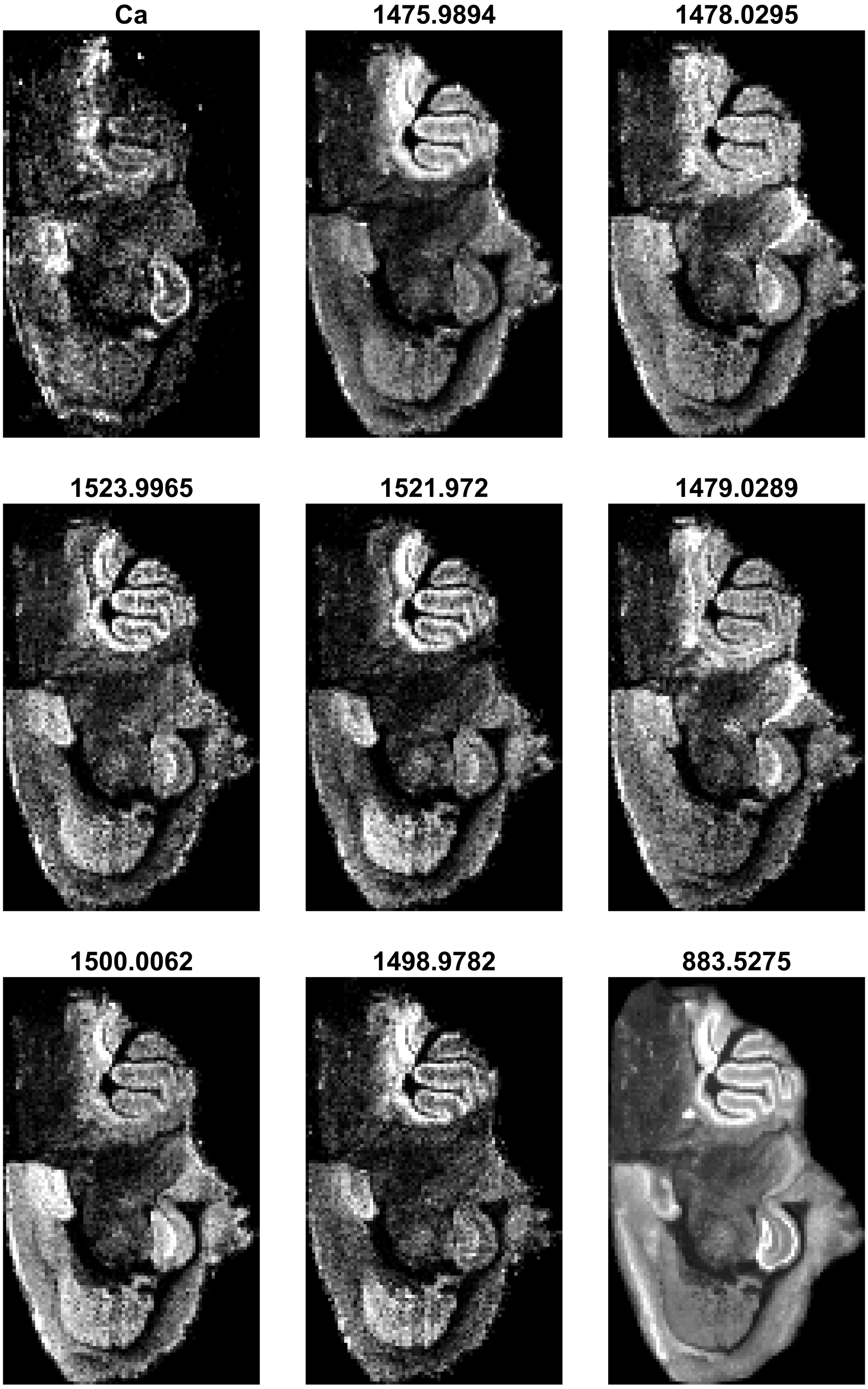


Figure 77: Ca chemical map and the m/z channel map with the highest SSIM

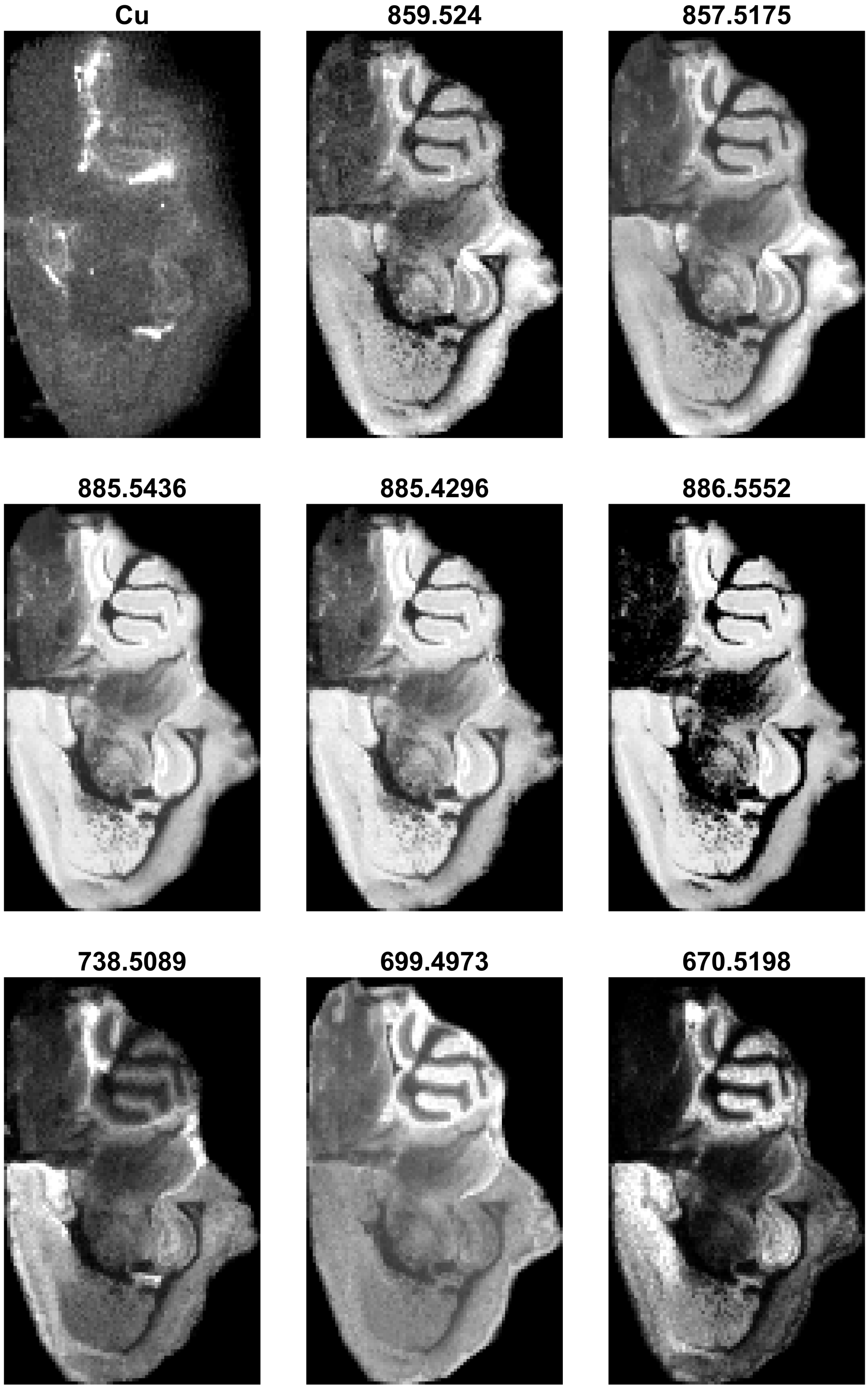


Figure 78: Cu chemical map and the m/z channel map with the highest SSIM

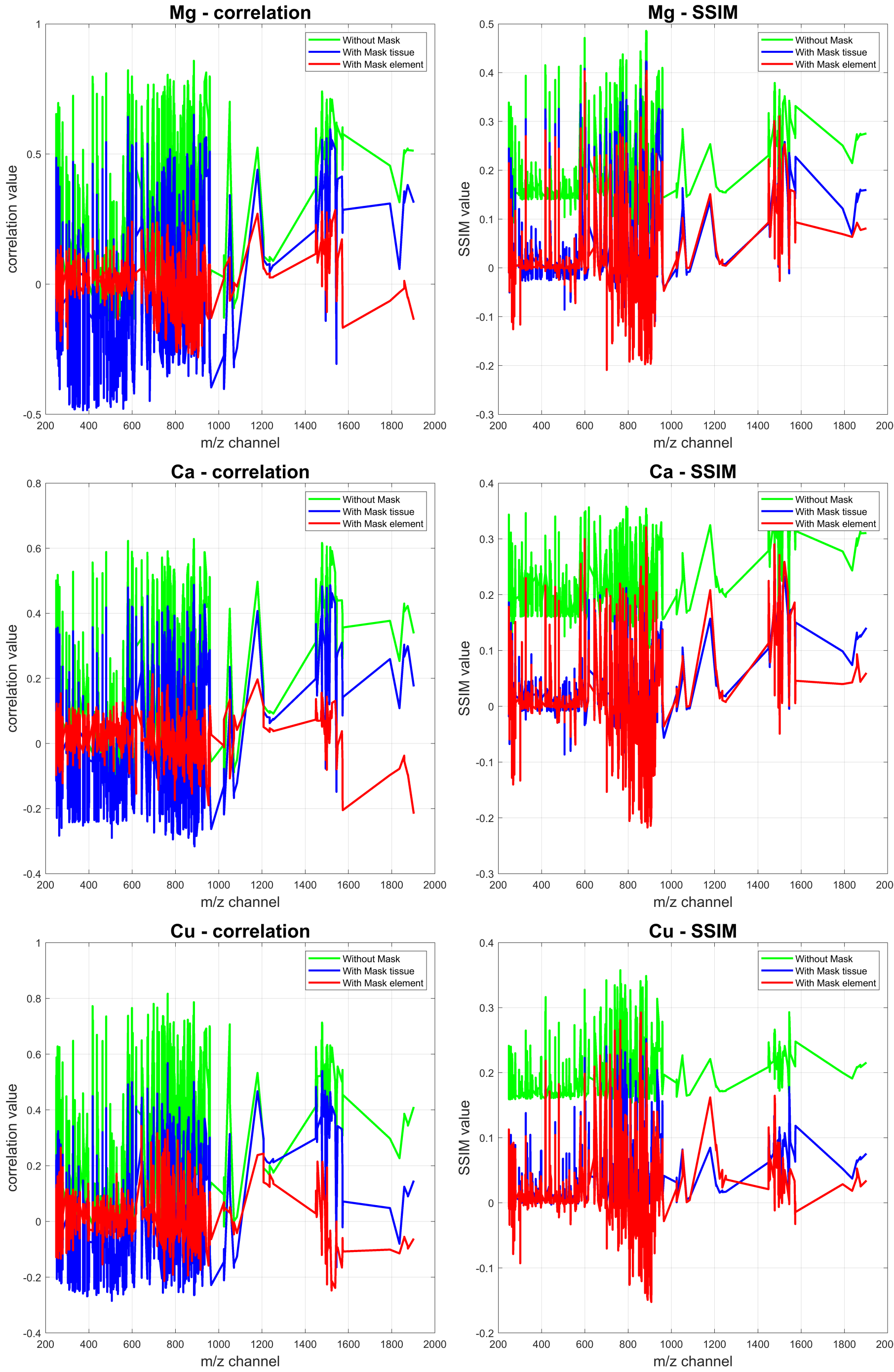


Figure 79: Plots of the criteria without the mask (green), with a mask of the tissue (blue), and with the maks of the element (red). It is possible to see that using the mask the value decrease.

3.2.2 Conclusions and future perspectives of chapter 3

This work was a feasibility study aimed at proving that the fusion of elemental and molecular imaging, in this case, MALDI and LIBS, can improve the interpretation of biological samples by allowing possible correlations between elements to be appreciated. As both techniques are destructive, acquiring the same section is not possible. Due to this, two consecutive sections of the tissue were analyzed, with the morphological differences considered small and mitigated by the spatial binning performed during preprocessing. The initial study involved applying PCA to the two blocks of data individually and then to the fused data. However, the results obtained from PCA were difficult to interpret. The MALDI block almost covered all the explained variance, even with normalization, and it was difficult to see patterns within the LIBS data. Due to this, a different approach was applied to study the relationship between molecular and elemental information. In short, a few elements were extracted by calculating the chemical map, obtaining a vector for each element. The idea was to calculate the correlation between the elements and the different m/z maps. However, for minor elements such as copper, the correlation value was mostly driven by pixels where copper was not present. This resulted in m/z values that did not seem to follow a similar spatial distribution. To avoid this, the correlation was calculated only on pixels showing presence of the element, e.g. Cu. This made the results more representative and produced an m/z channel map that was more similar to the elemental one. Another metric was also used to calculate the relationship between the chemical map (LIBS) and the m/z channel map: SSIM. This parameter is used in image processing to calculate the similarity between images. The idea was to analyze and compare the results obtained by calculating the correlation on a vector with this parameter, which is calculated directly on the image. However, even in this case, the initial results were not ideal as they seemed to be mostly driven by pixels where the considered element was not present. As the algorithm uses a moving filter to calculate a value for each pixel and then considers the mean of these values as the global similarity of the image, an intermediate step was added. Instead of calculating the mean across the entire image, the mean was calculated using a mask based on the element in question. As shown, this improved the results when looking at the values, but when looking at the images using correlation, the m/z channel maps were closer to the elemental one.

Although this study did not reveal any significant innovations in terms of the results obtained, the potential for combining MALDI and LIBS techniques has been demonstrated by studying a rat brain. Further studies using more interesting and informative samples are worth considering. In future work, it will be particularly valuable to apply this fusion framework to biomedical questions in which aberrant elemental homeostasis is believed to have a causative or diagnostic role. Neurodegenerative disorders such as Alzheimer's [92] and Parkinson's [93, 94] disease, for instance, exhibit spatially heterogeneous accumulations of metals like Fe, Cu and Zn. By correlating LIBS-derived elemental maps with MALDI-MS molecular signatures in the affected tissue, it should be possible to determine whether distinct lipid, peptide or metabolite patterns co-localise with metal hotspots, offering fresh insight into disease mechanisms and potential biomarkers. A similar strategy could be extended to cancers, where trace-element dysregulation is implicated in tumour progression and therapy resistance, due to the role of certains elements in cancer progression proven [95] and the ability to see those elements by LIBS in this context [96]. Accordingly, the approach demonstrated here could evolve into a powerful platform for resolving outstanding questions on element–molecular relationships in a range of pathological conditions.

Chapter 4

New Challenges in Processing Data from Kilohertz LIBS Imaging

Noise is one of the biggest enemy of a chemical analysis. These unwanted fluctuation of the signal are not related to chemical (or physical) properties of interest, and can cover interesting signals, usually traces as they can have a similar intensities.

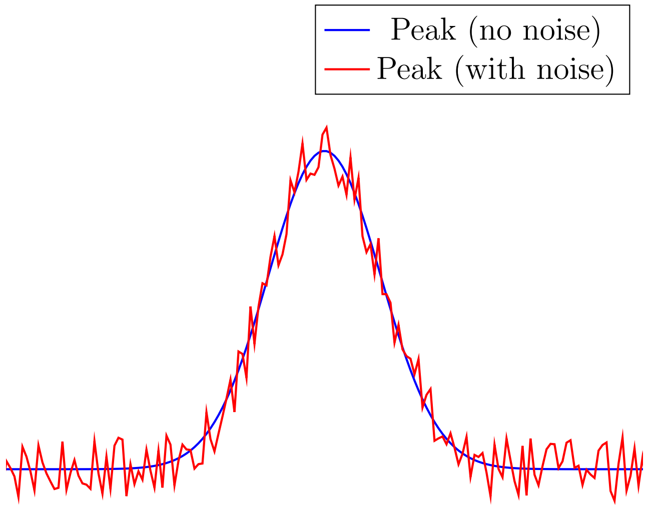


Figure 80: Example of peak with a gaussian shape affected by noise

There are different type of noise, depending on the instrumentation, the spectroscopy used, etc. In analytical chemistry, noise has effect directly on Limit Of Detection (LOD) and Limit Of Quantification (LOQ), as their definition is based on noise intensity. Usually is defined Signal-To-Noise-Ratio (SNR) as

$$\text{SNR} = \frac{\text{Amplitude of signal}}{\text{Mean amplitude of noise}} \tag{29}$$

There are different way to reduce the noise, and increase SNR, including instrumentation improvement, however they are out of the aim of this work.

Currently, most of the μLIBS-imaging setups have lasers with a shooting rage lower than 100 Hz [97]. However, kHz lasers could be a significant breakthrough for elemental imaging analysis. Nowadays, the use of kHz laser on LIBS is not widespread, although literature presents several examples mainly focused on industrial [98] or geological [99] application. Using such sampling frequencies, kHz range, offers a significant benefit in terms of analysis time reduction, enabling mapping 1 million spectra - equivalent to 1 cm^2 with a lateral resolution (e.g., shot-to-shot distance) of 10 μm - in roughly 17 min. Building on these advancements, μLIBS-Imaging at kHz rate is attracting the interest of new fields of application. In addition, it requires compact and transportable equipment, with no need for consumables, and reduced economic cost and environmental impacts, making it an interesting candidate to be implemented in fields combining research and routine requirements, such as biomedicine, where specimens can be incredibly diverse and complex, composed of multiple tissues such as bone tissue, muscular or blood vessels [97]. Nonetheless, its implementation has diverse challenges. Firstly, as a direct consequence of the reduction in the analysis time, the measure of elemental distribution maps with a considerable quantity of pixels (1 pixel in the elemental map corresponds to 1 spectrum) becomes easily achievable, resulting in an increase in the complexity of the data treatment process since the number of spectra would be increased in, at least, one order of magnitude. To address this challenge, new and more efficient tools are required; previous work has proposed different solutions, such as the use of novel artificial intelligence techniques, including Facebook libraries for clustering [18], or new mask-creation operations via logical relationships for mineral phase identification [99]. Moreover, to avoid compromising the lateral resolution of the analysis, it is crucial to minimize the induced thermal or stress damage due to laser-matter interaction, especially for the biological specimen, as they are typically more fragile than mineral or metal samples. Thus, low-energy laser pulses (1 mJ or lower) are recommended. Unfortunately, this approach results in a weaker plasma emission.

Additionally, detectors capable of achieving an analysis rate in the kHz range are necessary. For kHz LIBS, two-dimensional sensors, such as sCMOS, are typically employed. However, the effective area of the sensor (sensor pixels) must be reduced to achieve the desired acquisition rate, reducing the light amount recorded. The combination of these factors inevitably causes the worsening of the signal-to-noise ratio (SNR) of the acquired spectra and, therefore, a decrease in the elemental image quality generated from them. The noise could strongly impact the quality and quantitative analysis of minor compounds, affecting the limit of detection (LOD). In biological sample analysis, this could reduce the capability of detecting specific minor compounds in the tissue that can have a relevant role in some diseases. Losing this information plays a decisive (negative) role in the study of bioclinical samples. Then, identifying and reducing the noise contributions affecting the analysis is of paramount importance. Numerous methods are available for reducing the influence of noise on acquired signals and enhancing the signal-to-noise ratio (SNR) in analytical measurements. One option is to optimize the instrumental hardware, such as the detector cooling system or adequate gain level selection, which can minimize the noise generated by the detector reading process. We can indeed always make efforts at the instrumental level to reduce noise, but ultimately, this remains quite limited in this particular framework, and therefore, another approach is required: using signal processing and machine learning tools to clean up the signal as much as possible and improve the quality of the analytical information [100]. Overall, selecting a specific method for reducing noise and enhancing SNR depends on the nature of the data and the specific analytical application. In this context, denoising can be applied either spectrum by spectrum or across the entire dataset. In the first case, denoising can be applied almost in real time; however, if optimization is needed, especially for a heterogeneous sample, a complete spectral dataset is required. Quite intuitively, we can imagine that a denoising method based on analyzing the entire dataset will better understand the structure of the signals, and therefore perform more effectively in this filtering task. In the LIBS domain, the most commonly used methods for denoising are Savitzky-Golay smoothing, Fast Fourier Transform, wavelet-based filtering, and the Whittaker Filter (Whitsm), all based on working on each spectrum individually [101] to complement these four widely used techniques, we propose also exploring Principal Component Analysis (PCA) in this denoising framework. It's worth noting that PCA has already been utilized in combination with other methods to explore LIBS imaging dataset [102][103] however, to our knowledge, this is the first time this approach has been applied for denoising purposes in the LIBS domain.

4.1 Experimental part

4.1.1 LIBS experimental set-up

The LIBS experimental set-up comprised a kHz laser (Cobolt Tor XE, $\lambda = 1064$ nm) capable of achieving a shooting rate of 1000 Hz and the laser-focusing optics (x10 bean expander followed by an x5 objective). The sample is places on a set of XYZ linear motorized stages to displace precisely the sample during the analysis. The elemental images presented in this work were recorder with a lateral resolution (e.g., shot-to-shot distance) of 10 μm, with a typical crater diameter aroung 7 μm. Each of them were composed of 1400 by 1500 pixels, derived from 2.1 milion acquired spectra. The detection system is based on an Andor-s iStar sCMOS sensor coupled to an Andor's Kymera Czerny-Turner spectrograph. The experimental conditions were fixed to 5 μs as integration time without delay time.

4.1.2 Data Acquisition

The sample analyzed was a rat kidney, collected 1 hour after intravenous injection of Au NP (size < 10 μm). The sample was fixed in paraformaldehyde 4% solution for 1 hour, before epoxy/embedding following previously reported procedure [104]. The sample was cut following the transversal axis. Then, the surface of the sample was polished before analysis.

4.1.3 Generation of chemical images from LIBS data

The typical μLIBS-imaging data processing workflow involves extracting the analytical signal from the spectra dataset. This analytical signal could be the emission line's maximum intensity, area, or net area for each element detected. Once the signal is extracted, it is used to create a chemical imaging map of the inspected sample. In this work, the net peak ares was used as analytical signal calculated as the signal region's mean minus the background region's mean, calculated as shown in figure

81 below. In figure 81 is showed a gaussian peak as a reference as a signal for an

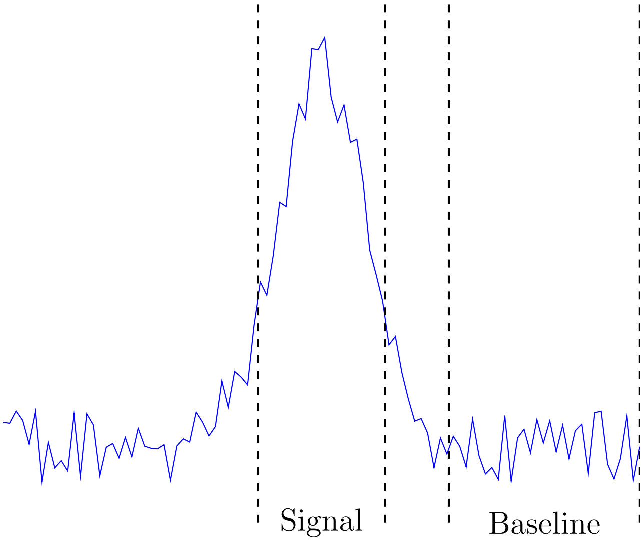


Figure 81: The integration of the *baseline* area is substracted to the integration of the *signal* area. In this way the signal is related only to the element that is examinated.

element of interest. Usually to display the chemical map of the element. The region *signal* is integrated and the region *baseline* is subtracted from it. This will give a net quantity of the signal. In the sample presented in this work, three elements have been considered: Phosphorus (P), presented in the whole tissue; Iron (Fe), related to the blood vessels; and Gold (Au), an exogenous element due to the presence of gold nanoparticles. The purpose of this work is not to provide an exhaustive exploration of this tissue but rather to demonstrate the validity of our approach on a set of elements with varying signal qualities within a single sample. Figure 82-a shows a visible image of the kidney analyzed in this study. On this same image, three specific position (denoted as L1, L2, and L3) were randomly selected so that it is possible to observe the associated LIBS spectra before and after applying the denoising strategies. The raw spectra acquired at these three positions are shown in figure82-c Figure 82-b shows the chemical maps of the three elements.

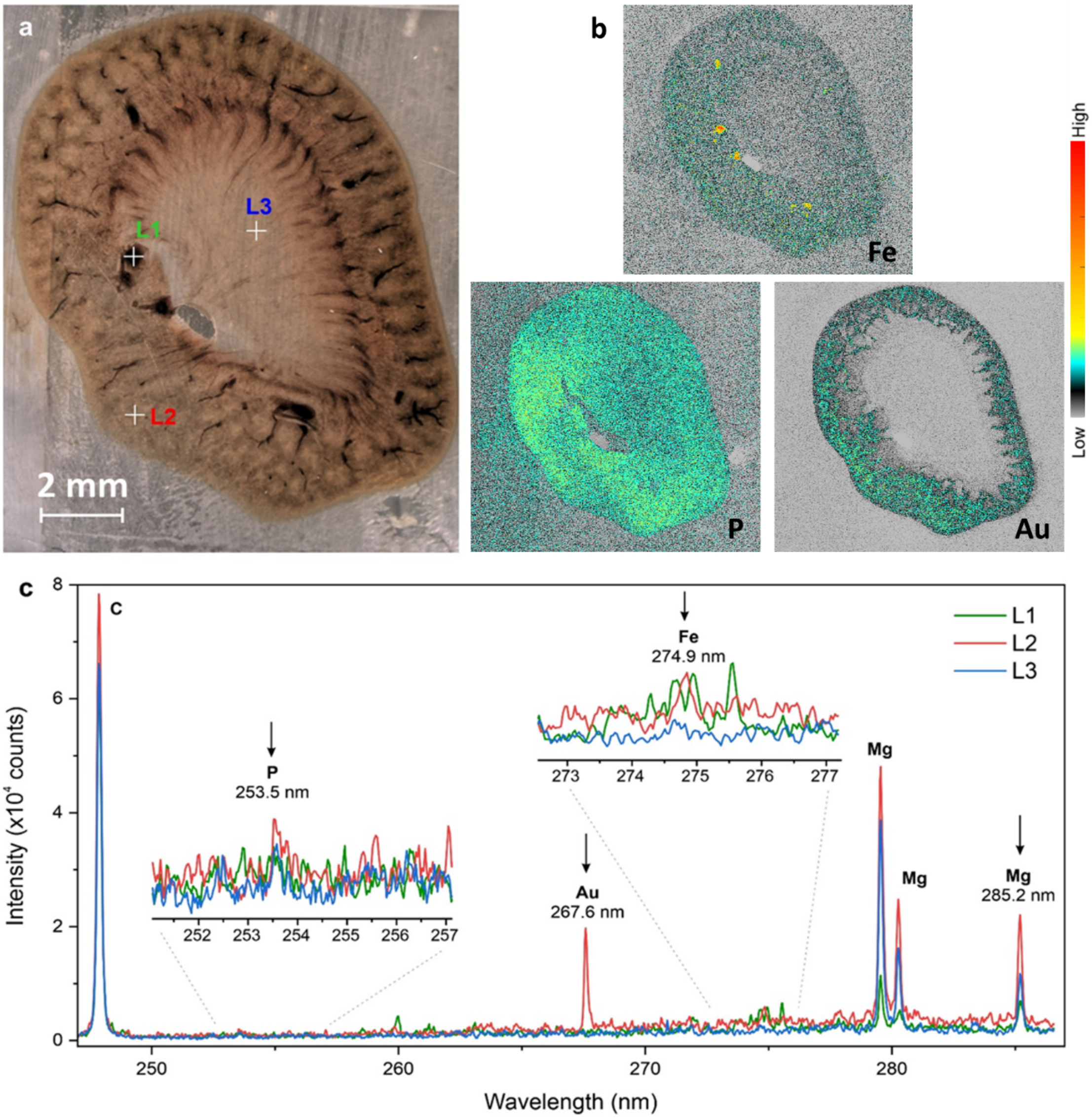


Figure 82: a) optical image of the analyzed sample b) Elemental distribution for different elements: P (253 nm), Au (267 nm), Fe (274 nm). c) Spectra obtained from the three selected positions, along with the specific spectral regions corresponding to the elements of interest

4.1.4 Denoising methods

In this work, different techniques for denoising were compared, with a focus on fast μLIBS-imaging for biomedical applications, particularly for analyzing endogenous and exogenous elements in tissue. Our approach is adapted to imaging applications:

rather than seeking a global denoising method for the entire spectral range, we aim to optimize denoising parameters for specific elemental emission peaks. This targeted strategy allows us to enhance the Signal-to-Noise Ratio (SNR) for individual elements of interest, recognizing that different elements may require distinct denoising parameters or methods for optimal results. The aim was to compare different denoising method used in LIBS imaging domain.

4.1.5 PCA

Principal Component Analysis [105] is a well-known method for the reduction of dimensionality, allowing a more straightforward representation and description of complex, multivariate datasets. The dataset X, applying PCA, is rewritten as a linear combination of the original variables to explain most of the variance, and it can be described as follows:

$$\widehat{X} = T \cdot P^T + E$$

where T is the scores matrix, i.e., the projection of the original data into the low-dimensional space; P is the loadings matrix, and E is the residuals one. Because noise contributes to a small percentage of the total variance, reconstructing the original matrix after selecting a given number of the very first principal components (PCs) with the higher explained variance will potentially provide a denoised dataset. As can be seen, this denoising procedure applies to the entire dataset simultaneously. The entire optimization of such a denoising method lies in selecting an optimal number of principal components. Indeed, selecting too few components would risk losing part of the signal of interest, while selecting too many would only introduce additional noise.

4.1.6 Savitzky-Golay

The Savitzky–Golay (SG) filter [106] smooths discrete data by locally fitting low-degree polynomials. For each point x_i, a symmetric window of $2n+1$ neighbouring samples, spanning x_{i-n} to x_{i+n}, is selected. A polynomial of order p is least-squares fitted to the values $y(x_{i-n}), \ldots, y(x_{i+n})$, and its value at x_i yields the smoothed estimate $y'(x_i)$. The two hyper-parameters that require tuning are the window length $2n+1$ and the polynomial order p. If the window is too narrow, high-frequency noise is retained; if it is too wide, genuine features (e.g. sharp peaks) may be blurred or lost. In this work the SG filter is applied independently to each spectrum in the dataset.

4.1.7 Fast Fourier Transform (FFT)

The Fast Fourier Transform (FFT) decomposes the original signal into a series of frequency components. The spectra are decomposed into a base of sinusoids with different frequencies.
Typically, spectral noise is associated with small fluctuations at high frequencies. On the other hand, relevant signals are associated with lower frequencies with higher contributions. By transforming the spectra into the frequency domain, if the noise has different frequencies compared to those of interest, reducing these frequencies to zero and applying the inverse FFT will produce a spectrum with less noise. The formula to calculate the DFT and the inverse IDFT is expressed in the following equation:

$$y_k = \sum_{n=0}^{N-1} x_n e^{-i\frac{2\pi}{N}nk} \text{ for } k = 0, 1, ..., N-1 \tag{30}$$

$$x_n = \frac{1}{N}\sum_{n=0}^{N-1} x_n e^{\frac{2\pi}{N}nk} \text{ for } n = 0, 1, ..., N-1 \tag{31}$$

To calculate the contributions to the signal for each frequencies is possible to use the *Power Spectrum Density (PSD)*, expressed by the equation 32

$$PSD = \text{fft}(x) * \text{conj}(\text{fft}(x))/n \tag{32}$$

As the results will show, this method is limited by the possibility of having the same frequency for meaningful and meaningless information. In this case, information will be lost.

4.1.8 Wavelet transform

Wavelets are used in chemometrics for different purposes [107]. They enables the decomposition at various levels of the spectra (figure 83 below). High-frequency noise, low-frequency baselines and intermediate components at different frequency levels can be captured. Wavelet decomposition involves recursively applying a matrix of wavelet filter coefficients to the signal [100]. After decomposition, the small coefficients are typically related to noise, while the large coefficients are associated with the significant features. Removing or reducing the smallest coefficients makes it possible to reduce noise without affecting the mail signal features. Wavelet Threshold Denoising (WTD) method has already been used in the LIBS domain by Schenke et al. [108] as a spectral denoising tool. The parameters to optimize with this approach are the type of wavelet function, the decomposition level, and the threshold choice. The decomposition level determines the number of times the signal is decomposed into approximate and detailed components. The selection of the threshold is crucial and challenging. However, we used in this work the universal threshold method proposed by Donoho et al. [109].

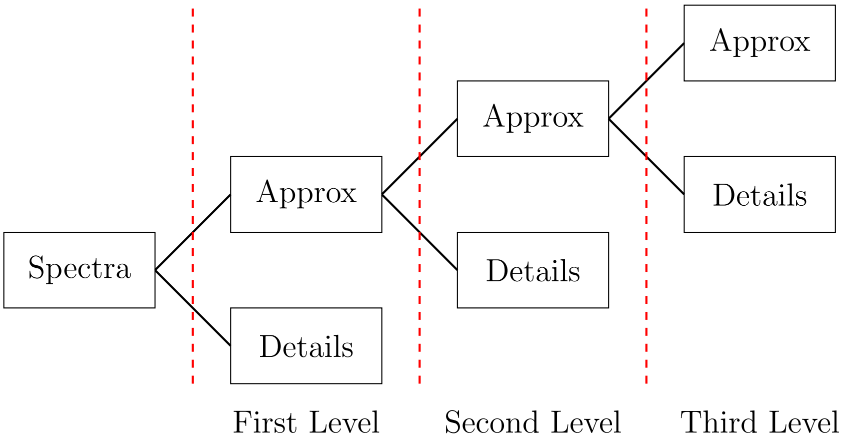


Figure 83: Wavelet decomposition

The method involves three steps: (1) Apply the Wavelet transform with a decomposition level chosen in advance (in this case four); (2) Select a proper threshold for each level; (3) Filter the coefficients below the threshold; and (4) reconstruct the signal with the inverse wavelet transform. There are two main types of filtering, both evaluated during this works: *hard* and *soft*.

With the hard filtering, all the coefficients below the selected threshold are forced to be zero. With soft filtering, those components are reduces be a given threshold. Regarding the type of wavelet, we limited ourselves to Daubechies wavelets, which have proven effective in many spectroscopies beyond LIBS [110].

4.1.9 Whittaker Smoothing

The Whittaker filter[111] function is based on penalized least squares. Considering the original signal as y and the smoothed one as z, Whittaker Smoothing modifies the signal, balancing two conflicting objectives: (1) data fidelity and (2) roughness of z.

$$\underset{z}{min} |\sum_{i}(y_i - z_i)^2 + \lambda \sum_{i}(z_i - z_{i-1})^2| \tag{33}$$

A larger λ increased the weight of $\sum_i (z_i - z_{i-1})^2$ that is the distance of two consecutive points. As results, the discrepancy between two consecutive points should be as less as possible, increasing lambda.

4.1.10 Figures of merit

One of the critical points for better selecting and optimizing the denoising method is the definition of a signal-to-noise ratio (SNR) criterion. Different methods for the calculation of the SNR available to evaluate the quality of a spectral denoising procedure. The approach used in this work for the calculation of SNR based on the spectra. It will be denoted as SNR_{sp} from this point. Considering a given elements, the SNR was calculated as follow:

1. **Estimation of the noise level.** A specific spectral range is selected for the estimation. The 269.2-270.8 nm range is chosen in our case as it does not exhibit significant spectral contribution. Thus, for each spectrum in the dataset, the standard deviation of the measured values within this spectral range is calculated. This dispersion provides an estimate of the noise level for a given spectrum. Then the mean of all these standard deviations is calculated, as a global estimation of noise in the dataset.

2. **Estimation of the signal level.** A spectral region around the emission line of an element of interest is selected. The average of these integration values for all the spectra in the dataset, retaining only those within the 99.99th percentile beforehand. This allow us to avoid biasing the estimation due to potentially outlier values while ensuring the detection of weak signals.

3. **SNR estimation** SNR is then calculated by taking the ratio of the two previously estimated values.

Spectra and images are examinated before and after the correction to better understand the impact of the method on both spectral and spatial levels.

4.2 Results and discussion

The following section will present the optimization for the different denoising methods. We want to show here that beyond selecting a given method, each one uses parameters that must be optimized to achieve the best denoising. Figure 84 represents a schema summarizing the data treatment process used in this work.

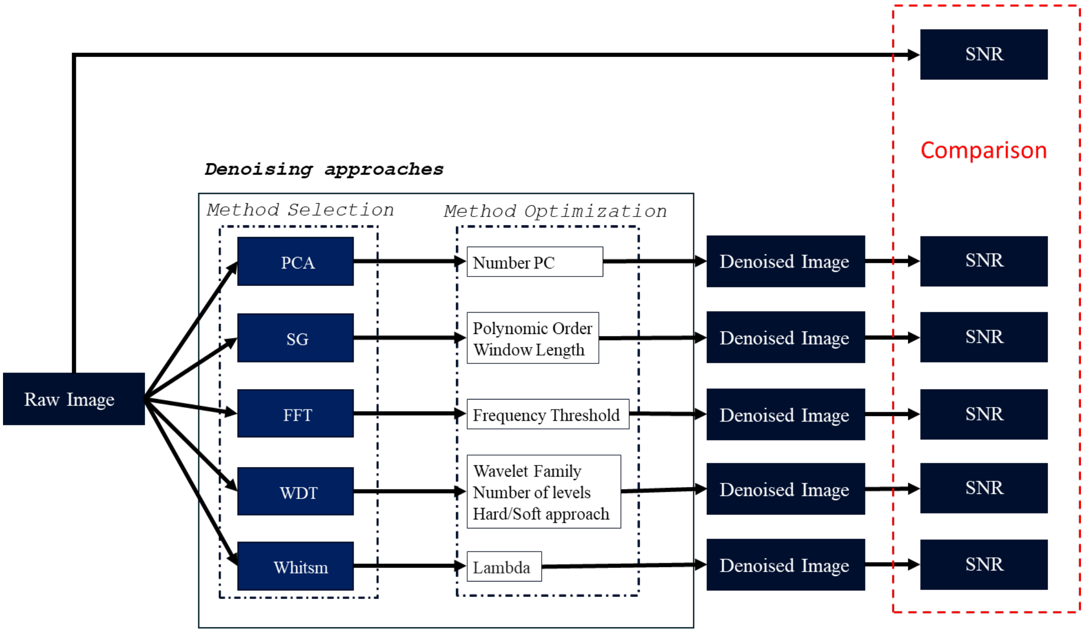


Figure 84: Data strategy denoising

Typically, in the case of imaging results based on spectroscopic techniques, we generally carry out the denoising, before or after extracting the analytical signal. The denoising after extracting the analytical signal is based on the use of image treatment methods; however, in this work, we will only focus on denoising the spectral information before extracting the analytical signal. The original SNR values are: P 119, Au 435, Fe 103. It can see that the highest values naturally come from the elements with the highest average emission level, e.g., Au, across the entire dataset. These values will enable us to assess the effectiveness of the different denoising methods. The proposed methods require an optimization process to obtain the optimum $\mathrm{SNR}_{s}p$ in order to obtain the best image reconstruction possible, e.g., we have some a priori knowledge of the elements to be optimized, allowing us to evaluate the SNR improvement for each of the element's analytical signals separately. This optimization step would need to be redone if we manage spectral data from a new sample with a different composition.

4.2.1 PCA

PCA was applied to the entire dataset also considering the whole spectral domain. As previously stated, the denoised approach based on PCA is based on the fact that the first principal components (PCs) extracted reflect the chemical variances, and beyond a specific component, only noise is captured. Consequently, selecting an appropriate number of PCs is an essential optimization step. The number of PCs selected is optimized for each element based on the highest SNR improvement (ratio between the calculated SNR and the raw spectra SNR), which varies as a function of the number of retained PCs, as shown in Fig. 85-a. The signal-to-noise ratio is improved by a factor of approximately 5 for the three elements when PCA is used, which is quite remarkable. Another way to look at it would be to say that, theoretically, the signal-to-noise ratio observed after this denoising could potentially be achieved with instrumentation 25 times slower, operating not in the kHz range but at 40 Hz. Beyond this initial observation, it must be highlighted that an optimal number of different components is obtained for each element. As a result, optimal denoising is achieved for the elements P, Au, and Fe with a total number of principal components equal to 10, 9, and 23, respectively. This result is quite logical, as each one exhibits varying degrees of variance within the spectral dataset and is, by definition, described by a different number of principal components. Beyond these signal-to-noise ratio values measured on the raw data and then on the potentially denoised data, it also seemed important to us to assess the relevance of a denoising strategy in terms of preserving the spectral information contained in the spectra. Indeed, such processing could, in extreme cases, distort or even eliminate an emission line of interest. Figure 86 shows the evolution of the L1, L2, and L3 spectra, whose locations were presented in Fig. 82, before and after the application of a given denoising method.

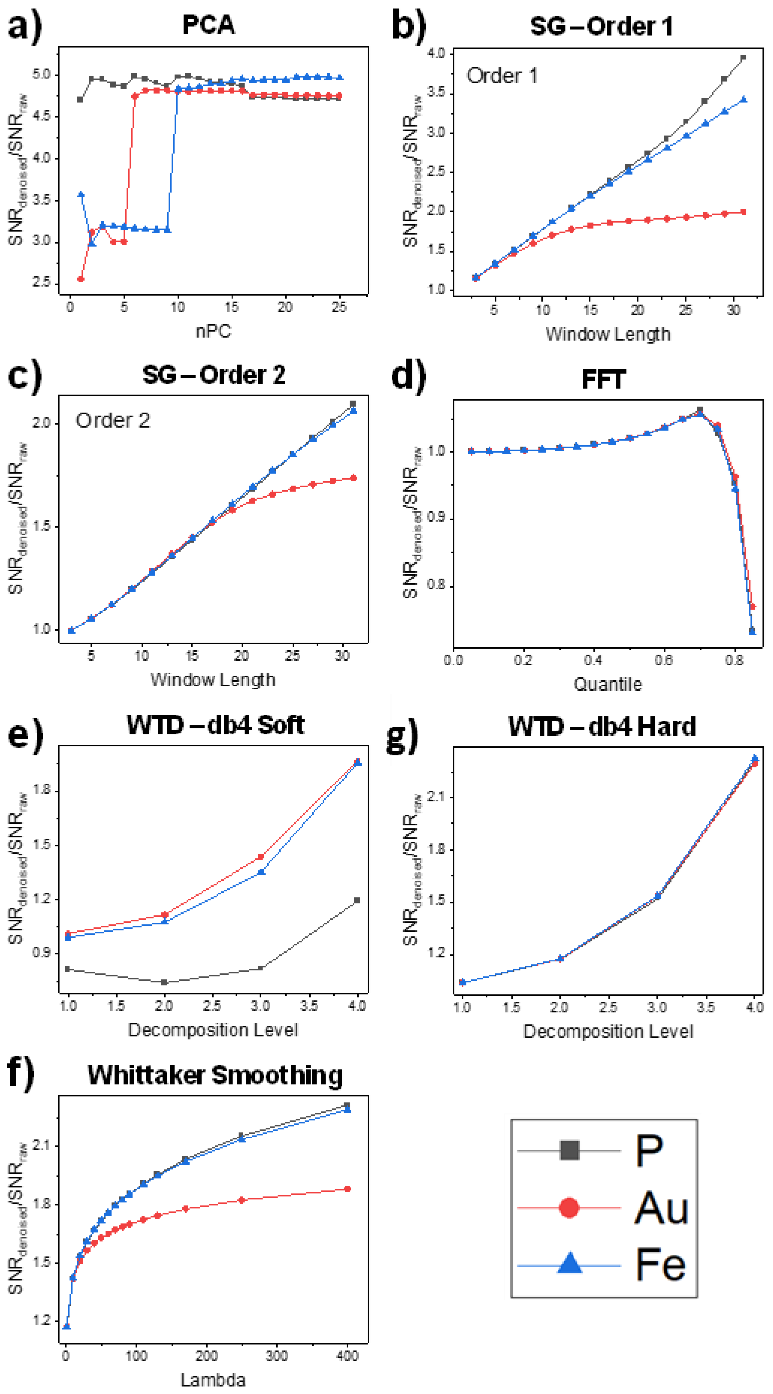


Figure 85: SNR evolution for the different methods.

To be more precise, only the spectral regions of interest are considered for each element. The visual comparison of the spectra enables us to observe the denoising

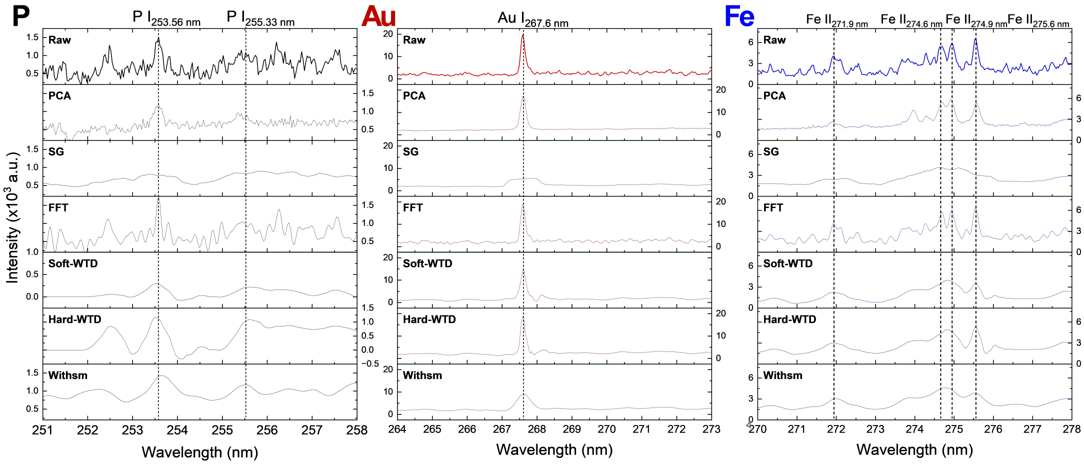


Figure 86: Denoised Spectra for the SNR optimal values with all the studied denoising methods. Fe spectrum corresponds to the point-of-interest L1, Au to L2, and P to the point-of-interest L3, denoted in Figure 87

efficiency of PCA for all elements while preserving the associated emission lines, ensuring no distortion.

At first glance, some readers might argue that, on the contrary, distortions are indeed observed when PCA is applied for denoising. We are not actually referring to distortions across the entire spectrum but only to the emission line related to the element of interest, which is quite different. Thus, referring to the emission lines of P, Au, and Fe at 253.56 nm, 267.6 nm, and 274.9 nm respectively, it is possible to observe that, compared to the raw data, their profiles are preserved after PCA denoising. Regarding the emission line observed around 253.5 nm in the raw data, it is not inconsistent for it to disappear after PCA denoising. Indeed, this line is not associated with phosphorus but with another element, which is likely represented by principal components beyond the optimal number chosen for denoising in this specific context of generating a phosphorus image. Building on these results, it is worth pausing to consider the applicability of this denoising approach to a dataset of this scale, comprising 2.1 million spectra. The computations, performed in a MATLAB environment, required in our case 23 GB of RAM, which remains entirely feasible within the computational frameworks used in spectroscopic imaging. Should memory constraints arise for certain users, adopting an HDF5 data format is recommended, as it allows for sequential access to smaller subsets of the spectral dataset as needed. Under such conditions, an alternative implementation of the

PCA algorithm, such as incremental PCA, should be considered to accommodate this segmented data exploration [112].

4.2.2 Savitzky-Golay smoothing

The parameters to optimize for the Savitzky-Golay smoothing are the polynomial order and the window size. The first and second orders were tested, while the window length were between 2 and 31 (with a step of 2). All combinations of these two parameters have thus been studied. When examining the results presented in figure 85-b and c, it is noticeable that the signal-to-noise ratio tends to stabilize when the window size reaches approximately 20, regardless of the polynomial order for Au. The improvement in the signal-to-noise ratio is, however, relatively limited in this case to a value of approximately 2, whereas it was 5 for PCA. The signal-to-noise ratio appears to increase consistently for the elements P and Fe, but this trend does not reflect a true improvement in spectral quality in terms of noise. Indeed, in Fig. 86, we observe completely distorted emission lines for the elements of interest following such correction, which is unacceptable. Therefore, Savitzky-Golay smoothing is not optimal for our LIBS imaging data. This approach is indeed very effective for vibrational spectroscopies, for example, but the bands observed are much broader than the noise structure. This is absolutely not the case in LIBS, where we observe very sharp emission lines.

4.2.3 Fast Fourier Transform

Once the signal is decomposed into the frequency domain with the FFT approach, the frequencies that are not related to the signal of interest are set to zero. To achieve this, the Power Spectrum Density (PSD) was calculated and all the frequencies below a pre-determinated threshold value was set to zero to eliminate all frequencies that do not significantly contribute to the signal. The different values used as thresholds are quantiles of the PSD. Various quantiles, ranging from 0.05 to 0.85 with a step of 0.05, were evaluated. Figure 85-d shows an optimal $\mathrm{SNR}_{\mathrm{sp}}$ value of 1.06 for a 0.70 quantile threshold regardless of the element considered. Consequently, the improvement in the SNR is very minimal, if not negligible, which is also evident in the spectra shown in figure 86.

4.2.4 Wavelet Threshold Denoising (WTD)

The wavelet family used in this case was db4 (Daubechies 4), with a hard and soft denoising approach. This wavelet family was selected as it has demonstrated its effectiveness in numerous spectroscopic techniques [113, 114]. Up to the fourth level of decomposition was evaluated for both hard and soft denoising, as going further in the decomposition would introduce artifacts in the spectra. Figure 85-e and f present a situation quite comparable to the previously introduced methods. For both strategies, the SNR consistently increases with the level of decomposition, though the improvement remains modest, barely reaching 2.5. Even worse, figure 86 shows that applying such a denoising method removes nearly all the chemical information originally present in the spectra.

4.2.5 Whittaker smoother

The lambda value was optimized in this case, ranging from 1 to 400. Fig. 85-g demonstrates that we quickly reached a plateau in the improvement of the signal-to-noise ratio for all three elements of interest. This improvement remains limited, as it averages around 2 times. It can be observed in Fig. 86 that, even though the denoised spectra using this approach appear significantly better than those obtained for SG, FFT, and WTD, the emission lines are noticeably broadened compared to the raw data.

Based on these results, PCA-based denoising procedure is undoubtedly more suited for LIBS imaging. This is a particularly interesting finding, as the other techniques studied here generally perform quite well in the context of other spectroscopic methods. This can, of course, be explained by the specific nature of LIBS data, especially the presence of particularly narrow emission lines. There is the need to emphasize the importance of optimizing the number of components for each element. The final step of this work now involves observing the effects of the denoising procedure using PCA on the integration images. Figure 87 thus presents a comparison of the integration images obtained from the raw spectra and the spectra denoised using PCA for the three elements considered. Starting with a global observation of this figure, we can see that the largest differences are observed for Fe and P. The differences are indeed less noticeable for Au. This is naturally explained by the fact that the weakest signals are associated with Fe and P, and it is

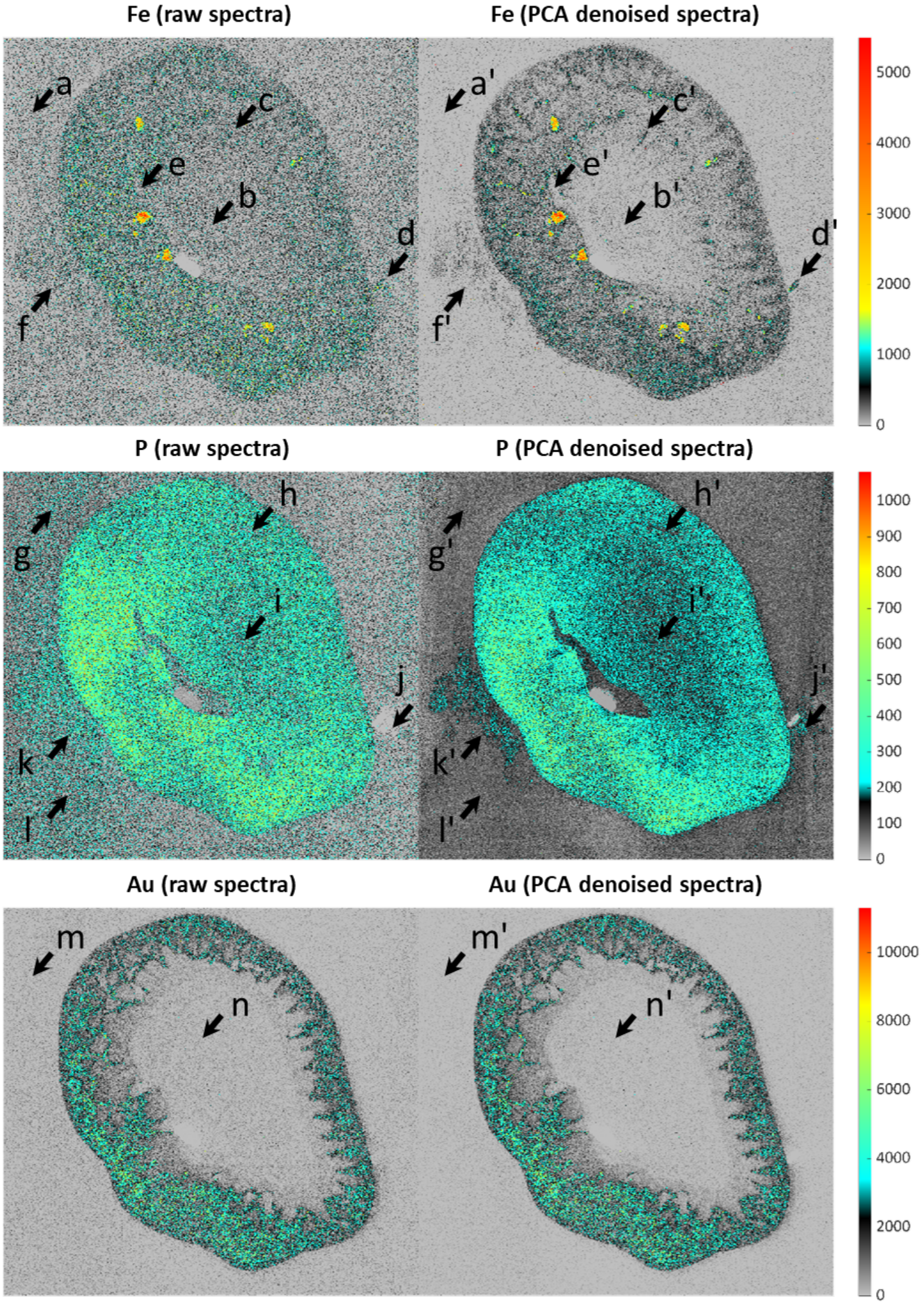


Figure 87: comparison between the original and denoised elemental distribution maps for iron, phosphorus and gold.

precisely under these conditions that an increase in the signal-to-noise ratio has a significant impact. Arrows have also been added to these images to highlight details or areas that differ significantly between them and will therefore be discussed in greater detail. Analyzing the iron maps, it can be observed an overestimation of concentrations when raw spectra are used, as seen, for example, in positions a vs. a' and b vs. b'. This iron originates from blood and is particularly present in the vascularized areas of the organ. Therefore, it cannot be found outside the organ or at its center, which is much more consistent in the image obtained after PCA denoising. A low dynamic range and very low contrast are also observed in the iron image derived from the raw spectra—values that are significantly improved after denoising, allowing the observation of previously unseen details (d vs. d', e vs. e', and f vs. f'). Fairly similar observations can be made for P. However, these differences are less striking, as it is known that phosphorus is present in numerous cells and is distributed throughout the organ. As a result, overestimations of phosphorus concentrations are once again observed outside outside the organ and at its center (g vs. g', i vs. i', and l vs. l') when raw spectra are used. Imperceptible details from the raw data are also revealed when the spectra are denoised (h vs. h', j vs. j' and k vs. k'). Regarding gold, our biological understanding of the issue allows us to state that it can only be present at the periphery of the organ. As a result, an overestimation of concentrations is observed both outside the organ and at its center, which is much more consistent when denoised spectra are used. Through these three elements with varying concentration ranges and locations within this complex organ, we were able to demonstrate that applying PCA as a denoising method generates elemental images with improved contrast, greater dynamic range, and, most importantly, reduced bias, allowing us to better highlight the biological reality of the sample after processing.

4.2.6 Conclusion and perspective

The use of new experimental approaches that enable higher throughput and higher resolution analysis makes the parallel development of new chemometric tools mandatory. As the complexity and size of spectral data increases - with hypercubes that can hold millions of spectra - it is necessary to develop algorithms and workflows for spectral processing to handle, analyze, and extract analytical information from these data. In this work, the application of kHz μLIBS-imaging for the analysis of samples of bio-clinical interest was highlighted, with a focus on a comparative evaluation of 5 different denoising methods. Furthermore, to our knowledge, this research applies principal component analysis (PCA) and Whittaker Smoothing to LIBS data for the first time, opening new ways to improve the accuracy of such analyses. The results shows that PCA is by far the most effective method in this specific LIBS framework, offering a better enhancement than the other methods. Specifically, PCA provides an important SNR enhancement of approximately 5 times for the three elements under study compared to the raw data; moreover, no distortion of the emission line of a given element has been found, unlike the other denoising methods studied. In this work, PCA was applied to the entire available spectral range, but in more delicate cases, we could consider using the restricted spectral range around the emission line of interest for the PCA calculation. This possibility was explored but any significant improvement was observed, at least for this particular dataset. In conclusion, this enhancement in the quality of the kHz μLIBS-imaging highlights the PCA value for the data treatment of LIBS-based applications, particularly where the experimental conditions limit the quality of spectral data. Recently, the explosion of deep learning has also contaminated chemical research. Among the different tools that can be used, autoencoder[115, 116] it is one of the most suitable, talking about denoising approaches. There can be two kinds of strategies to train an autoencoder. The matrix X' can be simulated and noise is added (X). In this way the original matrix without noise (X') is used as reference. The autoencoder will learn the noise. This modality implies that there is a good knowledge of the type of noise. If this knowledge is not available, the other way, maybe the most interesting for this case, it is to use the original data ad input (X) and output (X') at the same time. However, in this case, if the net is not well built, it will learn the identity, so it won't denoise the data, as it will try to reconstruct exactly the data. Different strategy can be applied to avoid this. For example, using a proper cross-validation to test the autoencoder to spectra that was not used to build the autoencoder, ad in this way will be selected the autoencoder that predict properly the common factor and not the noise of the training data. It is a methodology all the more justified since deep-learning approaches require large amounts of data to perform well, which is exactly what we have in kilohertz LIBS imaging, where acquiring several million spectra in a reasonable time is easy nowadays.

Chapter 5

General conclusions and perspectives

In recent years, the need for robust and unbiased methods to analyze complex hyperspectral imaging data has emerged as a key challenge. An increase in spatial and spectral resolution leads to larger data sets, which makes analysis more challenging. If the data processing workflow is not well suited, small variations may be overlooked and the interpretation of the results may be biased. This can be problematic in cases where these small variations are meaningful. For instance, when analyzing biological tissue, there may be different stages between healthy and diseased tissue, each of which may differ slightly from the previous or subsequent stage. For this reason, this PhD project began with the specific aim of developing advanced clustering methods to enhance the analysis of cancerous tissue. This was achieved by addressing the challenge in several parallel ways, with the aim of analyzing the impact of each method or approach. The decision to work on them separately was made because it represented a progressive evolution of knowledge about the problem, rather than being based on the OVAT approach. The outcomes of this PhD work can be grouped into five key areas: (1) a comparison between bisecting kmeans and QDD kmeans; (2) the development of stopping criteria for a fully automatic hierarchical clustering algorithm; (3) the integration of spatial information into clustering processes; (4) the combined use of MALDI and LIBS imaging to enhance tissue interpretation; and (5) facing new challenges in processing data from kHz LIBS imaging. Although these areas were explored in parallel rather than sequentially, advances in one often informed progress in another, reflecting the iterative and interdisciplinary nature of the research.

During the first part, the Quality-Driven Divisive (QDD) kmeans algorithm was introduced, an adaptive hierarchical-clustering strategy expressly designed for the high-dimensional spectral cubes produced by MALDI-MSI. By determining the optimal number of clusters at every split with the silhouette score, QDD prevents the arbitrary mergers and oversplits that plague conventional bisecting kmeans, thereby maximising intracluster compactness and intercluster separation. When applied to lipid images of oral tongue squamous-cell carcinoma, the algorithm clearly segregated cancer cells, inflamed stroma, nerves and edematous muscle while generating masks that exclude matrix only pixels, a level of anatomical fidelity the bisecting approach could not reach. These gains translate into molecular maps that are sharper, less biased and more biologically interpretable, positioning QDD kmeans as a robust platform for future high throughput, untargeted tissue characterisation in biomedical and clinical research.

In continuation with the previous part, the necessity of an automated hierarchical approach emerged to enable a proper, unbiased data exploration. During this part, the split at each division level was performed using kmeans clustering with $k = 2$, for simplicity. The goal was not to compare the clusters based on conventional descriptors, but rather to address the question: *“Are the generated clusters actually different from each other?”*. If the answer was no, then the division would have no meaningful justification. The results presented in this work were obtained by applying PCA to investigate this question. However, as stated earlier, further methods for properly comparing the extracted information could be explored. Additionally, discussing the outcomes with a bio-expert is an essential step, however was not feasible due to time constraints.

HSI play a key role among the analytical techniques however, the spatial resolution power of these techniques is not always used properly. During the standard procedure, the data cube is unfolded, resulting in a loss of spatial information. This can limit the analysis, as the pixels are not fully independent, in fact considering a pixel as a cell in biological samples, adjacent cells are connected. Two possible situations were introduced as examples. This demonstrates that using only spectral information is not always sufficient to answer the research question, and highlights the importance of the right viewpoint. In both cases, the answer to the question ”How many clusters are there?” is either 2 or 3, depending on the point of view.

The same information can have different meanings in different environments. The purpose of this section was to demonstrate that, by considering neighbouring pixels, the heterogeneity of tissue can be explored more effectively.

These three works are connected each other. As affirmed during the general conclusions of chapter 2, the next step will be to try to fuse them, using spatial information applied to an automatic hierarchical QDD kmeans.

The fourth part involved the fusion of RGB and LIBS imaging for HSI fusion. This work was interesting due to the possibility of enhancing the information without increasing the cost of the analysis thanks to visible images, which are normally acquired by all light microscopes. This was demonstrated by showing an example where, even though two quartz crystals looked very similar in the LIBS data and were almost impossible to distinguish, there were some chemical differences, probably due to traces. The fusion of RGB and LIBS made it possible to extract this information. This study has proven that even simple-looking RGB has high potential for enhancing the interpretation of complex samples without additional cost.

Techniques such as MALDI and LIBS are among the most widely used in molecular and elemental imaging, respectively. However, they have the disadvantage of being destructive and based on different principles. Furthermore, to our knowledge, there is no instrument that can acquire the two signals simultaneously. That's why two consecutive sections were used, and to mitigate small differences that can occurs between them a spatial binning was done. The results has shown a simply but efficient method to study and observe the relationship between molecular and elemental information . To well establish the approach, new samples need to be analyzed, additionally trying to do a deeper interpretation of the results, collaborating with biologists.

The last but not the least, with the denoising work the aim was to demonstrate that integrating advanced chemometric methods is indispensable for fully exploiting kilohertz-rate μLIBS imaging, which can capture millions of spectra in just a few minutes. Across the five denoising strategies compared, principal component analysis emerged as by far the most effective, preserving spectral line shapes and boosting the signal-to-noise ratio for P, Fe and Au by roughly a factor of five over the raw data. This optimization produces high-contrast, wide-dynamic-range elemental maps that reveal features previously invisible and reduce false positives in low-concentration regions. Overall, the workflow established there lays the foundation for ultra-sensitive, high-throughput biomedical analyses and opens the way to future quantitative studies on fragile specimens.

Besides all that has been proposed, several ideas and ways to improve the approaches were planned. However, due to time constraints not all of them were fully explored. As expressed more than one time during this thesis, the first three parts need to be combined, to have a more robust and proper automated approach for the exploration of biological samples. The aim is to develop a hierarchical clustering method based on QDD with an optimization step to determine the optimal number of clusters for each division. This will be achieved by using the evaluation (based on PCA) to accept or reject the optimal division. Once this method has been developed, it would be interesting to assess the impact of spatial analysis on segmentation using this automated approach, by testing different neighbourhood pixel descriptors and grid sizes. At the same time, on both fusion RGB-LIBS and LIBS-MALDI, other samples need to be analyzed to establish the robustness of the approaches, or potentially improve them based on the results. Ultimately, to improve the data obtained by kHz LIBS systems, a deep learning method such as an autoencoder could be employed to propose new approaches that are not well explored in this field.

In conclusion, the foundations laid in this project provide a solid basis for further investigation. Although this work has reached important milestones, further development is needed to fully realize its potential, particularly through improved interdisciplinary collaboration. Better communication with the biological team is needed to interpret the results in a biological context. This will be the work required to complete the research initiated during this PhD

Bibliography

(1) Jolivet, L.; Leprince, M.; Moncayo, S.; Sorbier, L.; Lienemann, C.-P.; Motto-Ros, V. Review of the recent advances and applications of LIBS-based imaging. *Spectrochimica Acta Part B: Atomic Spectroscopy* **2019**, *151*, 41–53.

(2) Pagnotta, S.; Lezzerini, M.; Ripoll-Seguer, L.; Hidalgo, M.; Grifoni, E.; Legnaioli, S.; Lorenzetti, G.; Poggialini, F.; Palleschi, V. Micro-laser-induced breakdown spectroscopy (micro-LIBS) study on ancient Roman mortars. *Applied Spectroscopy* **2017**, *71*, 721–727.

(3) Chang, C.-I., *Hyperspectral data processing: algorithm design and analysis*; John Wiley & Sons, Inc: Hoboken, NJ, 2013.

(4) Shehab, M.; Abualigah, L.; Shambour, Q.; Abu-Hashem, M. A.; Shambour, M. K. Y.; Alsalibi, A. I.; Gandomi, A. H. Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine* **2022**, *145*, 105458.

(5) Aichler, M. MALDI Imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. *Laboratory Investigation* **2015**, *95*, 10.

(6) Cremers, D. A., *Handbook of Laser-Induced Breakdown Spectroscopy*; John wiley & Sons: 2006.

(7) Sommella, E. e. a. MALDI Mass Spectrometry Imaging Highlights Specific Metabolome and Lipidome Profiles in Salivary Gland Tumor Tissues. *Metabolites* **2022**, *12*, Number: 6, 530.

(8) Balluff, B.; Hopf, C.; Porta Siegel, T.; Grabsch, H. I.; Heeren, R. M. A. Batch Effects in MALDI Mass Spectrometry Imaging. *Journal of the American Society for Mass Spectrometry* **2021**, *32*, 628–635.

(9) He, L.; Long, L. R.; Antani, S., et al. Histology image analysis for carcinoma detection and grading. *Computer Methods and Programs in Biomedicine* **2012**, *107*, 538–556.

(10) He, L.; Long, L. R.; Antani, S.; Thoma, G. R. Histology image analysis for carcinoma detection and grading. *Computer Methods and Programs in Biomedicine* **2012**, *107*, 538–556.

(11) Liu, X. e. a. Spatial multi-omics: deciphering technological landscape of integration of multi-omics and its applications. *Journal of Hematology & Oncology* **2024**, *17*, DOI: `10.1186/s13045-024-01596-9`.

(12) *MALDI mass spectrometry imaging: from fundamentals to spatial omics*; Siegel, T., Ed.; New developments in mass spectrometry 12; Royal Society of Chemistry: Cambridge, 2021.

(13) Gilmore, I. S.; Heiles, S.; Pieterse, C. L. Metabolic Imaging at the Single-Cell Scale: Recent Advances in Mass Spectrometry Imaging. *Annual Review of Analytical Chemistry* **2019**, *12*, 201–224.

(14) Ščupáková, K.; Balluff, B.; Tressler, C.; Adelaja, T.; Heeren, R. M. A.; Glunde, K.; Ertaylan, G. Cellular resolution in clinical MALDI mass spectrometry imaging: the latest advancements and current challenges. *Clinical Chemistry and Laboratory Medicine* **2020**, *58*, 914–929.

(15) 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* **2023**, *42*, 1927–1964.

(16) Alexandrov, T. a. METASPACE: A community-populated knowledge base of spatial metabolomes in health and disease. *bioRxiv* **2019**, DOI: `10.1101/539478`.

(17) Wenzel, T. J. Douglas A. Skoog, Donald M. West, F. James Holler, and Stanley R. Crouch: Fundamentals of analytical chemistry, 9th ed., international ed. *Analytical and Bioanalytical Chemistry* **2013**, *405*, 7903–7904.

(18) Duponchel, L.; Guerrini, R.; Ferreira, V. H.; Llamas, C. A.; Dujardin, C.; Motto-Ros, V. When Social Media Empowers Analytical Chemists to Explore Millions of Spectra Derived from a Complex Sample. *Analytical Chemistry* **2024**, *96*, 3994–3998.

(19) Alexandrov, T.; Becker, M.; Deininger, S.-O.; Ernst, G.; Wehder, L.; Grasmair, M.; von Eggeling, F.; Thiele, H.; Maass, P. Spatial Segmentation of Imaging Mass Spectrometry Data with Edge-Preserving Image Denoising and Clustering. *Journal of Proteome Research* **2010**, *9*, 6535–6546.

(20) McCann, A.; Rappe, S.; La Rocca, R.; Tiquet, M.; Quinton, L.; Eppe, G.; Far, J.; De Pauw, E.; Kune, C. Mass shift in mass spectrometry imaging: comprehensive analysis and practical corrective workflow. *Analytical and Bioanalytical Chemistry* **2021**, *413*, 2831–2844.

(21) Kumar, P.; Soumyashree, S.; Rao Epuru, N.; Banerjee, S. B.; Singh, R. P.; Subramanian, K. P. Determination of Stark Shifts and Widths Using Time Resolved Laser-Induced Breakdown Spectroscopy (LIBS) Measurements. *Applied Spectroscopy* **2020**, *74*, 913–920.

(22) Han, J., *Data Mining: Concepts and techniques*; Morgan Kaufmann: 2012.

(23) Bellman, R. Dynamic programming and Lagrange multipliers. *Proceedings of the National Academy of Sciences of the United States of America* **1956**, *42*, 767–769.

(24) Arthur, D.; Vassilvitskii, S. k-means++: The Advantages of Careful Seeding.

(25) Todeschini, R.; Ballabio, D.; Termopoli, V.; Consonni, V. Extended multivariate comparison of 68 cluster validity indices. A review. *Chemometrics and Intelligent Laboratory Systems* **2024**, *251*, 105117.

(26) Pakhira, M. K.; Bandyopadhyay, S.; Maulik, U. Validity index for crisp and fuzzy clusters. *Pattern Recognition* **2004**, *37*, 487–501.

(27) Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **1987**, *20*, 53–65.

(28) Calinski, T.; Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics* **1974**, *3*, 1–27.

(29) Deininger, S.-O.; Ebert, M. P.; Fütterer, A.; Gerhard, M.; Röcken, C. MALDI Imaging Combined with Hierarchical Clustering as a New Tool for the Interpretation of Complex Human Cancers. *Journal of Proteome Research* **2008**, *7*, PMID: 19367705, 5230–5236.

(30) Steinbach, M.; Karypis, G.; Kumar, V. A Comparison of Document Clustering Techniques.

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

(32) Verbeeck, N.; Caprioli, R. M.; Van de Plas, R. Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. *Mass Spectrometry Reviews* **2020**, *39*, Number: 3, 245–291.

(33) Trede, D.; Schiffler, S.; Becker, M.; Wirtz, S.; Steinhorst, K.; Strehlow, J.; Aichler, M.; Kobarg, J. H.; Oetjen, J.; Dyatlov, A.; Heldmann, S.; Walch, A.; Thiele, H.; Maass, P.; Alexandrov, T. Exploring Three-Dimensional Matrix-Assisted Laser Desorption/Ionization Imaging Mass Spectrometry Data: Three-Dimensional Spatial Segmentation of Mouse Kidney. *Analytical Chemistry* **2012**, *84*, 6079–6087.

(34) Bouveyron, C.; Girard, S.; Schmid, C. High-dimensional data clustering. *Computational Statistics & Data Analysis* **2007**, *52*, 502–519.

(35) Alexandrov, T.; Kobarg, J. H. Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. *Bioinformatics* **2011**, *27*, i230–i238.

(36) Gustafson, D. E.; Kessel, W. C. In *1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, 1978, pp 761–766.

(37) Bruand, J.; Alexandrov, T.; Sistla, S.; Wisztorski, M.; Meriaux, C.; Becker, M.; Salzet, M.; Fournier, I.; Macagno, E.; Bafna, V. AMASS: Algorithm for MSI analysis by semi-supervised segmentation. *Journal of Proteome Research* **2011**, *10*, 4734–4743.

(38) Chernyavsky, I.; Alexandrov, T.; Maass, P.; Nikolenko, S. I. In *Proceedings of the German Conference on Bioinformatics 2012*, 2012.

(39) Bemis, K. D.; Harry, A.; Eberlin, L. S.; Ferreira, C. R.; van de Ven, S. M.; Mallick, P.; Stolowitz, M.; Vitek, O. Probabilistic Segmentation of Mass Spectrometry (MS) Images Helps Select Important Ions and Characterize Confidence in the Resulting Segments. *Molecular & Cellular Proteomics* **2016**, *15*, 1761–1772.

(40) Ogrinc, N. Mass Spectrometry-Based Differentiation of Oral Tongue Squamous Cell Carcinoma and Nontumor Regions With the SpiderMass Technology. **2022**, *3*, 11.

(41) Yang, C.; He, Z.; Yu, W. Comparison of Public Peak Detection Algorithms for MALDI Mass Spectrometry Data Analysis. *BMC Bioinformatics* **2009**, *10*, 4.

(42) Han, J., *Data Mining: Concepts and techniques*; Morgan Kaufmann: 2012.

(43) Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: an overview. *WIREs Data Mining and Knowledge Discovery* **2012**, *2*, 86–97.

(44) Boyaval, F. et al. N-Glycomic Signature of Stage II Colorectal Cancer and Its Association With the Tumor Microenvironment. *Molecular & Cellular Proteomics* **2021**, *20*, 100057.

(45) Holst, S.; Heijs, B.; de Haan, N.; van Zeijl, R. J. M.; Briaire-de Bruijn, I. H.; van Pelt, G. W.; Mehta, A. S.; Angel, P. M.; Mesker, W. E.; Tollenaar, R. A.; Drake, R. R.; Bovée, J. V. M. G.; McDonnell, L. A.; Wuhrer, M. Linkage-specific in situ sialic acid derivatization for N-glycan mass spectrometry imaging of formalin-fixed paraffin-embedded tissues. *Analytical Chemistry* **2016**, *88*, 5904–5913.

(46) Strohalm, M.; Hassman, M.; Kosata, B.; Kodícek, M. mMass data miner: an open source alternative for mass spectrometric data analysis. *Rapid Communications in Mass Spectrometry* **2008**, *22*, 905–908.

(47) Lavine, B. In *Comprehensive Chemometrics*, Brown, S. D., Tauler, R., Walczak, B., Eds.; Elsevier: 2009, pp 601–607.

(48) J.M. Roger a, e. a. CovSel: Variable selection for highly multivariate and multi-response calibration: Application to IR spectroscopy. *Chemometrics and Intelligent Laboratory Systems* **2011**, *106*, Chimiométrie 2009, Paris, France, 30 November - 1 December 2009, 216–223.

(49) Nardecchia, A.; Vitale, R.; Duponchel, L. Fusing spectral and spatial information with 2-D stationary wavelet transform (SWT 2-D) for a deeper exploration of spectroscopic images. *Talanta* **2021**, *224*, 121835.

(50) Ahmad, M.; Vitale, R.; Silva, C. S.; Ruckebusch, C.; Cocchi, M. Exploring local spatial features in hyperspectral images. *Journal of Chemometrics* **2020**, *34*, e3295.

(51) Hu, H.; Yin, R.; Brown, H. M.; Laskin, J. Spatial Segmentation of Mass Spectrometry Imaging Data by Combining Multivariate Clustering and Univariate Thresholding. *Analytical Chemistry* **2021**, *93*, Number: 7, 3477–3485.

(52) Alexandrov, T.; Kobarg, J. H. Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. *Bioinformatics* **2011**, *27*, i230–i238.

(53) Bharati, M. H.; MacGregor, J. F. In ed. by McCann, H.; Scott, D. M., Boston, MA, 2001, p 27.

(54) Prats-Montalbán, J.; Ferrer, A. Integration of colour and textural information in multivariate image analysis: defect detection and classification issues. *Journal of Chemometrics* **2007**, *21*, 10–23.

(55) Jamme, F.; Duponchel, L. Neighbouring pixel data augmentation: a simple way to fuse spectral and spatial information for hyperspectral imaging data analysis. *Journal of Chemometrics* **2017**, *31*, e2882.

(56) Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347.

(57) *Data fusion methodology and applications*; Cocchi, M., Ed.; Data handling in science and technology volume 31; Elsevier: Amsterdam, Netherlands ; Cambridge, MA, 2019.

(58) Nardecchia, A.; de Juan, A.; Motto-Ros, V.; Fabre, C.; Duponchel, L. LIBS and Raman image fusion: An original approach based on the use of chemometric methodologies. *Spectrochimica Acta Part B: Atomic Spectroscopy* **2022**, *198*, 106571.

(59) Hoehse, M.; Gornushkin, I.; Merk, S.; Panne, U. Assessment of suitability of diode pumped solid state lasers for laser induced breakdown and Raman spectroscopy. *Journal of Analytical Spectrometry* **2011**, *26*, 414–424.

(60) Nardecchia, A.; de Juan, A.; Motto-Ros, V.; Gaft, M.; Duponchel, L. Data fusion of LIBS and PIL hyperspectral imaging: Understanding the luminescence phenomenon of a complex mineral sample. *Analytica Chimica Acta* **2022**, *1192*, 339368.

(61) Bedia, C.; Sierra, À.; Tauler, R. Multimodal multisample spectroscopic imaging analysis of tumor tissues using multivariate curve resolution. *Chemometrics and Intelligent Laboratory Systems* **2021**, *215*, 104366.

(62) Van de Plas, R.; Yang, J.; Spraggins, J.; Caprioli, R. M. Image fusion of mass spectrometry and microscopy: a multimodality paradigm for molecular tissue mapping. *Nature Methods* **2015**, *12*, 366–372.

(63) Tuck, M.; Blanc, L.; Touti, R.; Patterson, N. H.; Van Nuffel, S.; Villette, S.; Taveau, J.-C.; Römpp, A.; Brunelle, A.; Lecomte, S.; Desbenoit, N. Multimodal Imaging Based on Vibrational Spectroscopies and Mass Spectrometry Imaging Applied to Biological Tissue: A Multiscale and Multiomics Review. *Analytical Chemistry* **2021**, *93*, Publisher: American Chemical Society, 445–477.

(64) Tuck, M.; Grélard, F.; Blanc, L.; Desbenoit, N. MALDI-MSI Towards Multimodal Imaging: Challenges and Perspectives. *Frontiers in Chemistry* **2022**, *10*, 904688.

(65) Trichard, F.; Moncayo, S.; Devismes, D.; Pelascini, F.; Maurelli, J.; Feugier, A.; Sasseville, C.; Surma, F.; Motto-Ros, V. Evaluation of a compact VUV spectrometer for elemental imaging by laser-induced breakdown spectroscopy: application to mine core characterization. *Journal of Analytical Atomic Spectrometry* **2017**, *32*, 1527–1534.

(66) Manard, B.; Quarles, C. D.; Wylie, E.; Xu, N. Laser ablation – inductively coupled plasma – mass spectrometry/laser induced breakdown spectroscopy. *Journal of Analytical Atomic Spectrometry* **2017**, *32*, 1680–1687.

(67) Bonta, M.; Török, S.; Döme, B.; Limbeck, A. Tandem LA-LIBS coupled to ICP-MS for comprehensive analysis of tumor samples. *Spectroscopy Online* **2017**, *32*, 42–46.

(68) Mattes, D.; Haynor, D.; Vesselle, H.; Lewellen, T.; Eubank, W. In *Medical Imaging 2001: Image Processing*, SPIE Publications: 2001, pp 1609–1620.

(69) Rahunathan, S.; Stredney, D.; Schmalbrock, P.; Clymer, B. In *MMVR13: The 13th Annual Medicine Meets Virtual Reality Conference*; Poster presentation, Long Beach, CA, 2005.

(70) Vergara, J. R.; Estévez, P. A. A review of feature selection methods based on mutual information. *Neural Computing and Applications* **2014**, *24*, 175–186.

(71) R. L. Ornberg B. M. Woerner, D. A. E. Analysis of stained objects in histological sections by spectral imaging and differential absorption. *J. Histochem. Cytochem.* **1999**, *47*, 1307–1313.

(72) Foca, G.; Masino, F.; Antonelli, A.; Ulrici, A. Prediction of compositional and sensory characteristics using RGB digital images and multivariate calibration techniques. *Analytica Chimica Acta* **2011**, *706*, 238–245.

(73) Sadeghi, A.; Khani, S.; Sabourian, R.; Hajimahmoodi, M.; Ghasemi, J. B. Integrating CNNs and chemometrics for analyzing NIR spectra and RGB images in turmeric adulterant detection. *Journal of Food Composition and Analysis* **2025**, *141*, 107324.

(74) Kucheryavski, S. Extracting useful information from images. *Chemometrics and Intelligent Laboratory Systems* **2011**, *108*, Analytical Platforms for Providing and Handling Massive Chemical Data, 2–12.

(75) Vitale, R.; Prats-Montalbán, J. M.; López-García, F.; Blasco, J.; Ferrer, A. Segmentation techniques in image analysis: A comparative study. *Journal of Chemometrics* **2016**, *30*, 749–758.

(76) Bharati, M. H.; MacGregor, J. F. Multivariate Image Analysis for Real-Time Process Monitoring and Control. *Industrial & Engineering Chemistry Research* **1998**, *37*, 4715–4724.

(77) Antonelli, A.; Cocchi, M.; Fava, P.; Foca, G.; Franchini, G. C.; Manzini, D.; Ulrici, A. Automated evaluation of food colour by means of multivariate image analysis coupled to a wavelet-based classification algorithm. *Analytica Chimica Acta* **2004**, *515*, 3–13.

(78) Gowda, S. N.; Yuan, C. In *Computer Vision – ACCV 2018*, Jawahar, C., Li, H., Mori, G., Schindler, K., Eds.; Lecture Notes in Computer Science, Vol. 11364; Springer, Cham: 2019, pp 475–490.

(79) Busser, B.; Bulin, A.-L.; Gardette, V.; Elleaume, H.; Pelascini, F.; Bouron, A.; Motto-Ros, V.; Sancey, L. Visualizing the cerebral distribution of chemical elements: A challenge met with LIBS elemental imaging. *Journal of Neuroscience Methods* **2022**, *379*, 109676.

(80) Janovszky, P.; Kéri, A.; Palásti, D. J.; Bencze, I.; Pálinkás, J.; Geretovszky, Z.; Bozóki, Z.; Szabó, G. Quantitative elemental mapping of biological tissues by laser-induced breakdown spectroscopy using matrix recognition. *Scientific Reports* **2023**, *13*, 10089.

(81) Sancey, L.; Motto-Ros, V.; Kotb, S.; Wang, X.; Lux, F.; Panczer, G.; Yu, J.; Tillement, O. Laser-induced breakdown spectroscopy: a new approach for nanoparticle's mapping and quantification in organ tissue. *Journal of Visualized Experiments* **2014**, e51353.

(82) Skalny, A. V.; Korobeinikova, T. V.; Aschner, M.; Baranova, O. V.; Barbounis, E. G.; Tsatsakis, A.; Tinkov, A. A. Medical application of laser-induced breakdown spectroscopy (LIBS) for assessment of trace element and mineral in biosamples: Laboratory and clinical validity of the method. *Journal of Trace Elements in Medicine and Biology* **2023**, *79*, 127241.

(83) Lin, Q.; Wang, S.; Duan, Y.; Tuchin, V. V. Ex vivo three-dimensional elemental imaging of mouse brain tissue block by laser-induced breakdown spectroscopy. *Journal of Biophotonics* **2021**, *14*, e202000479.

(84) Wu, B.; Becker, J. S. Bioimaging of metals in rat brain hippocampus by laser microdissection inductively coupled plasma mass spectrometry (LMD-ICP-MS) using high-efficiency laser ablation chambers. *International Journal of Mass Spectrometry* **2012**, *323-324*, 34–40.

(85) Sabine Becker, J.; Matusch, A.; Palm, C.; Salber, D.; Morton, K. A.; Susanne Becker, J. Bioimaging of metals in brain tissue by laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS) and metallomics. *Metallomics* **2010**, *2*, 104–111.

(86) Becker, J.; Matusch, A.; Wu, B. Bioimaging mass spectrometry of trace elements – recent advance and applications of LA-ICP-MS: A review. *Analytica Chimica Acta* **2014**, *835*, 1–18.

(87) Hahn, D. W.; Omenetto, N. Laser-Induced Breakdown Spectroscopy (LIBS), Part II: Review of Instrumental and Methodological Approaches to Material Analysis and Applications to Different Fields. *Applied Spectroscopy* **2012**, *66*, 347–419.

(88) Gardette, V. et al. Quantifying Titanium Exposure in Lung Tissues: A Novel Laser-Induced Breakdown Spectroscopy Elemental Imaging-Based Analytical Framework for Biomedical Applications. *Small Science* **2024**, *4*, 2300307.

(89) Nardecchia, A.; Motto-Ros, V.; Duponchel, L. Saturated signals in spectroscopic imaging: why and how should we deal with this regularly observed phenomenon? *Analytica Chimica Acta* **2021**, *1157*, 338389.

(90) Gómez-Sánchez, A.; Vitale, R.; Ruckebusch, C.; de Juan, A. Solving the missing value problem in PCA by Orthogonalized-Alternating Least Squares (O-ALS). *Chemometrics and Intelligent Laboratory Systems* **2024**, *250*, 105153.

(91) Wang, Z.; Bovik, A. C.; Sheikh, H. R.; Simoncelli, E. P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* **2004**, *13*, 600–612.

(92) Finnegan, M. E.; Visanji, N. P.; Romero-Canelon, I.; House, E.; Rajan, S.; Mosselmans, J. F. W.; Hazrati, L.; Dobson, J.; Collingwood, J. F. Synchrotron XRF imaging of Alzheimer's disease basal ganglia reveals linear dependence of high-field magnetic resonance microscopy on tissue iron concentration. *Journal of Neuroscience Methods* **2019**, *319*, Epub 2019-03-06, 28–39.

(93) Popescu, B. F.; George, M. J.; Bergmann, U.; Garachtchenko, A. V.; Kelly, M. E.; McCrea, R. P.; Lüning, K.; Devon, R. M.; George, G. N.; Hanson, A. D.; Harder, S. M.; Chapman, L. D.; Pickering, I. J.; Nichol, H. Mapping metals in Parkinson's and normal brain using rapid-scanning X-ray fluorescence. *Physics in Medicine and Biology* **2009**, *54*, Epub 2009-01-09, 651–663.

(94) Genoud, S.; Roberts, B. R.; Gunn, A. P.; Halliday, G. M.; Lewis, S. J. G.; Ball, H. J.; Hare, D. J.; Double, K. L. Subcellular compartmentalisation of copper, iron, manganese, and zinc in the Parkinson's disease brain. *Metallomics* **2017**, *9*, 1447–1455.

(95) Guo, Z.; Chen, D.; Yao, L.; Sun, Y.; Li, D.; Le, J.; Dian, Y.; Zeng, F.; Chen, X.; Deng, G., et al. The molecular mechanism and therapeutic landscape of copper and cuproptosis in cancer. *Signal Transduction and Targeted Therapy* **2025**, *10*, Published online 9 May 2025, 149.

(96) Kopřivová, H.; Kiss, K.; Krbal, L.; Stejskal, V.; Buday, J.; Pořízka, P.; Kaška, M.; Ryška, A.; Kaiser, J. Imaging the elemental distribution within human malignant melanomas using Laser-Induced Breakdown Spectroscopy. *Analytica Chimica Acta* **2024**, *1310*, Epub 2024-05-03, 342663.

(97) Gardette, V.; Motto-Ros, V.; Alvarez-Llamas, C.; Sancey, L.; Duponchel, L.; Busser, B. Laser-Induced Breakdown Spectroscopy Imaging for Material and Biomedical Applications: Recent Advances and Future Perspectives. *Analytical Chemistry* **2023**, *95*, 49–69.

(98) Boué-Bigne, F. Analysis of Oxide Inclusions in Steel by Fast Laser-Induced Breakdown Spectroscopy Scanning: An Approach to Quantification. *Applied Spectroscopy* **2007**, *61*, 333–337.

(99) Alvarez-Llamas, C.; Tercier, A.; Ballouard, C.; Fabre, C.; Hermelin, S.; Margueritat, J.; Duponchel, L.; Dujardin, C.; Motto-Ros, V. Ultrafast micro-LIBS imaging for the multiscale mineralogical characterization of pegmatite rocks. *J. Anal. At. Spectrom.* **2024**, *39*, 1077–1086.

(100) Roger, J.-M. e. a. In *Comprehensive Chemometrics*; Elsevier: 2020, pp 1–75.

(101) Zhang, B.; Sun, L.; Yu, H.; Xin, Y.; Cong, Z. A Method for Improving Wavelet Threshold Denoising in Laser-Induced Breakdown Spectroscopy. *Spectrochimica Acta Part B: Atomic Spectroscopy* **2015**, *107*, 32–44.

(102) Finotello, R.; Tamaazousti, M.; Sirven, J.-B. HyperPCA: A Powerful Tool to Extract Elemental Maps from Noisy Data Obtained in LIBS Mapping of Materials. *Spectrochimica Acta Part B: Atomic Spectroscopy* **2022**, *192*, 106418.

(103) Pořízka, P.; Klus, J.; Képeš, E.; Prochazka, D.; Hahn, D. W.; Kaiser, J. On the Utilization of Principal Component Analysis in Laser-Induced Breakdown Spectroscopy Data Analysis: A Review. *Spectrochimica Acta Part B: Atomic Spectroscopy* **2018**, *148*, 65–82.

(104) Sancey, L. et al. Long-term in vivo clearance of gadolinium-based AGuIX nanoparticles and their biocompatibility after systemic injection. *ACS Nano* **2015**, *9*, 2477–2488.

(105) Bro, R.; Smilde, A. K. Principal component analysis. *Anal. Methods* **2014**, *6*, 2812–2831.

(106) Savitzky, A.; Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* **1964**, *36*, 1627–1639.

(107) Jetter, K.; Depczynski, U.; Molt, K.; Niemöller, A. Principles and applications of wavelet transformation to chemometrics. *Analytica Chimica Acta* **2000**, *420*, 169–180.

(108) Schlenke, J.; Hildebrand, L.; Moros, J.; Laserna, J. Adaptive approach for variable noise suppression on laser-induced breakdown spectroscopy responses using stationary wavelet transform. *Analytica Chimica Acta* **2012**, *754*, 8–19.

(109) Donoho, D. L.; Johnstone, I. M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, *81*, DOI: `10.1093/biomet/81.3.425`.

(110) Aberkane, S. M.; Melikechi, N.; Yahiaoui, K. In Palleschi, V., Ed., 2022.

(111) Eilers, P. H. C. A Perfect Smoother. *Analytical Chemistry* **2003**, *75*, 3631–3636.

(112) Ross, D. A.; Lim, J.; Lin, R.-S.; Yang, M.-H. Incremental learning for robust visual tracking. *International Journal of Computer Vision* **2008**, *77*, 125–141.

(113) Chen, D.; Shao, X.; Hu, B.; Su, Q. A Background and noise elimination method for quantitative calibration of near infrared spectra. *Analytica Chimica Acta* **2004**, *511*, 37–45.

(114) Ma, X.-G.; Zhang, Z.-X. Application of wavelet transform to background correction in inductively coupled plasma atomic emission spectrometry. *Analytica Chimica Acta* **2003**, *485*, 233–239.

(115) Zhang, C.; Zhou, L.; Zhao, Y.; Zhu, S.; Liu, F.; He, Y. Noise reduction in the spectral domain of hyperspectral images using denoising autoencoder methods. *Chemometrics and Intelligent Laboratory Systems* **2020**, *203*, 104063.

(116) Lan, R.; Li, Z.; Liu, Z.; Gu, T.; Luo, X. Hyperspectral image classification using k-sparse denoising autoencoder and spectral–restricted spatial characteristics. *Applied Soft Computing* **2019**, *74*, 693–708.

List of pubblications and conferences

Articles published

Related to the thesis

- **Optimization of denoising approaches in the context of ultra-fast LIBS imaging** Ruggero Guerrini, Cesar Alvarez-Llamas, Lucie Sancey, Vincent Motto-Ros, Ludovic Duponchel. *Spectrochimica Acta B* **2025** 227.

Not related to the thesis

- **When Social Media Empowers Analytical Chemists to Explore Millions of Spectra Derived from a Complex Sample.** Ludovic Duponchel, Ruggero Guerrini, Victor H.C. Ferreira, César Alvarez Llamas, Christophe Dujardin, and Vincent Motto-Ros. *Analytical Chemistry* **2024** 96 (10), 3994-3998 DOI: 10.1021/acs.analchem.3c05724

- **NMR Metabolomics of Arctium lappa L., Taraxacum officinale and Melissa officinalis: A Comparison of Spontaneous and Organic Ecotypes.** Ambroselli D, Masciulli F, Romano E, Guerrini R, Ingallina C, Spano M, Mannina L. *Foods 2024* 13(11) doi: 10.3390/foods13111642.

Articles submitted

Related to the thesis

- **Quality-Driven Divisive kmeans: A New Clustering Strategy for MALDI Imaging Data For a More Precise and Less Biased Characterization of Complex Biological Tissues** Guerrini, Ruggero; Ogrinc ,

Nina; COLIN, Emilien; TEBBAKHA, Riad; Attencourt, Christophe; Boudahi, Ahmed; testelin, Sylvie; Dakpe, Stephanie; Fournier, Isabelle; Duponchel, Ludovic. *Talanta*

Articles on going

Related to the thesis

- **Boosting clustering in MALDI imaging coupling spatial and spectral information** Guerrini, Ruggero; Ogrinc , Nina; Marini, Federico; Duponchel, Ludovic.

- **Towards automatic clustering in MALDI imagin for a real unsupervised exploration of complex biological tissues** Guerrini, Ruggero; Ogrinc , Nina; Marini, Federico; Duponchel, Ludovic.

- **Can a simple visible image guide us in our chemometric explorations?** Guerrini, Ruggero; Marini, Federico; Herreyre, Nicholas; Motto-Ros, Vincent; Duponchel, Ludovic.

- **Bridging the Gap Between Molecular and Elemental Maps in Tissue Imaging** Guerrini, Ruggero; Ogrinc , Nina; Marini, Federico; Duponchel, Ludovic.

Conferences

International:

- Fusion of LIBS and RGB imaging for enhance marble analysis: A chemometrics approach. *Ruggero Guerrini, Vincent Motto-Ros, Federico Marini, Ludovic Duponchel.* . Raleigh (USA) 20-25/10/2024. **SCIX 2024** *Invited speaker*

- Towards automatic clustering in MALDI imagin for a real unsupervised exploration of complex biological tissues. *Ruggero Guerrini, Nina Ogrinc, Ludovic Duponchel.* . Padova (Italy) 27-30/06/2023. **CCM23** - *oral*

International (ACCEPTED):

- Can a simple visible image guide us in our chemometric explorations? *Ruggero Guerrini, Federico Marini, Herreyre Nicolas, Vincent Motto-Ros, Ludovic Duponchel.* . Porquerolles (France) 09-12/09/2025. **CCM 2025** - *oral*

Nationals

- MALDI-imaging enhanced by automatic clustering for exploration of tumor heterogeneity. *Ruggero Guerrini, Nina Ogrinc and Ludovic Duponchel.* . Nantes (France) 26-08/02/2024. **CHIMIOMETRIE 2024** - *oral*

- Embedded kmeans: an improvement in MALDI-imaging segmentation on biological tissues. *Ruggero Guerrini, Nina Ogrinc and Ludovic Duponchel.*. Villeneuve D'Ascq (France), 12-13/06/2023. **GDR-MSI 2023** - *oral*

Appendix

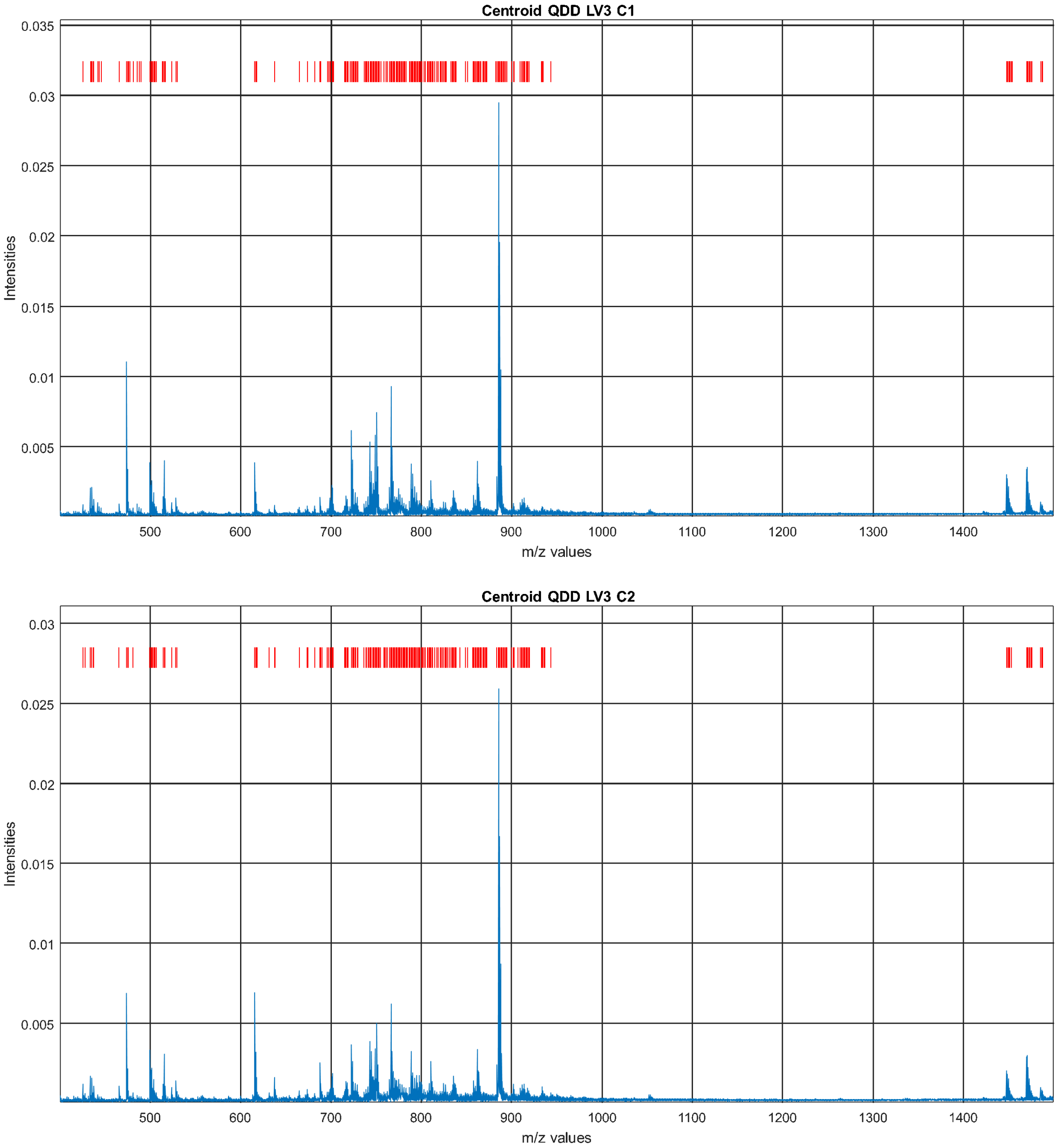
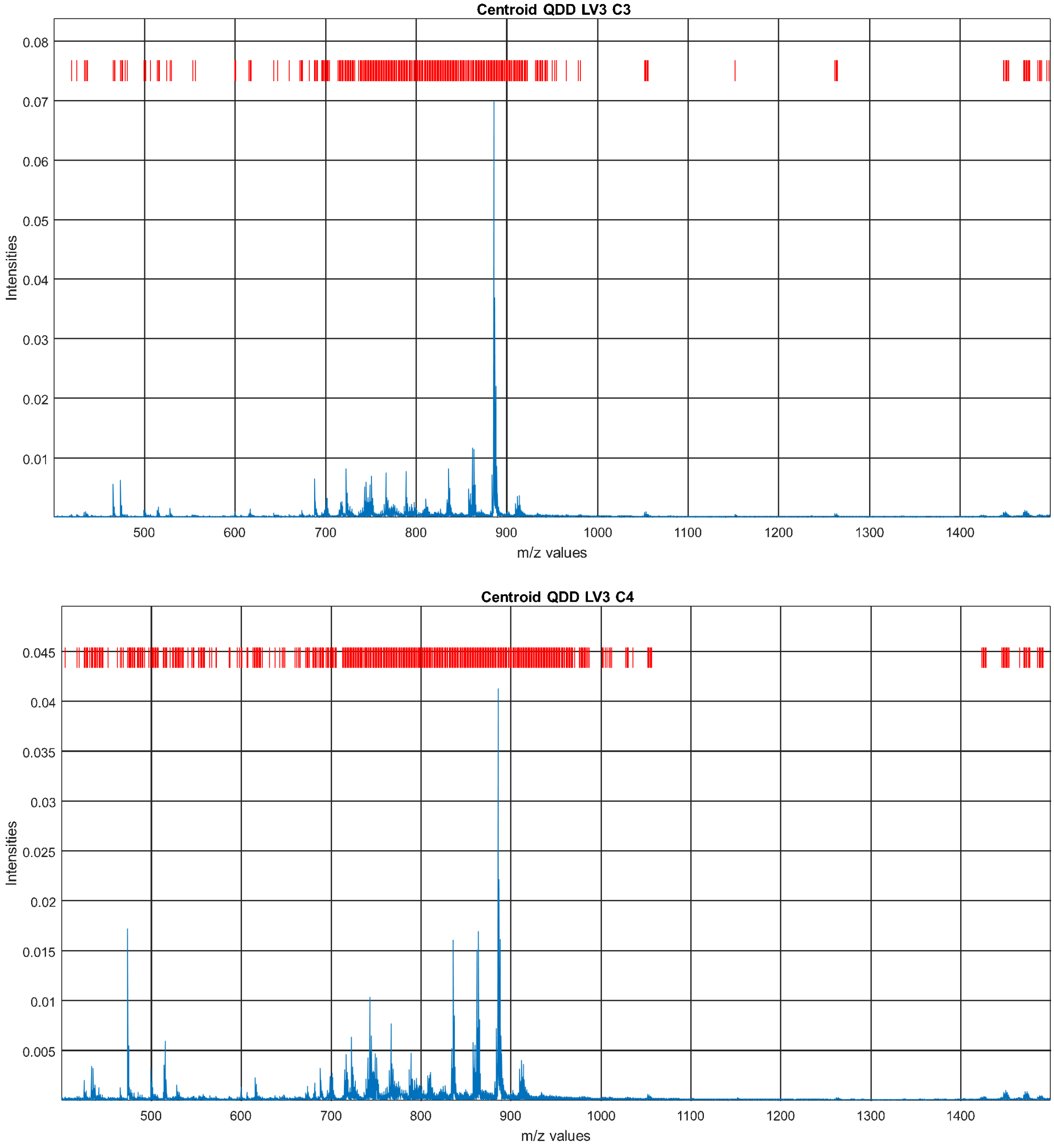


Figure 88: Spectra of the QDD_LV3_C1 and QDD_LV3_C2 centroids. The red lines indicate the positions of peaks detected with intensities greater than three times the estimated noise level around m/z = 1300.



Centroid QDD LV3 C3

Centroid QDD LV3 C4

Figure 89: Spectra of the QDD_LV3_C3 and QDD_LV3_C4 centroids. The red lines indicate the positions of peaks detected with intensities greater than three times the estimated noise level around m/z = 1300.

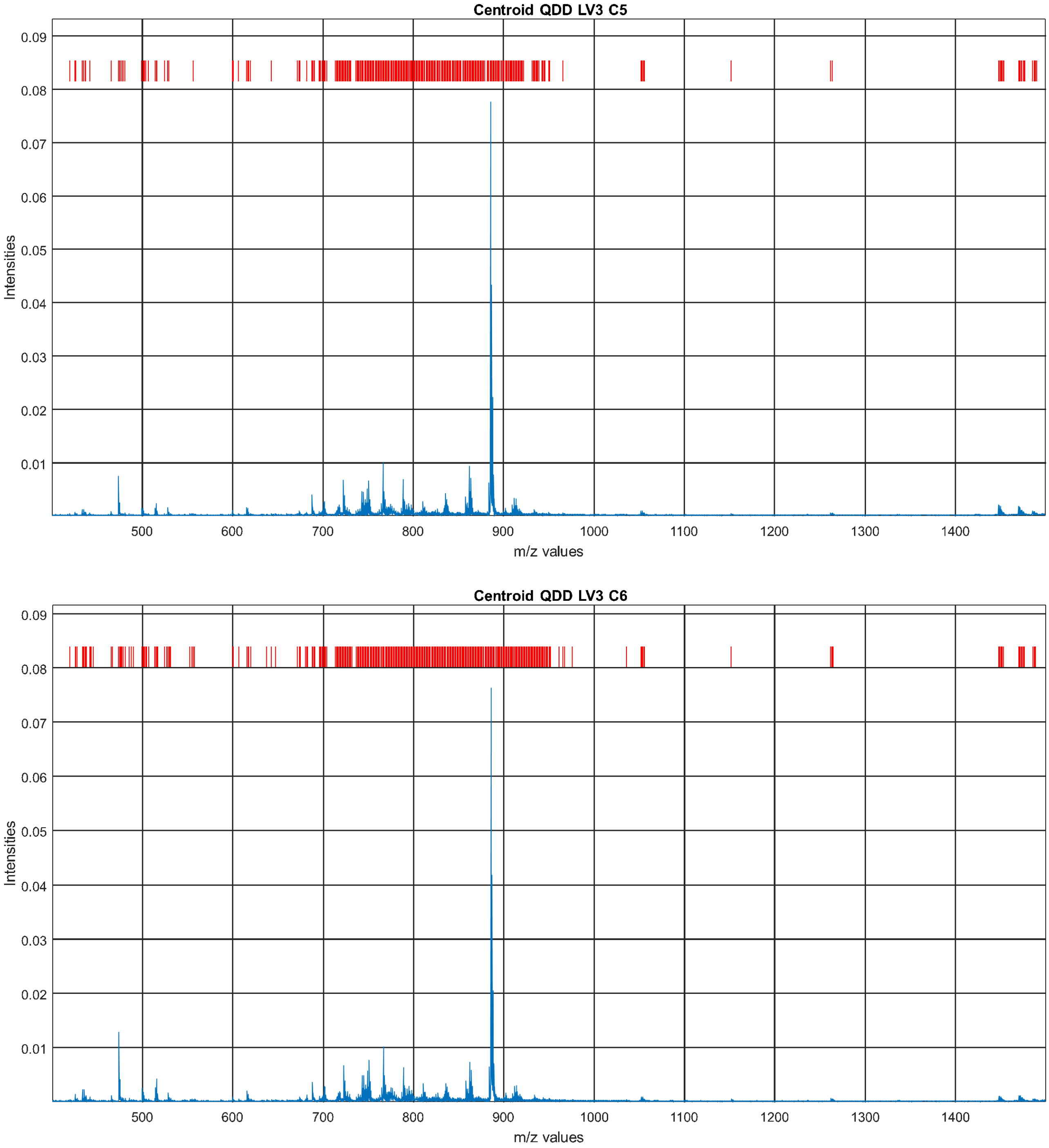


Figure 90: Spectra of the QDD_LV3_C5 and QDD_LV3_C6 centroids. The red lines indicate the positions of peaks detected with intensities greater than three times the estimated noise level around m/z = 1300.

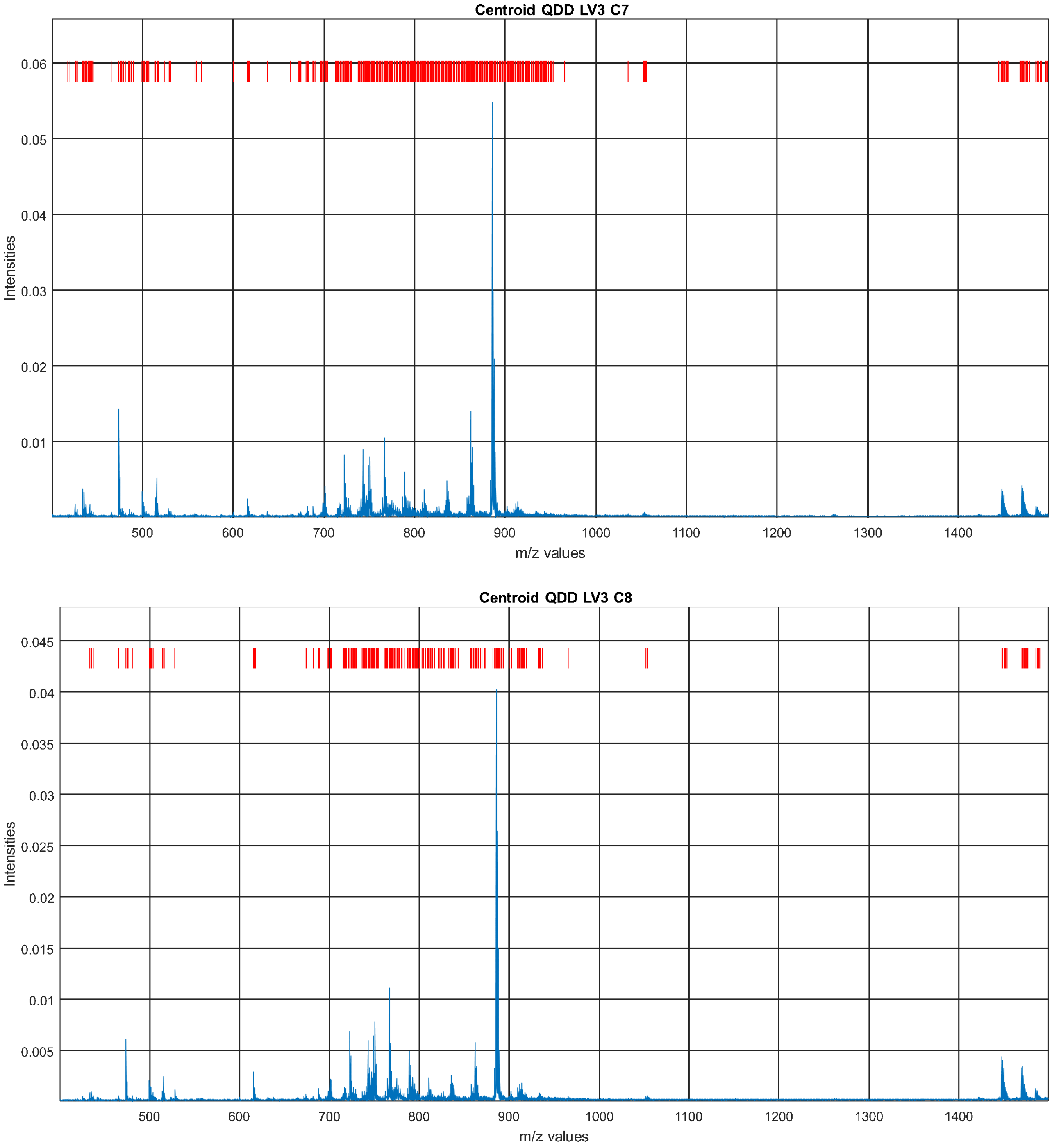


Figure 91: Spectra of the QDD_LV3_C7 and QDD_LV3_C8 centroids. The red lines indicate the positions of peaks detected with intensities greater than three times the estimated noise level around m/z = 1300.

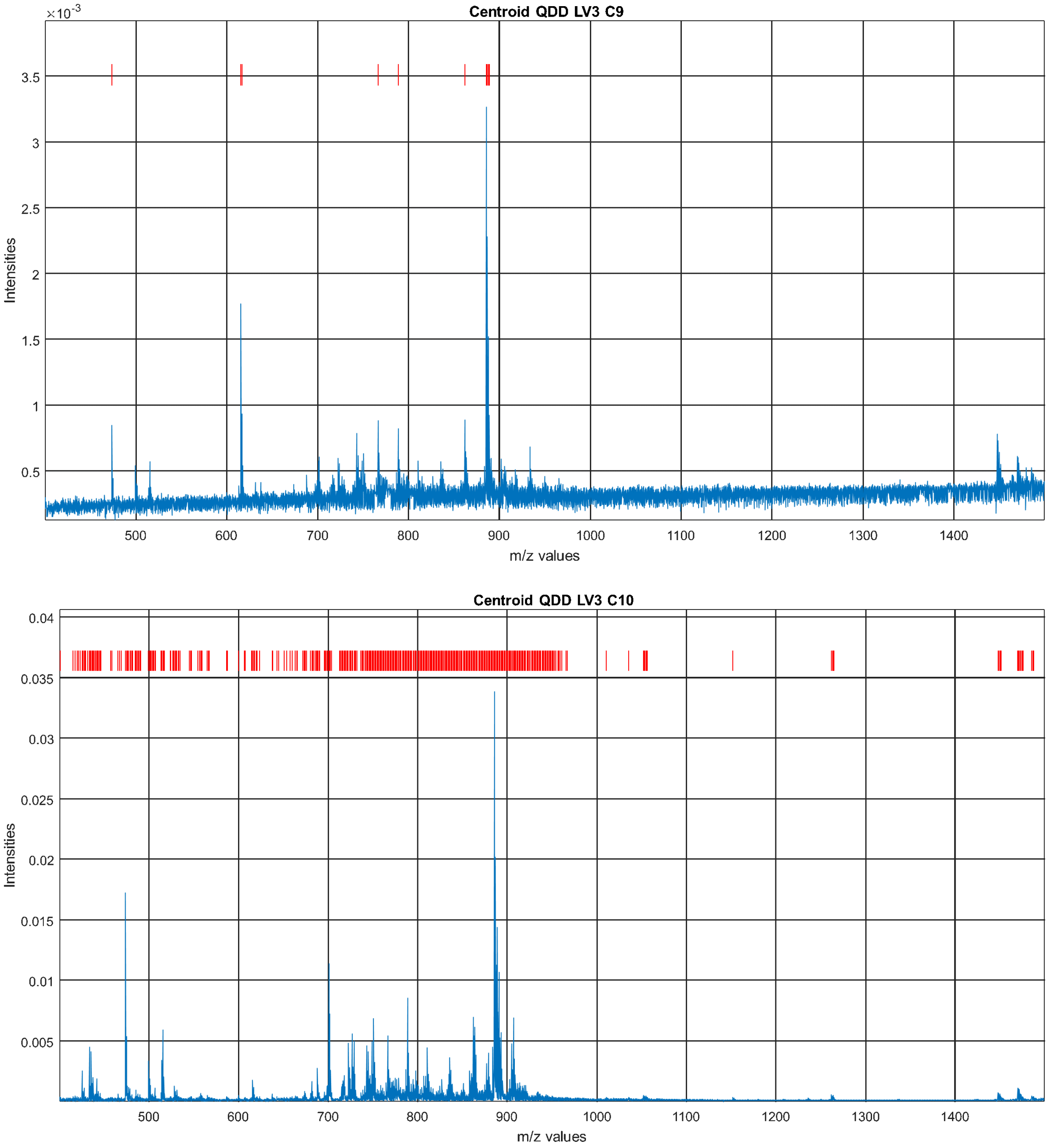


Figure 92: Spectra of the QDD_LV3_C9 and QDD_LV3_C10 centroids. The red lines indicate the positions of peaks detected with intensities greater than three times the estimated noise level around m/z = 1300.

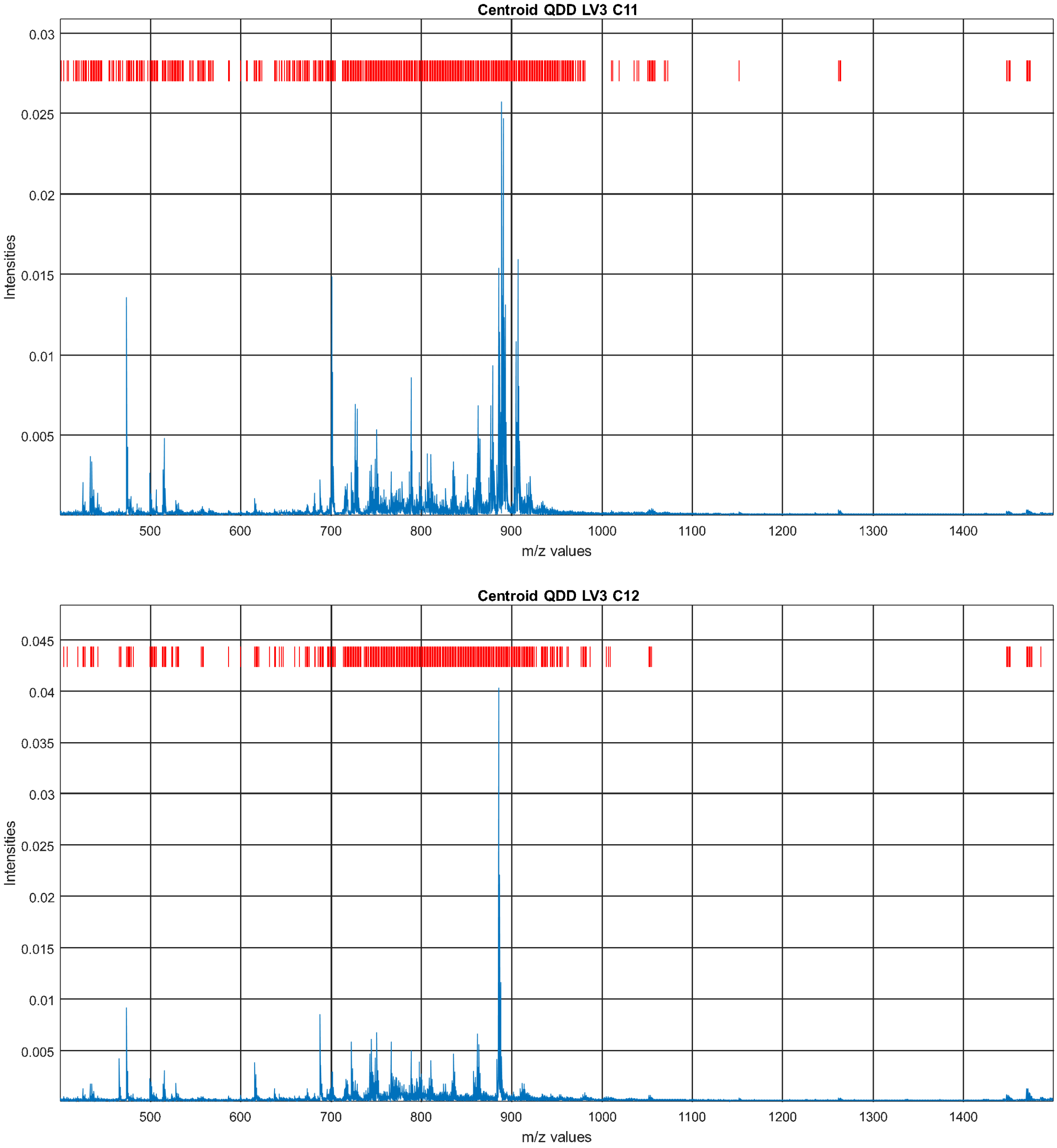


Figure 93: Spectra of the QDD_LV3_C11 and QDD_LV3_C12 centroids. The red lines indicate the positions of peaks detected with intensities greater than three times the estimated noise level around m/z = 1300.

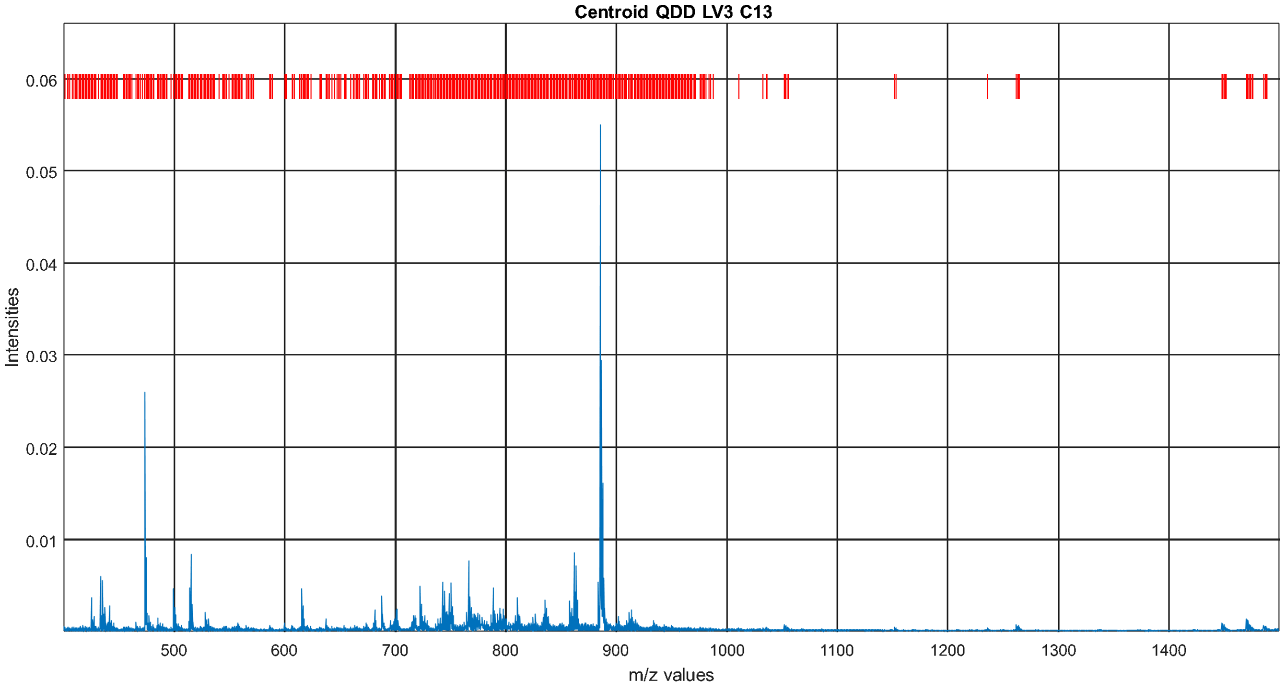


Figure 94: Spectra of the QDD_LV3_C13. The red lines indicate the positions of peaks detected with intensities greater than three times the estimated noise level around m/z = 1300.

m/z	QDD_LV3_C1	QDD_LV3_C2	QDD_LV3_C3	QDD_LV3_C4	QDD_LV3_C5	QDD_LV3_C6	QDD_LV3_C7	QDD_LV3_C8	QDD_LV3_C9	QDD_LV3_C10	QDD_LV3_C11	QDD_LV3_C12	QDD_LV3_C13	m/z
401,13										0,00036	0,00027		0,00049	401,13
403,11													0,00037	403,11
404,18				0,00036							0,00024	0,00048	0,00050	404,18
405,21													0,00035	405,21
407,20											0,00023	0,00045	0,00044	407,20
409,16											0,00025		0,00042	409,16
410,17													0,00041	410,17
411,14													0,00036	411,14
412,15													0,00039	412,15
413,16													0,00034	413,16
414,13													0,00038	414,13
415,12										0,00039	0,00029		0,00053	415,12
416,15													0,00040	416,15
417,20				0,00040			0,00039			0,00045	0,00034		0,00063	417,20
418,17											0,00024		0,00041	418,17
419,21			0,00044	0,00044	0,00044	0,00040	0,00039			0,00042	0,00030	0,00050	0,00056	419,21
420,16													0,00036	420,16
421,17										0,00034	0,00025		0,00046	421,17
422,16													0,00035	422,16
423,17										0,00035	0,00026		0,00047	423,17
424,16													0,00035	424,16
425,15	0,00089	0,00119	0,00050	0,00200	0,00068	0,00134	0,00175			0,00254	0,00204	0,00127	0,00362	425,15
426,16				0,00083	0,00041	0,00057	0,00063			0,00085	0,00066	0,00055	0,00128	426,16
427,15		0,00063		0,00094		0,00070	0,00096			0,00114	0,00086	0,00073	0,00161	427,15
428,16				0,00035						0,00038	0,00029		0,00054	428,16
429,17				0,00033						0,00039	0,00030		0,00052	429,17
431,17				0,00033						0,00036	0,00026		0,00050	431,17
433,14	0,00205	0,00171	0,00085	0,00342	0,00113	0,00225	0,00371	0,00089		0,00450	0,00367	0,00177	0,00598	433,14
434,15	0,00057			0,00102		0,00068	0,00105			0,00131	0,00106	0,00055	0,00174	434,15
435,16	0,00209	0,00159	0,00092	0,00322	0,00115	0,00223	0,00326	0,00097		0,00415	0,00334	0,00174	0,00548	435,16
436,17	0,00081	0,00066	0,00065	0,00124	0,00063	0,00093	0,00126			0,00150	0,00120	0,00079	0,00186	436,17
437,14	0,00124	0,00107	0,00053	0,00153	0,00063	0,00115	0,00174	0,00057		0,00201	0,00161	0,00105	0,00259	437,14
438,15				0,00052		0,00042	0,00054			0,00068	0,00053		0,00088	438,15
439,16				0,00046			0,00050			0,00059	0,00046		0,00078	439,16
440,17				0,00047			0,00050			0,00066	0,00045		0,00100	440,17
441,18	0,00103			0,00128	0,00049	0,00089	0,00167			0,00187	0,00138	0,00055	0,00274	441,18
442,17				0,00049		0,00039	0,00056			0,00064	0,00047		0,00094	442,17
443,18	0,00074			0,00064		0,00046	0,00078			0,00087	0,00067		0,00121	443,18
444,15				0,00032			0,00038			0,00042	0,00032		0,00059	444,15
445,17	0,00062			0,00058		0,00041	0,00064			0,00067	0,00052		0,00094	445,17
446,14				0,00031						0,00036	0,00027		0,00050	446,14
447,17													0,00037	447,17
448,12													0,00033	448,12
452,22				0,00033										452,22
454,14											0,00024		0,00043	454,14
455,15											0,00023		0,00044	455,15
456,14													0,00043	456,14
457,16										0,00034	0,00025		0,00047	457,16
458,15										0,00035	0,00027		0,00051	458,15
459,16											0,00022		0,00040	459,16
460,15													0,00035	460,15
461,16													0,00038	461,16
462,15				0,00029							0,00029			462,15
464,28											0,00029			464,28
465,42	0,00091	0,00107	0,00551	0,00129	0,00085	0,00117	0,00062	0,00058		0,00060	0,00042	0,00424	0,00103	465,42
466,34			0,00178	0,00047		0,00045				0,00033	0,00024	0,00141	0,00053	466,34
467,25			0,00066									0,00060	0,00038	467,25
469,12				0,00030						0,00037	0,00029		0,00053	469,12
471,15													0,00034	471,15
472,26													0,00043	472,26
473,42	0,01104	0,00688	0,00621	0,01722	0,00741	0,01283	0,01430	0,00610	0,00085	0,01723	0,01354	0,00917	0,02600	473,42
474,41	0,00341	0,00218	0,00196	0,00548	0,00242	0,00409	0,00525	0,00192		0,00539	0,00424	0,00297	0,00800	474,41
475,30	0,00105	0,00079	0,00050	0,00127	0,00061	0,00096	0,00108	0,00059		0,00130	0,00102	0,00081	0,00193	475,30
476,25				0,00077		0,00054	0,00062			0,00066	0,00053	0,00051	0,00093	476,25
477,17	0,00060			0,00111	0,00042	0,00072	0,00122			0,00120	0,00101	0,00052	0,00173	477,17
478,19			0,00046	0,00079	0,00044	0,00060	0,00071			0,00106	0,00116	0,00062	0,00098	478,19
479,18				0,00038						0,00046	0,00044		0,00053	479,18
480,31	0,00063	0,00067	0,00054	0,00075	0,00056	0,00060	0,00058	0,00059		0,00060	0,00050	0,00077	0,00076	480,31
481,22				0,00037						0,00032	0,00025		0,00043	481,22
484,19				0,00036			0,00039			0,00041	0,00030		0,00064	484,19
485,16	0,00091			0,00082		0,00066	0,00104			0,00095	0,00072		0,00145	485,16
486,17				0,00037			0,00042			0,00045	0,00035		0,00071	486,17
487,16	0,00071			0,00052		0,00042	0,00068			0,00062	0,00048		0,00093	487,16
488,17				0,00031						0,00036	0,00028		0,00054	488,17
489,18	0,00062			0,00054		0,00043	0,00054			0,00058	0,00046		0,00084	489,18
490,20				0,00032						0,00034	0,00027		0,00047	490,20
491,18											0,00023		0,00042	491,18
492,22				0,00029							0,00025		0,00043	492,22
493,27													0,00036	493,27
497,23				0,00032							0,00025		0,00045	497,23
499,28	0,00386	0,00335	0,00112	0,00322	0,00141	0,00255	0,00338	0,00209		0,00337	0,00265	0,00229	0,00464	499,28
500,29	0,00228	0,00208	0,00132	0,00192	0,00138	0,00182	0,00199	0,00147		0,00188	0,00136	0,00191	0,00261	500,29
501,28	0,00259	0,00219	0,00075	0,00129	0,00087	0,00119	0,00143	0,00144		0,00134	0,00098	0,00146	0,00181	501,28
502,29	0,00109	0,00096		0,00060	0,00042	0,00053	0,00065	0,00063		0,00060	0,00045	0,00071	0,00081	502,29
503,26	0,00173	0,00151		0,00062	0,00044	0,00057	0,00073	0,00092		0,00064	0,00049	0,00089	0,00084	503,26
504,29	0,00093	0,00079		0,00064		0,00041	0,00050			0,00044	0,00035	0,00061	0,00059	504,29
505,28	0,00073	0,00065		0,00043			0,00040			0,00037	0,00028	0,00051	0,00051	505,28
506,34	0,00072	0,00068	0,00053	0,00054	0,00048	0,00053	0,00048			0,00111	0,00162	0,00077	0,00066	506,34
507,30				0,00033						0,00047	0,00056		0,00041	507,30
508,32											0,00025			508,32
513,27	0,00057			0,00081		0,00064	0,00071			0,00076	0,00064	0,00045	0,00108	513,27
514,28	0,00143	0,00121	0,00119	0,00355	0,00139	0,00261	0,00260	0,00102		0,00340	0,00284	0,00174	0,00476	514,28
515,29	0,00399	0,00308	0,00179	0,00593	0,00230	0,00423	0,00518	0,00244		0,00594	0,00480	0,00306	0,00838	515,29
516,28	0,00131	0,00110	0,00072	0,00206	0,00087	0,00158	0,00178	0,00085		0,00207	0,00169	0,00120	0,00296	516,28
517,23				0,00067		0,00053	0,00058			0,00068	0,00054	0,00048	0,00096	517,23
518,24													0,00035	518,24
519,15											0,00023		0,00043	519,15
520,28				0,00029							0,00026		0,00040	520,28
521,25													0,00035	521,25
522,27											0,00026			522,27
m/z	QDD_LV3_C1	QDD_LV3_C2	QDD_LV3_C3	QDD_LV3_C4	QDD_LV3_C5	QDD_LV3_C6	QDD_LV3_C7	QDD_LV3_C8	QDD_LV3_C9	QDD_LV3_C10	QDD_LV3_C11	QDD_LV3_C12	QDD_LV3_C13	m/z

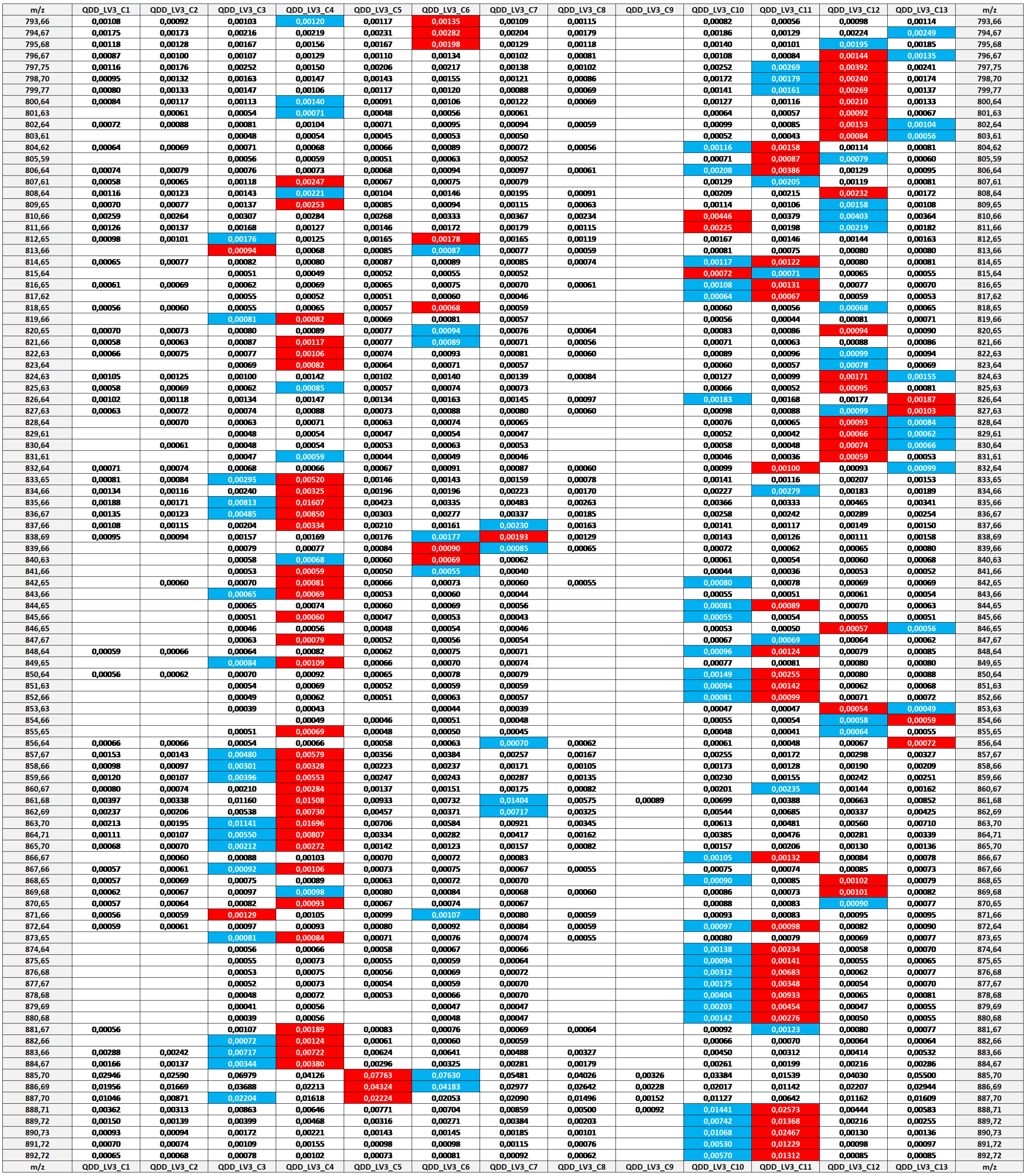
Figure 95: List of pointed peaks for each of the 13 centroids (m/z range: 401.13 – 522.27) associated with Figures S3-S9. A value in a cell corresponds to the observed intensity at a given m/z value for a specific cluster. The cell is red if the cluster in question shows the highest intensity among the 13 clusters for the given m/z value. On the same row, a blue cell indicates the second highest intensity among the 13 clusters. An empty cell means that no peak was detected for a particular cluster at this specific m/z value.

m/z	QDD_LV3_C1	QDD_LV3_C2	QDD_LV3_C3	QDD_LV3_C4	QDD_LV3_C5	QDD_LV3_C6	QDD_LV3_C7	QDD_LV3_C8	QDD_LV3_C9	QDD_LV3_C10	QDD_LV3_C11	QDD_LV3_C12	QDD_LV3_C13	m/z
523,28	0,00100	0,00099		0,00033			0,00041			0,00043	0,00033	0,00076	0,00052	523,28
524,31			0,00044	0,00039	0,00043	0,00047				0,00039	0,00031	0,00056	0,00051	524,31
525,24											0,00023		0,00038	525,24
526,33				0,00042		0,00041				0,00033	0,00025		0,00047	526,33
527,30				0,00036						0,00034	0,00027		0,00047	527,30
528,33	0,00136	0,00143	0,00152	0,00154	0,00149	0,00164	0,00122	0,00115		0,00129	0,00094	0,00182	0,00203	528,33
529,32	0,00076	0,00073	0,00061	0,00090	0,00063	0,00083	0,00078			0,00085	0,00066	0,00083	0,00125	529,32
530,27				0,00057		0,00046	0,00044			0,00050	0,00040	0,00049	0,00072	530,27
531,30				0,00079		0,00063	0,00072			0,00098	0,00078	0,00074	0,00135	531,30
532,29				0,00037						0,00044	0,00036		0,00058	532,29
533,29				0,00035						0,00043	0,00033		0,00057	533,29
534,28				0,00029						0,00037	0,00037		0,00042	534,28
535,25				0,00030							0,00026		0,00041	535,25
536,28											0,00027		0,00034	536,28
540,20				0,00033									0,00033	540,20
543,27											0,00022		0,00038	543,27
544,30													0,00035	544,30
545,28				0,00038						0,00038	0,00030		0,00053	545,28
546,27				0,00043						0,00043	0,00035		0,00058	546,27
547,26				0,00040						0,00039	0,00031		0,00053	547,26
549,26													0,00035	549,26
552,33				0,00038		0,00045					0,00025		0,00049	552,33
553,30			0,00039	0,00047							0,00024		0,00046	553,30
554,33				0,00040		0,00048				0,00035	0,00031		0,00054	554,33
555,33				0,00030									0,00039	555,33
556,36			0,00041	0,00035	0,00043	0,00047				0,00041	0,00038	0,00052	0,00056	556,36
557,26				0,00060		0,00047	0,00054			0,00070	0,00056	0,00046	0,00095	557,26
558,26				0,00044						0,00051	0,00040	0,00048	0,00071	558,26
559,25				0,00041			0,00040			0,00044	0,00034		0,00059	559,25
560,26											0,00023		0,00038	560,26
561,25													0,00033	561,25
564,28											0,00028			564,28
565,23				0,00047			0,00041			0,00053	0,00044		0,00067	565,23
566,26										0,00033	0,00033		0,00035	566,26
567,25				0,00031						0,00038	0,00031		0,00047	567,25
568,26											0,00031			568,26
569,25											0,00026		0,00037	569,25
570,27													0,00042	570,27
571,30				0,00045									0,00036	571,30
572,29				0,00030										572,29
586,28				0,00039						0,00043	0,00032	0,00056	0,00060	586,28
587,29				0,00036						0,00036	0,00027		0,00050	587,29
588,26													0,00037	588,26
595,36				0,00033										595,36
597,34				0,00038										597,34
599,39			0,00099	0,00133	0,00104	0,00096	0,00067			0,00055	0,00032	0,00061	0,00094	599,39
600,36			0,00040	0,00052	0,00043	0,00043							0,00047	600,36
601,28													0,00033	601,28
606,23				0,00079	0,00045	0,00052				0,00033	0,00025		0,00062	606,23
607,27				0,00041						0,00038	0,00028		0,00059	607,27
608,30													0,00038	608,30
613,31				0,00031									0,00051	613,31
614,32				0,00036						0,00036	0,00028		0,00055	614,32
615,29	0,00385	0,00691	0,00063	0,00228	0,00161	0,00196	0,00240	0,00291	0,00177	0,00176	0,00106	0,00383	0,00464	615,29
616,30	0,00178	0,00322	0,00135	0,00162	0,00139	0,00138	0,00143	0,00137	0,00094	0,00118	0,00074	0,00268	0,00279	616,30
617,28	0,00093	0,00161	0,00060	0,00082	0,00067	0,00068	0,00073	0,00072		0,00058	0,00037	0,00134	0,00133	617,28
618,31		0,00072		0,00039								0,00062	0,00059	618,31
619,32				0,00048	0,00044	0,00048				0,00034	0,00024	0,00055	0,00058	619,32
620,35				0,00036						0,00040	0,00031		0,00063	620,35
621,38				0,00035							0,00023		0,00049	621,38
623,30				0,00035						0,00041	0,00031		0,00053	623,30
631,29		0,00067		0,00029								0,00046	0,00050	631,29
632,30													0,00039	632,30
633,27													0,00037	633,27
637,29	0,00083	0,00164		0,00047		0,00046	0,00074			0,00061	0,00038	0,00131	0,00132	637,29
638,28		0,00088					0,00040			0,00036	0,00026	0,00073	0,00068	638,28
639,31											0,00022		0,00049	639,31
640,31													0,00033	640,31
642,50			0,00067	0,00041	0,00046	0,00040				0,00040	0,00032	0,00076	0,00051	642,50
644,48										0,00033	0,00035	0,00047	0,00033	644,48
645,41				0,00043							0,00023			645,41
647,30			0,00042	0,00054		0,00039						0,00045	0,00045	647,30
648,31				0,00033							0,00030		0,00036	648,31
649,30													0,00035	649,30
650,36										0,00034	0,00037		0,00039	650,36
652,29											0,00024			652,29
653,28										0,00034	0,00028		0,00062	653,28
654,28											0,00023		0,00041	654,28
655,35													0,00035	655,35
657,42										0,00038	0,00039			657,42
658,43											0,00026			658,43
659,50			0,00048	0,00040						0,00039	0,00035	0,00060	0,00045	659,50
661,50				0,00030							0,00023		0,00038	661,50
663,44				0,00035			0,00043			0,00047	0,00033		0,00060	663,44
664,45	0,00069	0,00079		0,00037						0,00039	0,00030	0,00055	0,00049	664,45
665,36				0,00039						0,00039	0,00030		0,00053	665,36
666,37											0,00030		0,00037	666,37
667,38													0,00034	667,38
668,39											0,00022			668,39
670,42											0,00024			670,42
671,51			0,00047	0,00073	0,00044	0,00042	0,00057			0,00039	0,00027	0,00057	0,00051	671,51
672,52			0,00043	0,00063			0,00045			0,00050	0,00046	0,00063	0,00043	672,52
673,53	0,00077	0,00086	0,00122	0,00140	0,00091	0,00086	0,00073	0,00071		0,00082	0,00069	0,00130	0,00094	673,53
674,53		0,00059	0,00077	0,00084	0,00062	0,00063	0,00066	0,00055		0,00067	0,00059	0,00089	0,00072	674,53
675,49				0,00039						0,00035	0,00029	0,00047	0,00042	675,49
679,41				0,00030									0,00039	679,41
680,36				0,00071		0,00055	0,00054			0,00069	0,00057		0,00092	680,36
681,35	0,00078	0,00064	0,00055	0,00174	0,00068	0,00124	0,00142	0,00056		0,00169	0,00138	0,00081	0,00237	681,35
682,36				0,00086		0,00064	0,00072			0,00085	0,00072	0,00048	0,00117	682,36
683,38				0,00038						0,00039	0,00031		0,00049	683,38
m/z	QDD_LV3_C1	QDD_LV3_C2	QDD_LV3_C3	QDD_LV3_C4	QDD_LV3_C5	QDD_LV3_C6	QDD_LV3_C7	QDD_LV3_C8	QDD_LV3_C9	QDD_LV3_C10	QDD_LV3_C11	QDD_LV3_C12	QDD_LV3_C13	m/z

Figure 96: List of pointed peaks for each of the 13 centroids (m/z range: 523.28 – 683.38) associated with Figures S3-S9. A value in a cell corresponds to the observed intensity at a given m/z value for a specific cluster. The cell is red if the cluster in question shows the highest intensity among the 13 clusters for the given m/z value. On the same row, a blue cell indicates the second highest intensity among the 13 clusters. An empty cell means that no peak was detected for a particular cluster at this specific m/z value.

m/z	QDD_LV3_C1	QDD_LV3_C2	QDD_LV3_C3	QDD_LV3_C4	QDD_LV3_C5	QDD_LV3_C6	QDD_LV3_C7	QDD_LV3_C8	QDD_LV3_C9	QDD_LV3_C10	QDD_LV3_C11	QDD_LV3_C12	QDD_LV3_C13	m/z
685,50				0,00042						0,00039	0,00037	0,00055	0,00039	685,50
686,53				0,00034						0,00035	0,00035			686,53
687,69	0,00139	0,00255	0,00652	0,00321	0,00395	0,00357	0,00148	0,00127		0,00277	0,00222	0,00851	0,00385	687,69
688,64	0,00079	0,00124	0,00275	0,00199	0,00170	0,00169	0,00085	0,00072		0,00137	0,00108	0,00354	0,00190	688,64
689,67		0,00073	0,00135	0,00097	0,00087	0,00083	0,00045			0,00074	0,00058	0,00175	0,00091	689,67
690,58			0,00068	0,00070	0,00051	0,00050				0,00042	0,00031	0,00083	0,00055	690,58
691,57				0,00033								0,00046		691,57
694,40				0,00034							0,00025		0,00039	694,40
695,43	0,00059	0,00072	0,00081	0,00100	0,00076	0,00100	0,00068			0,00079	0,00061	0,00121	0,00120	695,43
696,42			0,00047	0,00056	0,00046	0,00056	0,00045			0,00048	0,00038	0,00067	0,00067	696,42
697,49	0,00059	0,00069	0,00060	0,00076	0,00066	0,00063	0,00064	0,00058		0,00052	0,00034	0,00073	0,00074	697,49
698,59	0,00135	0,00088	0,00115	0,00243	0,00086	0,00091	0,00189	0,00107		0,00131	0,00111	0,00124	0,00105	698,59
699,62	0,00118	0,00095	0,00140	0,00235	0,00110	0,00111	0,00180	0,00107		0,00132	0,00108	0,00137	0,00133	699,62
700,65	0,00228	0,00165	0,00323	0,00277	0,00237	0,00284	0,00412	0,00222		0,01142	0,01482	0,00291	0,00229	700,65
701,64	0,00208	0,00189	0,00318	0,00230	0,00263	0,00276	0,00316	0,00215		0,00724	0,00890	0,00294	0,00246	701,64
702,65	0,00094	0,00094	0,00148	0,00103	0,00124	0,00129	0,00132	0,00100		0,00259	0,00306	0,00145	0,00118	702,65
703,60			0,00064	0,00052	0,00055	0,00054	0,00046			0,00074	0,00078	0,00066	0,00051	703,60
704,57				0,00038							0,00028	0,00048	0,00036	704,57
705,52				0,00030									0,00033	705,52
712,56				0,00048			0,00038			0,00037	0,00035		0,00035	712,56
713,49			0,00044	0,00049	0,00051	0,00045	0,00042			0,00044	0,00040	0,00054	0,00049	713,49
714,60	0,00080	0,00073	0,00128	0,00306	0,00094	0,00093	0,00115	0,00077		0,00123	0,00121	0,00119	0,00099	714,60
715,63	0,00077	0,00085	0,00118	0,00204	0,00107	0,00098	0,00115	0,00077		0,00147	0,00179	0,00138	0,00108	715,63
716,65	0,00150	0,00137	0,00248	0,00465	0,00168	0,00170	0,00192	0,00141		0,00174	0,00158	0,00220	0,00175	716,65
717,59	0,00125	0,00128	0,00268	0,00279	0,00207	0,00186	0,00171	0,00129		0,00216	0,00199	0,00209	0,00167	717,59
718,61	0,00076	0,00089	0,00188	0,00209	0,00159	0,00164	0,00089	0,00082		0,00125	0,00099	0,00170	0,00148	718,61
719,62			0,00071	0,00085	0,00065	0,00071				0,00044	0,00030	0,00075	0,00067	719,62
720,59	0,00072		0,00076	0,00131	0,00063	0,00062	0,00098	0,00075		0,00050	0,00030	0,00067	0,00059	720,59
721,56			0,00055	0,00082	0,00051	0,00055	0,00062			0,00040	0,00024	0,00054	0,00053	721,56
722,63	0,00613	0,00366	0,00816	0,00634	0,00668	0,00664	0,00825	0,00687		0,00486	0,00266	0,00585	0,00486	722,63
723,64	0,00406	0,00263	0,00408	0,00333	0,00379	0,00382	0,00448	0,00446		0,00279	0,00155	0,00322	0,00297	723,64
724,63	0,00198	0,00133	0,00228	0,00255	0,00184	0,00188	0,00260	0,00199		0,00187	0,00141	0,00191	0,00162	724,63
725,60	0,00090	0,00079	0,00119	0,00133	0,00106	0,00109	0,00130	0,00094		0,00102	0,00078	0,00104	0,00110	725,60
726,65	0,00171	0,00127	0,00186	0,00196	0,00150	0,00188	0,00248	0,00143		0,00559	0,00691	0,00207	0,00173	726,65
727,64	0,00099	0,00078	0,00093	0,00094	0,00080	0,00096	0,00123	0,00085		0,00288	0,00341	0,00107	0,00087	727,64
728,66	0,00138	0,00115	0,00135	0,00110	0,00116	0,00142	0,00165	0,00121		0,00500	0,00660	0,00171	0,00114	728,66
729,65	0,00079	0,00072	0,00070	0,00063	0,00063	0,00074	0,00085	0,00069		0,00225	0,00302	0,00095	0,00064	729,65
730,62			0,00059	0,00061	0,00050	0,00057	0,00051			0,00093	0,00111	0,00072	0,00053	730,62
731,59			0,00041	0,00044		0,00038				0,00043	0,00042	0,00060	0,00039	731,59
732,57				0,00043							0,00026	0,00047	0,00035	732,57
733,55				0,00030										733,55
734,53				0,00033							0,00024		0,00033	734,53
735,50													0,00033	735,50
736,62	0,00094	0,00071	0,00090	0,00079	0,00089	0,00095	0,00108	0,00099		0,00075	0,00045	0,00088	0,00079	736,62
737,61	0,00064		0,00059	0,00054	0,00061	0,00064	0,00069	0,00065		0,00049	0,00031	0,00067	0,00057	737,61
738,62	0,00111	0,00088	0,00116	0,00213	0,00113	0,00140	0,00138	0,00106		0,00097	0,00059	0,00106	0,00132	738,62
739,57	0,00068	0,00064	0,00072	0,00115	0,00075	0,00088	0,00083	0,00066		0,00058	0,00035	0,00068	0,00084	739,57
740,63	0,00144	0,00103	0,00165	0,00430	0,00130	0,00133	0,00241	0,00151		0,00127	0,00090	0,00134	0,00147	740,63
741,62	0,00084	0,00067	0,00094	0,00208	0,00081	0,00079	0,00134	0,00088		0,00075	0,00054	0,00082	0,00086	741,62
742,67	0,00535	0,00387	0,00514	0,01033	0,00454	0,00480	0,00897	0,00597		0,00463	0,00278	0,00469	0,00539	742,67
743,66	0,00246	0,00206	0,00274	0,00489	0,00243	0,00249	0,00426	0,00273		0,00252	0,00176	0,00283	0,00281	743,66
744,67	0,00323	0,00324	0,00587	0,00651	0,00449	0,00480	0,00435	0,00329		0,00410	0,00312	0,00608	0,00437	744,67
745,66	0,00164	0,00165	0,00257	0,00285	0,00199	0,00222	0,00193	0,00163		0,00189	0,00150	0,00288	0,00204	745,66
746,63	0,00242	0,00168	0,00330	0,00294	0,00309	0,00316	0,00286	0,00294		0,00218	0,00174	0,00260	0,00218	746,63
747,64	0,00193	0,00147	0,00257	0,00282	0,00244	0,00235	0,00223	0,00237		0,00144	0,00103	0,00186	0,00180	747,64
748,65	0,00582	0,00341	0,00545	0,00468	0,00503	0,00577	0,00688	0,00644		0,00495	0,00352	0,00427	0,00415	748,65
749,64	0,00251	0,00153	0,00265	0,00215	0,00239	0,00281	0,00337	0,00277		0,00228	0,00166	0,00200	0,00210	749,64
750,67	0,00743	0,00502	0,00697	0,00428	0,00661	0,00764	0,00800	0,00780		0,00685	0,00534	0,00675	0,00529	750,67
751,65	0,00358	0,00243	0,00323	0,00207	0,00309	0,00368	0,00374	0,00369		0,00326	0,00261	0,00306	0,00272	751,65
752,64	0,00159	0,00134	0,00195	0,00161	0,00174	0,00203	0,00187	0,00157		0,00192	0,00178	0,00207	0,00170	752,64
753,60	0,00069	0,00068	0,00081	0,00073	0,00078	0,00078	0,00069	0,00068		0,00074	0,00068	0,00084	0,00073	753,60
754,60	0,00061	0,00062	0,00070	0,00072	0,00066	0,00071	0,00072	0,00058		0,00104	0,00122	0,00080	0,00069	754,60
755,58			0,00049	0,00051	0,00047	0,00047	0,00046			0,00063	0,00071	0,00060	0,00049	755,58
756,64			0,00056	0,00075	0,00051	0,00058	0,00067			0,00091	0,00108	0,00075	0,00063	756,64
757,63			0,00047	0,00057	0,00041	0,00045	0,00050			0,00071	0,00082	0,00074	0,00053	757,63
758,62	0,00059	0,00069	0,00063	0,00091	0,00055	0,00066	0,00082			0,00132	0,00161	0,00101	0,00076	758,62
759,61		0,00059	0,00050	0,00058	0,00043	0,00051	0,00057			0,00089	0,00098	0,00097	0,00062	759,61
760,62	0,00066	0,00061	0,00098	0,00123	0,00074	0,00082	0,00085	0,00063		0,00110	0,00115	0,00097	0,00076	760,62
761,61			0,00055	0,00063	0,00045	0,00048	0,00046			0,00056	0,00054	0,00059	0,00046	761,61
762,62	0,00094	0,00074	0,00121	0,00137	0,00113	0,00128	0,00101	0,00102		0,00085	0,00064	0,00093	0,00103	762,62
763,61			0,00066	0,00073	0,00066	0,00072	0,00057	0,00055		0,00047	0,00033	0,00055	0,00063	763,61
764,65	0,00212	0,00150	0,00216	0,00304	0,00228	0,00261	0,00258	0,00226		0,00154	0,00098	0,00165	0,00207	764,65
765,62	0,00125	0,00093	0,00101	0,00149	0,00116	0,00134	0,00130	0,00131		0,00083	0,00053	0,00088	0,00112	765,62
766,67	0,00926	0,00620	0,00748	0,00769	0,00993	0,01019	0,01048	0,01110	0,00088	0,00541	0,00273	0,00586	0,00762	766,67
767,66	0,00496	0,00323	0,00337	0,00366	0,00457	0,00485	0,00525	0,00574		0,00264	0,00138	0,00276	0,00378	767,66
768,67	0,00253	0,00200	0,00287	0,00316	0,00305	0,00313	0,00280	0,00301		0,00192	0,00116	0,00249	0,00255	768,67
769,66	0,00100	0,00097	0,00140	0,00154	0,00131	0,00145	0,00132	0,00110		0,00111	0,00081	0,00150	0,00141	769,66
770,67	0,00147	0,00160	0,00162	0,00195	0,00163	0,00180	0,00182	0,00146		0,00163	0,00122	0,00215	0,00192	770,67
771,68	0,00124	0,00146	0,00155	0,00140	0,00153	0,00172	0,00158	0,00123		0,00159	0,00137	0,00236	0,00187	771,68
772,67	0,00142	0,00144	0,00189	0,00171	0,00171	0,00184	0,00172	0,00144		0,00173	0,00156	0,00234	0,00167	772,67
773,66	0,00123	0,00105	0,00159	0,00133	0,00157	0,00171	0,00156	0,00141		0,00127	0,00094	0,00153	0,00149	773,66
774,65	0,00203	0,00148	0,00216	0,00187	0,00221	0,00267	0,00226	0,00225		0,00194	0,00163	0,00205	0,00196	774,65
775,64	0,00109	0,00087	0,00160	0,00125	0,00133	0,00163	0,00120	0,00119		0,00105	0,00082	0,00136	0,00121	775,64
776,66	0,00161	0,00122	0,00201	0,00143	0,00184	0,00251	0,00186	0,00165		0,00181	0,00167	0,00184	0,00175	776,66
777,63	0,00082	0,00066	0,00094	0,00068	0,00089	0,00117	0,00091	0,00082		0,00090	0,00082	0,00088	0,00084	777,63
778,64	0,00134	0,00117	0,00147	0,00108	0,00146	0,00195	0,00161	0,00129		0,00192	0,00208	0,00156	0,00153	778,64
779,63	0,00071	0,00067	0,00077	0,00059	0,00076	0,00094	0,00076	0,00068		0,00094	0,00101	0,00084	0,00077	779,63
780,62	0,00098	0,00099	0,00101	0,00101	0,00090	0,00121	0,00126	0,00082		0,00120	0,00104	0,00166	0,00119	780,62
781,61	0,00059	0,00066	0,00061	0,00066	0,00057	0,00070	0,00071			0,00066	0,00052	0,00098	0,00075	781,61
782,62	0,00087	0,00095	0,00078	0,00100	0,00079	0,00096	0,00107	0,00077		0,00100	0,00082	0,00125	0,00106	782,62
783,63		0,00065	0,00059	0,00066	0,00055	0,00064	0,00063			0,00072	0,00067	0,00095	0,00072	783,63
784,64		0,00062	0,00061	0,00079	0,00054	0,00064	0,00076			0,00102	0,00109	0,00090	0,00075	784,64
785,67			0,00048	0,00052	0,00043	0,00048	0,00050			0,00082	0,00099	0,00085	0,00058	785,67
786,66	0,00114	0,00101	0,00188	0,00312	0,00148	0,00176	0,00264	0,00128		0,00265	0,00272	0,00161	0,00182	786,66
787,66	0,00064	0,00060	0,00094	0,00146	0,00078	0,00090	0,00130	0,00072		0,00128	0,00131	0,00083	0,00091	787,66
788,67	0,00378	0,00326	0,00775	0,00472	0,00679	0,00628	0,00593	0,00493	0,00082	0,00855	0,00855	0,00501	0,00469	788,67
789,66	0,00197	0,00170	0,00334	0,00217	0,00308	0,00298	0,00288	0,00254		0,00400	0,00400	0,00229	0,00226	789,66
790,65	0,00304	0,00193	0,00225	0,00215	0,00264	0,00249	0,00263	0,00358		0,00212	0,00175	0,00168	0,00187	790,65
791,64	0,00133	0,00092	0,00100	0,00103	0,00118	0,00111	0,00119	0,00150		0,00081	0,00059	0,00084	0,00090	791,64
792,67	0,00215	0,00157	0,00172	0,00204	0,00207	0,00239	0,00212	0,00239		0,00143	0,00099	0,00157	0,00190	792,67
m/z	QDD_LV3_C1	QDD_LV3_C2	QDD_LV3_C3	QDD_LV3_C4	QDD_LV3_C5	QDD_LV3_C6	QDD_LV3_C7	QDD_LV3_C8	QDD_LV3_C9	QDD_LV3_C10	QDD_LV3_C11	QDD_LV3_C12	QDD_LV3_C13	m/z

Figure 97: List of pointed peaks for each of the 13 centroids (m/z range: 685.50 – 792.67) associated with Figures S3-S9. A value in a cell corresponds to the observed intensity at a given m/z value for a specific cluster. The cell is red if the cluster in question shows the highest intensity among the 13 clusters for the given m/z value. On the same row, a blue cell indicates the second highest intensity among the 13 clusters. An empty cell means that no peak was detected for a particular cluster at this specific m/z value.

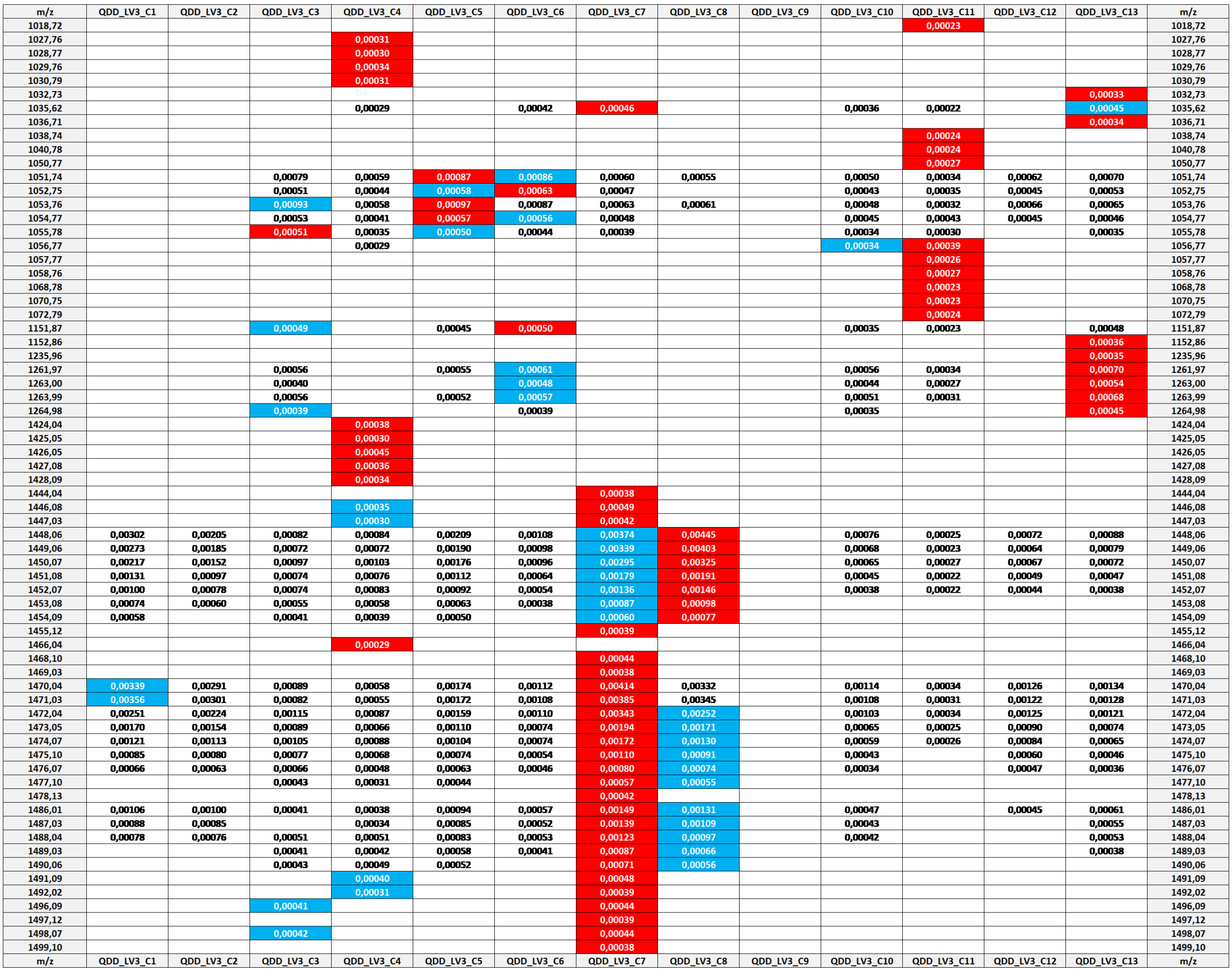


m/z	QDD_LV3_C1	QDD_LV3_C2	QDD_LV3_C3	QDD_LV3_C4	QDD_LV3_C5	QDD_LV3_C6	QDD_LV3_C7	QDD_LV3_C8	QDD_LV3_C9	QDD_LV3_C10	QDD_LV3_C11	QDD_LV3_C12	QDD_LV3_C13	m/z
793,66	0,00108	0,00092	0,00103	0,00120	0,00117	0,00135	0,00109	0,00115		0,00082	0,00056	0,00098	0,00114	793,66
794,67	0,00175	0,00173	0,00216	0,00219	0,00231	0,00282	0,00204	0,00179		0,00186	0,00129	0,00224	0,00249	794,67
795,68	0,00118	0,00128	0,00167	0,00156	0,00167	0,00198	0,00129	0,00118		0,00140	0,00101	0,00195	0,00185	795,68
796,67	0,00087	0,00100	0,00107	0,00129	0,00110	0,00134	0,00102	0,00081		0,00108	0,00084	0,00144	0,00135	796,67
797,75	0,00116	0,00176	0,00252	0,00150	0,00206	0,00217	0,00138	0,00102		0,00252	0,00269	0,00392	0,00241	797,75
798,70	0,00095	0,00132	0,00163	0,00147	0,00143	0,00155	0,00121	0,00086		0,00172	0,00179	0,00240	0,00174	798,70
799,77	0,00080	0,00133	0,00147	0,00106	0,00117	0,00120	0,00088	0,00069		0,00141	0,00161	0,00269	0,00137	799,77
800,64	0,00084	0,00117	0,00113	0,00140	0,00091	0,00106	0,00122	0,00069		0,00127	0,00116	0,00210	0,00133	800,64
801,63		0,00061	0,00054	0,00071	0,00048	0,00056	0,00061			0,00064	0,00057	0,00092	0,00067	801,63
802,64	0,00072	0,00088	0,00081	0,00104	0,00071	0,00095	0,00094	0,00059		0,00099	0,00085	0,00153	0,00104	802,64
803,61			0,00048	0,00054	0,00045	0,00053	0,00050			0,00052	0,00043	0,00084	0,00056	803,61
804,62	0,00064	0,00069	0,00071	0,00068	0,00066	0,00089	0,00072	0,00056		0,00116	0,00158	0,00114	0,00081	804,62
805,59			0,00056	0,00059	0,00051	0,00063	0,00052			0,00071	0,00087	0,00079	0,00060	805,59
806,64	0,00074	0,00079	0,00076	0,00073	0,00068	0,00094	0,00097	0,00061		0,00208	0,00386	0,00129	0,00095	806,64
807,61	0,00058	0,00065	0,00118	0,00247	0,00067	0,00075	0,00079			0,00129	0,00205	0,00119	0,00081	807,61
808,64	0,00116	0,00123	0,00143	0,00221	0,00104	0,00146	0,00195	0,00091		0,00209	0,00215	0,00232	0,00172	808,64
809,65	0,00070	0,00077	0,00137	0,00253	0,00085	0,00094	0,00115	0,00063		0,00114	0,00106	0,00158	0,00108	809,65
810,66	0,00259	0,00264	0,00307	0,00284	0,00268	0,00333	0,00367	0,00234		0,00446	0,00379	0,00403	0,00364	810,66
811,66	0,00126	0,00137	0,00168	0,00127	0,00146	0,00172	0,00179	0,00115		0,00225	0,00198	0,00219	0,00182	811,66
812,65	0,00098	0,00101	0,00176	0,00125	0,00165	0,00178	0,00165	0,00119		0,00167	0,00146	0,00144	0,00163	812,65
813,66			0,00094	0,00068	0,00085	0,00087	0,00077	0,00059		0,00081	0,00075	0,00080	0,00080	813,66
814,65	0,00065	0,00077	0,00082	0,00080	0,00087	0,00089	0,00085	0,00074		0,00117	0,00122	0,00080	0,00081	814,65
815,64			0,00051	0,00049	0,00052	0,00055	0,00052			0,00072	0,00071	0,00065	0,00055	815,64
816,65	0,00061	0,00069	0,00062	0,00069	0,00065	0,00075	0,00070	0,00061		0,00108	0,00131	0,00077	0,00070	816,65
817,62			0,00055	0,00052	0,00051	0,00060	0,00046			0,00064	0,00067	0,00059	0,00053	817,62
818,65	0,00056	0,00060	0,00055	0,00065	0,00057	0,00068	0,00059			0,00060	0,00056	0,00068	0,00065	818,65
819,66			0,00081	0,00082	0,00069	0,00081	0,00057			0,00056	0,00044	0,00081	0,00071	819,66
820,65	0,00070	0,00073	0,00080	0,00089	0,00077	0,00094	0,00076	0,00064		0,00083	0,00086	0,00094	0,00090	820,65
821,66	0,00058	0,00063	0,00087	0,00117	0,00077	0,00089	0,00071	0,00056		0,00071	0,00063	0,00088	0,00086	821,66
822,63	0,00066	0,00075	0,00077	0,00106	0,00074	0,00093	0,00081	0,00060		0,00089	0,00096	0,00099	0,00094	822,63
823,64			0,00069	0,00082	0,00064	0,00071	0,00057			0,00060	0,00057	0,00078	0,00069	823,64
824,63	0,00105	0,00125	0,00100	0,00142	0,00102	0,00140	0,00139	0,00084		0,00127	0,00099	0,00171	0,00155	824,63
825,63	0,00058	0,00069	0,00062	0,00085	0,00057	0,00074	0,00073			0,00066	0,00052	0,00095	0,00081	825,63
826,64	0,00102	0,00118	0,00134	0,00147	0,00134	0,00163	0,00145	0,00097		0,00183	0,00168	0,00177	0,00187	826,64
827,63	0,00063	0,00072	0,00074	0,00088	0,00073	0,00088	0,00080	0,00060		0,00098	0,00088	0,00099	0,00103	827,63
828,64		0,00070	0,00063	0,00071	0,00063	0,00074	0,00065			0,00076	0,00065	0,00093	0,00084	828,64
829,61			0,00048	0,00054	0,00047	0,00054	0,00047			0,00052	0,00042	0,00066	0,00062	829,61
830,64		0,00061	0,00048	0,00054	0,00053	0,00063	0,00053			0,00058	0,00048	0,00074	0,00066	830,64
831,61			0,00047	0,00059	0,00044	0,00049	0,00046			0,00046	0,00036	0,00059	0,00053	831,61
832,64	0,00071	0,00074	0,00068	0,00066	0,00067	0,00091	0,00087	0,00060		0,00099	0,00100	0,00093	0,00099	832,64
833,65	0,00081	0,00084	0,00295	0,00520	0,00146	0,00143	0,00159	0,00078		0,00141	0,00116	0,00207	0,00153	833,65
834,66	0,00134	0,00116	0,00240	0,00325	0,00196	0,00196	0,00223	0,00170		0,00227	0,00279	0,00183	0,00189	834,66
835,66	0,00188	0,00171	0,00813	0,01607	0,00423	0,00335	0,00483	0,00263		0,00366	0,00333	0,00465	0,00341	835,66
836,67	0,00135	0,00123	0,00485	0,00850	0,00303	0,00277	0,00337	0,00185		0,00258	0,00242	0,00289	0,00254	836,67
837,66	0,00108	0,00115	0,00204	0,00334	0,00210	0,00161	0,00230	0,00163		0,00141	0,00117	0,00149	0,00150	837,66
838,69	0,00095	0,00094	0,00157	0,00169	0,00176	0,00177	0,00193	0,00129		0,00143	0,00126	0,00111	0,00158	838,69
839,66			0,00079	0,00077	0,00084	0,00090	0,00085	0,00065		0,00072	0,00062	0,00065	0,00080	839,66
840,63			0,00058	0,00068	0,00060	0,00069	0,00062			0,00061	0,00054	0,00060	0,00068	840,63
841,66			0,00053	0,00059	0,00050	0,00055	0,00040			0,00044	0,00036	0,00053	0,00052	841,66
842,65		0,00060	0,00070	0,00081	0,00066	0,00073	0,00060	0,00055		0,00080	0,00078	0,00069	0,00069	842,65
843,66			0,00065	0,00069	0,00053	0,00060	0,00044			0,00055	0,00051	0,00061	0,00054	843,66
844,65			0,00065	0,00074	0,00060	0,00069	0,00056			0,00081	0,00089	0,00070	0,00063	844,65
845,66			0,00051	0,00060	0,00047	0,00053	0,00043			0,00055	0,00054	0,00055	0,00051	845,66
846,65			0,00046	0,00056	0,00048	0,00054	0,00046			0,00053	0,00050	0,00057	0,00056	846,65
847,67			0,00063	0,00079	0,00052	0,00056	0,00054			0,00067	0,00069	0,00064	0,00062	847,67
848,64	0,00059	0,00066	0,00064	0,00082	0,00062	0,00075	0,00071			0,00096	0,00124	0,00079	0,00085	848,64
849,65			0,00084	0,00109	0,00066	0,00070	0,00074			0,00077	0,00081	0,00080	0,00080	849,65
850,64	0,00056	0,00062	0,00070	0,00092	0,00065	0,00078	0,00079			0,00149	0,00255	0,00080	0,00088	850,64
851,63			0,00054	0,00069	0,00052	0,00059	0,00059			0,00094	0,00142	0,00062	0,00068	851,63
852,66			0,00049	0,00062	0,00051	0,00063	0,00057			0,00081	0,00099	0,00071	0,00072	852,66
853,63			0,00039	0,00043		0,00044	0,00039			0,00047	0,00047	0,00054	0,00049	853,63
854,66				0,00049	0,00046	0,00051	0,00048			0,00055	0,00054	0,00058	0,00059	854,66
855,65			0,00051	0,00069	0,00048	0,00050	0,00045			0,00048	0,00041	0,00064	0,00055	855,65
856,64	0,00066	0,00066	0,00054	0,00066	0,00058	0,00063	0,00070	0,00062		0,00061	0,00048	0,00067	0,00072	856,64
857,67	0,00153	0,00143	0,00480	0,00579	0,00356	0,00384	0,00257	0,00167		0,00255	0,00172	0,00298	0,00327	857,67
858,66	0,00098	0,00097	0,00301	0,00328	0,00223	0,00237	0,00171	0,00105		0,00173	0,00128	0,00190	0,00209	858,66
859,66	0,00120	0,00107	0,00396	0,00553	0,00247	0,00243	0,00287	0,00135		0,00230	0,00155	0,00242	0,00251	859,66
860,67	0,00080	0,00074	0,00210	0,00284	0,00137	0,00151	0,00175	0,00082		0,00201	0,00235	0,00144	0,00162	860,67
861,68	0,00397	0,00338	0,01160	0,01508	0,00933	0,00732	0,01404	0,00575	0,00089	0,00699	0,00388	0,00663	0,00852	861,68
862,69	0,00237	0,00206	0,00538	0,00730	0,00457	0,00371	0,00717	0,00325		0,00544	0,00685	0,00337	0,00425	862,69
863,70	0,00213	0,00195	0,01141	0,01696	0,00706	0,00584	0,00921	0,00345		0,00613	0,00481	0,00560	0,00710	863,70
864,71	0,00111	0,00107	0,00550	0,00807	0,00334	0,00282	0,00417	0,00162		0,00385	0,00476	0,00281	0,00339	864,71
865,70	0,00068	0,00070	0,00212	0,00272	0,00142	0,00123	0,00157	0,00082		0,00157	0,00206	0,00130	0,00136	865,70
866,67		0,00060	0,00088	0,00103	0,00070	0,00072	0,00083			0,00105	0,00132	0,00084	0,00078	866,67
867,66	0,00057	0,00061	0,00092	0,00106	0,00073	0,00075	0,00067	0,00055		0,00075	0,00074	0,00085	0,00073	867,66
868,65	0,00057	0,00069	0,00075	0,00089	0,00063	0,00072	0,00070			0,00090	0,00085	0,00102	0,00079	868,65
869,68	0,00062	0,00067	0,00097	0,00098	0,00080	0,00084	0,00068	0,00060		0,00086	0,00073	0,00101	0,00082	869,68
870,65	0,00057	0,00064	0,00082	0,00093	0,00067	0,00074	0,00067			0,00088	0,00083	0,00090	0,00077	870,65
871,66	0,00056	0,00059	0,00129	0,00105	0,00099	0,00107	0,00080	0,00059		0,00093	0,00083	0,00095	0,00095	871,66
872,64	0,00059	0,00061	0,00097	0,00093	0,00080	0,00092	0,00084	0,00059		0,00097	0,00098	0,00082	0,00090	872,64
873,65			0,00081	0,00084	0,00071	0,00076	0,00074	0,00055		0,00080	0,00079	0,00069	0,00077	873,65
874,64			0,00056	0,00066	0,00058	0,00067	0,00066			0,00138	0,00234	0,00058	0,00070	874,64
875,65			0,00055	0,00073	0,00055	0,00059	0,00064			0,00094	0,00141	0,00055	0,00065	875,65
876,68			0,00053	0,00075	0,00056	0,00069	0,00072			0,00312	0,00683	0,00062	0,00077	876,68
877,67			0,00052	0,00073	0,00054	0,00059	0,00070			0,00175	0,00348	0,00054	0,00070	877,67
878,68			0,00048	0,00072	0,00053	0,00066	0,00070			0,00404	0,00933	0,00065	0,00081	878,68
879,69			0,00041	0,00056		0,00047	0,00047			0,00203	0,00454	0,00047	0,00055	879,69
880,68			0,00039	0,00056		0,00048	0,00047			0,00142	0,00276	0,00050	0,00055	880,68
881,67	0,00056		0,00107	0,00189	0,00083	0,00076	0,00069	0,00064		0,00092	0,00123	0,00080	0,00077	881,67
882,66			0,00072	0,00124	0,00061	0,00060	0,00059			0,00066	0,00070	0,00064	0,00064	882,66
883,66	0,00288	0,00242	0,00717	0,00722	0,00624	0,00641	0,00488	0,00327		0,00450	0,00312	0,00414	0,00532	883,66
884,67	0,00166	0,00137	0,00344	0,00380	0,00296	0,00325	0,00281	0,00179		0,00261	0,00199	0,00216	0,00286	884,67
885,70	0,02946	0,02590	0,06979	0,04126	0,07763	0,07630	0,05481	0,04026	0,00326	0,03384	0,01539	0,04030	0,05500	885,70
886,69	0,01956	0,01669	0,03688	0,02213	0,04324	0,04183	0,02977	0,02642	0,00228	0,02017	0,01142	0,02207	0,02944	886,69
887,70	0,01046	0,00871	0,02204	0,01618	0,02224	0,02053	0,02090	0,01496	0,00152	0,01127	0,00642	0,01162	0,01609	887,70
888,71	0,00362	0,00313	0,00863	0,00646	0,00771	0,00704	0,00859	0,00500	0,00092	0,01441	0,02573	0,00444	0,00583	888,71
889,72	0,00150	0,00139	0,00399	0,00468	0,00316	0,00271	0,00384	0,00203		0,00742	0,01368	0,00216	0,00255	889,72
890,73	0,00093	0,00094	0,00172	0,00221	0,00143	0,00145	0,00185	0,00101		0,01068	0,02467	0,00130	0,00136	890,73
891,72	0,00070	0,00074	0,00109	0,00155	0,00098	0,00098	0,00115	0,00076		0,00530	0,01229	0,00098	0,00097	891,72
892,72	0,00065	0,00068	0,00078	0,00102	0,00073	0,00081	0,00092	0,00062		0,00570	0,01312	0,00085	0,00085	892,72
m/z	QDD_LV3_C1	QDD_LV3_C2	QDD_LV3_C3	QDD_LV3_C4	QDD_LV3_C5	QDD_LV3_C6	QDD_LV3_C7	QDD_LV3_C8	QDD_LV3_C9	QDD_LV3_C10	QDD_LV3_C11	QDD_LV3_C12	QDD_LV3_C13	m/z

Figure 98: List of pointed peaks for each of the 13 centroids (m/z range: 793.66 – 892.72) associated with Figures S3-S9. A value in a cell corresponds to the observed intensity at a given m/z value for a specific cluster. The cell is red if the cluster in question shows the highest intensity among the 13 clusters for the given m/z value. On the same row, a blue cell indicates the second highest intensity among the 13 clusters. An empty cell means that no peak was detected for a particular cluster at this specific m/z value.

m/z	QDD_LV3_C1	QDD_LV3_C2	QDD_LV3_C3	QDD_LV3_C4	QDD_LV3_C5	QDD_LV3_C6	QDD_LV3_C7	QDD_LV3_C8	QDD_LV3_C9	QDD_LV3_C10	QDD_LV3_C11	QDD_LV3_C12	QDD_LV3_C13	m/z
893,68		0,00060	0,00072	0,00086	0,00070	0,00070	0,00078	0,00057		0,00273	0,00580	0,00073	0,00075	893,68
894,65	0,00056	0,00068	0,00066	0,00073	0,00065	0,00077	0,00075			0,00175	0,00315	0,00096	0,00085	894,65
895,66			0,00062	0,00064	0,00058	0,00061	0,00059			0,00092	0,00131	0,00072	0,00065	895,66
896,64			0,00050	0,00054	0,00047	0,00050	0,00047			0,00061	0,00070	0,00055	0,00051	896,64
897,69			0,00062	0,00061	0,00059	0,00063	0,00054			0,00063	0,00059	0,00061	0,00062	897,69
898,66			0,00052	0,00057	0,00051	0,00057	0,00053			0,00058	0,00052	0,00057	0,00058	898,66
899,67	0,00059	0,00063	0,00080	0,00076	0,00087	0,00094	0,00080	0,00062		0,00073	0,00057	0,00070	0,00091	899,67
900,66			0,00059	0,00063	0,00065	0,00077	0,00070			0,00081	0,00084	0,00060	0,00079	900,66
901,67	0,00097	0,00120	0,00095	0,00104	0,00146	0,00164	0,00143	0,00106		0,00106	0,00074	0,00094	0,00164	901,67
902,68	0,00072	0,00086	0,00067	0,00076	0,00098	0,00110	0,00099	0,00075		0,00182	0,00305	0,00073	0,00111	902,68
903,67			0,00051	0,00062	0,00063	0,00068	0,00070			0,00107	0,00170	0,00051	0,00071	903,67
904,71			0,00043	0,00053	0,00052	0,00062	0,00066			0,00477	0,01081	0,00053	0,00064	904,71
905,72			0,00041	0,00052	0,00048	0,00057	0,00051			0,00263	0,00580	0,00047	0,00058	905,72
906,75		0,00060	0,00042	0,00055	0,00051	0,00065	0,00063			0,00693	0,01594	0,00057	0,00065	906,75
907,72			0,00054	0,00075	0,00056	0,00059	0,00055			0,00353	0,00804	0,00053	0,00060	907,72
908,71			0,00050	0,00074	0,00052	0,00060	0,00059			0,00222	0,00464	0,00054	0,00063	908,71
909,68	0,00103	0,00095	0,00232	0,00315	0,00204	0,00147	0,00109	0,00136		0,00153	0,00210	0,00127	0,00121	909,68
910,69	0,00085	0,00079	0,00138	0,00199	0,00125	0,00106	0,00091	0,00098		0,00103	0,00114	0,00099	0,00095	910,69
911,70	0,00126	0,00116	0,00353	0,00402	0,00327	0,00284	0,00178	0,00171		0,00152	0,00114	0,00178	0,00206	911,70
912,71	0,00103	0,00092	0,00201	0,00233	0,00195	0,00178	0,00124	0,00126		0,00113	0,00093	0,00122	0,00139	912,71
913,70	0,00134	0,00121	0,00360	0,00361	0,00320	0,00296	0,00209	0,00181		0,00155	0,00101	0,00177	0,00231	913,70
914,71	0,00098	0,00089	0,00209	0,00207	0,00191	0,00185	0,00137	0,00117		0,00128	0,00117	0,00121	0,00153	914,71
915,70	0,00066	0,00066	0,00122	0,00134	0,00114	0,00114	0,00094	0,00074		0,00088	0,00077	0,00083	0,00106	915,70
916,69	0,00070	0,00065	0,00088	0,00088	0,00083	0,00090	0,00087	0,00067		0,00130	0,00195	0,00079	0,00085	916,69
917,68	0,00083	0,00097	0,00092	0,00108	0,00126	0,00135	0,00109	0,00091		0,00114	0,00127	0,00091	0,00130	917,68
918,68	0,00071	0,00078	0,00077	0,00078	0,00092	0,00099	0,00084	0,00072		0,00135	0,00207	0,00082	0,00095	918,68
919,69		0,00061	0,00061	0,00071	0,00069	0,00073	0,00074	0,00056		0,00087	0,00114	0,00066	0,00076	919,69
920,70			0,00055	0,00057	0,00056	0,00059	0,00059			0,00133	0,00242	0,00065	0,00062	920,70
921,69			0,00042	0,00051	0,00043	0,00047	0,00049			0,00083	0,00132	0,00050	0,00052	921,69
922,70			0,00041	0,00047		0,00045	0,00046			0,00067	0,00093	0,00054	0,00049	922,70
923,69				0,00041		0,00040	0,00041			0,00045	0,00049	0,00044	0,00043	923,69
924,70				0,00044		0,00044	0,00046			0,00047	0,00045	0,00051	0,00046	924,70
925,71				0,00044		0,00040	0,00040			0,00038	0,00033		0,00041	925,71
926,70				0,00043		0,00043	0,00043			0,00050	0,00053	0,00046	0,00045	926,70
927,69				0,00040		0,00040				0,00038	0,00036		0,00040	927,69
928,68				0,00040		0,00044	0,00041			0,00054	0,00062		0,00046	928,68
929,67				0,00038		0,00042				0,00042	0,00041		0,00044	929,67
930,68				0,00042		0,00044	0,00041			0,00050	0,00059		0,00045	930,68
931,66			0,00040	0,00048	0,00056	0,00058	0,00045			0,00047	0,00042		0,00060	931,66
932,69	0,00060	0,00061	0,00047	0,00058	0,00061	0,00067	0,00060	0,00060		0,00068	0,00081	0,00052	0,00065	932,69
933,68	0,00072	0,00103	0,00069	0,00080	0,00121	0,00121	0,00086	0,00083		0,00079	0,00065	0,00078	0,00117	933,68
934,69	0,00070	0,00079	0,00060	0,00068	0,00089	0,00095	0,00076	0,00073		0,00081	0,00088	0,00067	0,00092	934,69
935,70		0,00062	0,00054	0,00057	0,00072	0,00070	0,00061			0,00061	0,00060	0,00057	0,00071	935,70
936,71		0,00059	0,00051	0,00057	0,00063	0,00066	0,00058	0,00055		0,00064	0,00065	0,00066	0,00070	936,71
937,70			0,00042	0,00045	0,00046	0,00048	0,00044			0,00045	0,00042	0,00049	0,00051	937,70
938,71			0,00044	0,00047	0,00046	0,00050	0,00050			0,00054	0,00050	0,00058	0,00053	938,71
939,68			0,00039	0,00041		0,00044	0,00042			0,00044	0,00039	0,00046	0,00046	939,68
940,69				0,00039		0,00044	0,00042			0,00042	0,00037		0,00044	940,69
941,68			0,00042	0,00041		0,00044	0,00041			0,00040	0,00032		0,00044	941,68
942,68			0,00040	0,00040	0,00042	0,00045	0,00041			0,00044	0,00042	0,00045	0,00045	942,68
943,66	0,00058	0,00066	0,00050	0,00053	0,00052	0,00072	0,00071			0,00057	0,00041	0,00073	0,00074	943,66
944,66			0,00040	0,00043	0,00043	0,00058	0,00054			0,00054	0,00050	0,00056	0,00058	944,66
945,65				0,00046	0,00042	0,00056	0,00056			0,00047	0,00037	0,00054	0,00058	945,65
946,68				0,00040		0,00046	0,00045			0,00043	0,00041	0,00045	0,00047	946,68
947,67				0,00042		0,00046	0,00041			0,00038	0,00031		0,00047	947,67
948,70				0,00038		0,00041				0,00038	0,00035		0,00042	948,70
949,69			0,00043	0,00048	0,00064	0,00068	0,00052			0,00044	0,00032	0,00047	0,00065	949,69
950,70				0,00042	0,00047	0,00052	0,00045			0,00040	0,00033	0,00047	0,00051	950,70
951,69				0,00037		0,00043	0,00046			0,00036	0,00026		0,00045	951,69
952,73			0,00041	0,00044			0,00043			0,00035	0,00028	0,00069	0,00041	952,73
953,72				0,00036								0,00052	0,00035	953,72
954,73			0,00042	0,00042						0,00034	0,00028	0,00068	0,00038	954,73
955,72				0,00034							0,00023	0,00048	0,00034	955,72
956,71				0,00034						0,00033	0,00026		0,00036	956,71
957,68				0,00032						0,00035	0,00029		0,00038	957,68
958,71				0,00034						0,00033	0,00028		0,00037	958,71
959,70				0,00033							0,00022		0,00036	959,70
960,75				0,00038		0,00039				0,00033	0,00027	0,00051	0,00039	960,75
961,70				0,00035							0,00022		0,00037	961,70
962,73				0,00034							0,00027	0,00044	0,00035	962,73
963,68				0,00033							0,00025		0,00037	963,68
964,74				0,00032							0,00030		0,00034	964,74
965,68			0,00046	0,00036	0,00055	0,00046	0,00043	0,00055		0,00036	0,00028		0,00045	965,68
966,68				0,00033		0,00039				0,00035	0,00031		0,00039	966,68
967,69				0,00030							0,00024		0,00036	967,69
968,70				0,00029							0,00024		0,00033	968,70
969,69													0,00033	969,69
970,70				0,00030							0,00024		0,00037	970,70
971,69													0,00033	971,69
972,72											0,00022			972,72
974,72											0,00023			974,72
975,69				0,00033		0,00038					0,00023		0,00043	975,69
976,73				0,00035							0,00023	0,00047	0,00038	976,73
977,72				0,00032									0,00038	977,72
978,75			0,00040	0,00041							0,00025	0,00069	0,00036	978,75
979,74				0,00033							0,00022	0,00054	0,00033	979,74
980,77			0,00047	0,00044							0,00022	0,00088	0,00034	980,77
981,76				0,00033								0,00060		981,76
982,77				0,00034								0,00061		982,77
983,70				0,00031									0,00036	983,70
985,68				0,00033									0,00037	985,68
986,73				0,00032								0,00046		986,73
987,68													0,00033	987,68
1001,72				0,00030										1001,72
1002,75				0,00032										1002,75
1004,77				0,00034								0,00049		1004,77
1006,79				0,00031								0,00048		1006,79
1008,77				0,00034								0,00047		1008,77
1010,71				0,00034						0,00037	0,00031		0,00037	1010,71
1011,70				0,00030							0,00024			1011,70
m/z	QDD_LV3_C1	QDD_LV3_C2	QDD_LV3_C3	QDD_LV3_C4	QDD_LV3_C5	QDD_LV3_C6	QDD_LV3_C7	QDD_LV3_C8	QDD_LV3_C9	QDD_LV3_C10	QDD_LV3_C11	QDD_LV3_C12	QDD_LV3_C13	m/z

Figure 99: List of pointed peaks for each of the 13 centroids (m/z range: 893.68 – 1011.70) associated with Figures S3-S9. A value in a cell corresponds to the observed intensity at a given m/z value for a specific cluster. The cell is red if the cluster in question shows the highest intensity among the 13 clusters for the given m/z value. On the same row, a blue cell indicates the second highest intensity among the 13 clusters. An empty cell means that no peak was detected for a particular cluster at this specific m/z value.



m/z	QDD_LV3_C1	QDD_LV3_C2	QDD_LV3_C3	QDD_LV3_C4	QDD_LV3_C5	QDD_LV3_C6	QDD_LV3_C7	QDD_LV3_C8	QDD_LV3_C9	QDD_LV3_C10	QDD_LV3_C11	QDD_LV3_C12	QDD_LV3_C13	m/z
1018,72											0,00023			1018,72
1027,76				0,00031										1027,76
1028,77				0,00030										1028,77
1029,76				0,00034										1029,76
1030,79				0,00031										1030,79
1032,73													0,00033	1032,73
1035,62				0,00029		0,00042	0,00046			0,00036	0,00022		0,00045	1035,62
1036,71													0,00034	1036,71
1038,74											0,00024			1038,74
1040,78											0,00024			1040,78
1050,77											0,00027			1050,77
1051,74			0,00079	0,00059	0,00087	0,00086	0,00060	0,00055		0,00050	0,00034	0,00062	0,00070	1051,74
1052,75			0,00051	0,00044	0,00058	0,00063	0,00047			0,00043	0,00035	0,00045	0,00053	1052,75
1053,76			0,00093	0,00058	0,00097	0,00087	0,00063	0,00061		0,00048	0,00032	0,00066	0,00065	1053,76
1054,77			0,00053	0,00041	0,00057	0,00056	0,00048			0,00045	0,00043	0,00045	0,00046	1054,77
1055,78			0,00051	0,00035	0,00050	0,00044	0,00039			0,00034	0,00030		0,00035	1055,78
1056,77				0,00029						0,00034	0,00039			1056,77
1057,77											0,00026			1057,77
1058,76											0,00027			1058,76
1068,78											0,00023			1068,78
1070,75											0,00023			1070,75
1072,79											0,00024			1072,79
1151,87			0,00049		0,00045	0,00050				0,00035	0,00023		0,00048	1151,87
1152,86													0,00036	1152,86
1235,96													0,00035	1235,96
1261,97			0,00056		0,00055	0,00061				0,00056	0,00034		0,00070	1261,97
1263,00			0,00040			0,00048				0,00044	0,00027		0,00054	1263,00
1263,99			0,00056		0,00052	0,00057				0,00051	0,00031		0,00068	1263,99
1264,98			0,00039			0,00039				0,00035			0,00045	1264,98
1424,04				0,00038										1424,04
1425,05				0,00030										1425,05
1426,05				0,00045										1426,05
1427,08				0,00036										1427,08
1428,09				0,00034										1428,09
1444,04							0,00038							1444,04
1446,08				0,00035			0,00049							1446,08
1447,03				0,00030			0,00042							1447,03
1448,06	0,00302	0,00205	0,00082	0,00084	0,00209	0,00108	0,00374	0,00445		0,00076	0,00025	0,00072	0,00088	1448,06
1449,06	0,00273	0,00185	0,00072	0,00072	0,00190	0,00098	0,00339	0,00403		0,00068	0,00023	0,00064	0,00079	1449,06
1450,07	0,00217	0,00152	0,00097	0,00103	0,00176	0,00096	0,00295	0,00325		0,00065	0,00027	0,00067	0,00072	1450,07
1451,08	0,00131	0,00097	0,00074	0,00076	0,00112	0,00064	0,00179	0,00191		0,00045	0,00022	0,00049	0,00047	1451,08
1452,07	0,00100	0,00078	0,00074	0,00083	0,00092	0,00054	0,00136	0,00146		0,00038	0,00022	0,00044	0,00038	1452,07
1453,08	0,00074	0,00060	0,00055	0,00058	0,00063	0,00038	0,00087	0,00098						1453,08
1454,09	0,00058		0,00041	0,00039	0,00050		0,00060	0,00077						1454,09
1455,12							0,00039							1455,12
1466,04				0,00029										1466,04
1468,10							0,00044							1468,10
1469,03							0,00038							1469,03
1470,04	0,00339	0,00291	0,00089	0,00058	0,00174	0,00112	0,00414	0,00332		0,00114	0,00034	0,00126	0,00134	1470,04
1471,03	0,00356	0,00301	0,00082	0,00055	0,00172	0,00108	0,00385	0,00345		0,00108	0,00031	0,00122	0,00128	1471,03
1472,04	0,00251	0,00224	0,00115	0,00087	0,00159	0,00110	0,00343	0,00252		0,00103	0,00034	0,00125	0,00121	1472,04
1473,05	0,00170	0,00154	0,00089	0,00066	0,00110	0,00074	0,00194	0,00171		0,00065	0,00025	0,00090	0,00074	1473,05
1474,07	0,00121	0,00113	0,00105	0,00088	0,00104	0,00074	0,00172	0,00130		0,00059	0,00026	0,00084	0,00065	1474,07
1475,10	0,00085	0,00080	0,00077	0,00068	0,00074	0,00054	0,00110	0,00091		0,00043		0,00060	0,00046	1475,10
1476,07	0,00066	0,00063	0,00066	0,00048	0,00063	0,00046	0,00080	0,00074		0,00034		0,00047	0,00036	1476,07
1477,10			0,00043	0,00031	0,00044		0,00057	0,00055						1477,10
1478,13							0,00042							1478,13
1486,01	0,00106	0,00100	0,00041	0,00038	0,00094	0,00057	0,00149	0,00131		0,00047		0,00045	0,00061	1486,01
1487,03	0,00088	0,00085		0,00034	0,00085	0,00052	0,00139	0,00109		0,00043			0,00055	1487,03
1488,04	0,00078	0,00076	0,00051	0,00051	0,00083	0,00053	0,00123	0,00097		0,00042			0,00053	1488,04
1489,03			0,00041	0,00042	0,00058	0,00041	0,00087	0,00066					0,00038	1489,03
1490,06			0,00043	0,00049	0,00052		0,00071	0,00056						1490,06
1491,09				0,00040			0,00048							1491,09
1492,02				0,00031			0,00039							1492,02
1496,09			0,00041				0,00044							1496,09
1497,12							0,00039							1497,12
1498,07			0,00042				0,00044							1498,07
1499,10							0,00038							1499,10
m/z	QDD_LV3_C1	QDD_LV3_C2	QDD_LV3_C3	QDD_LV3_C4	QDD_LV3_C5	QDD_LV3_C6	QDD_LV3_C7	QDD_LV3_C8	QDD_LV3_C9	QDD_LV3_C10	QDD_LV3_C11	QDD_LV3_C12	QDD_LV3_C13	m/z

Figure 100: List of pointed peaks for each of the 13 centroids (m/z range: 1018.72 – 1499.10) associated with Figures S3-S9. A value in a cell corresponds to the observed intensity at a given m/z value for a specific cluster. The cell is red if the cluster in question shows the highest intensity among the 13 clusters for the given m/z value. On the same row, a blue cell indicates the second highest intensity among the 13 clusters. An empty cell means that no peak was detected for a particular cluster at this specific m/z value.

Contents lists available at ScienceDirect

Spectrochimica Acta Part B: Atomic Spectroscopy

journal homepage: www.elsevier.com/locate/sab

Optimization of denoising approaches in the context of ultra-fast LIBS imaging

Ruggero Guerrini [a], Cesar Alvarez-Llamas [b,*], Lucie Sancey [c], Vincent Motto-Ros [b], Ludovic Duponchel [a]

[a] *Univ. Lille, CNRS, UMR 8516 – LASIRE – Laboratoire de Spectroscopie pour Les Interactions, La Réactivité et L'Environnement, Lille 59000, France*
[b] *Institut Lumière Matière UMR 5306, Université Lyon 1. CNRS, Villeurbanne, France*
[c] *Université Grenoble Alpes, INSERM U1209, CNRS UMR 5309, Institute for Advanced Biosciences (IAB), Grenoble 38000, France*

ARTICLE INFO

Keywords:
LIBS imaging
Denosing methods
Elemental analysis
PCA
kHz LIBS

ABSTRACT

Laser-Induced Breakdown Spectroscopy (LIBS) has emerged as a powerful analytical tool capable of providing multi-elemental information from a single laser pulse with minimal sample preparation. This technique generates a laser-induced, transient plasma on the sample surface, whose spectral emission is analyzed to determine its elemental composition. μLIBS-Imaging, a variant offering spatially resolved elemental analysis, holds promise for applications in diverse fields such as industry, geology, forensics, and biomedicine. Our drive to go ever faster and analyze increasingly larger areas of interest in samples now compels us to use kHz lasers for this elemental imaging. Despite its potential, implementing such lasers in μLIBS-imaging would face diverse challenges mainly related to weak plasma emission and signal-to-noise ratio (SNR) degradation, particularly when applied to delicate biological samples. This paper investigates methods to enhance SNR in fast μLIBS imaging, particularly for biomedical applications. We focus on denoising techniques suitable for high-frequency laser applications, comparing methods like Savitzky-Golay smoothing, Fast Fourier Transform, wavelet-based filtering, Whittaker Filtering, and Principal Component Analysis (PCA). Our strategy optimizes denoising parameters for specific elemental emission peaks, enhancing SNR for individual elements of interest. The results demonstrate significant improvements in data quality, paving the way for more accurate and efficient elemental imaging in complex biomedical specimens.

1. Introduction

Laser-Induced Breakdown Spectroscopy (LIBS) has gained attention in the analytical community due to its ability to provide simultaneous multi-elemental information from a single laser pulse with minimal sample preparation and no ambient requirements. This analytical technique uses short-laser pulses (typically in the order of nanoseconds or below) focused on a sample surface to obtain chemical information about its constituents. When the laser beam's fluence exceeds a specific threshold value, based on the material of the sample, a small amount of material is ablated from the surface, creating a transient plasma that emits specific spectral signatures resulting from the relaxation of the atoms, ions, or molecules presents in the laser-induced plasma. This spectral signature is then analyzed using a spectrograph to determine the sample's elemental composition. Furthermore, LIBS can detect all elements in the periodic table, with the capability to carry out in-lab, stand-off [1], or in-situ [2] measurements with detection limits in the order of μg.g^{-1}. Moreover, thanks to the small size of the laser-sample interaction area, it is possible to provide a spatially resolved characterization of different surfaces, i.e., elemental imaging. With a lateral resolution in the order of a few μm, μLIBS-imaging is successfully employed in various fields, such as industry [3], geology [4], forensics [5], and biology [6,7], to name just a few. A further appeal for μLIBS-imaging is the analysis speed, which is currently mainly limited by the laser shooting rate. Currently, most of the μLIBS-imaging setups have lasers with a shooting rate lower than 100 Hz [6]. However, the use of kHz lasers could be a significant breakthrough for elemental imaging analysis. Nowadays, the use of kHz laser on LIBS is not widespread, although literature presents several examples mainly focused on industrial [8,9] or geological applications [10]. Using such sampling frequencies, kHz range, offers a significant benefit in terms of analysis time reduction, enabling mapping 1 million spectra - equivalent to 1 cm^2 with

[*] Corresponding author.
E-mail address: cesar.alvarez-llamas@univ-lyon1.fr (C. Alvarez-Llamas).

https://doi.org/10.1016/j.sab.2025.107167
Received 18 November 2024; Received in revised form 23 February 2025; Accepted 24 February 2025
Available online 26 February 2025

a lateral resolution (i.e., shot-to-shot distance) of 10 μm - in roughly 17 min. Building on these advancements, μLIBS-Imaging at kHz rate is attracting the interest of new fields of application. In addition, it requires compact and transportable equipment, with no need for consumables, and reduced economic cost and environmental impacts, making it an interesting candidate to be implemented in fields combining research and routine requirements, such as biomedicine, where specimens can be incredibly diverse and complex, composed of multiple tissues such as bone tissue, muscular or blood vessels [6]. Nonetheless, its implementation has diverse challenges. Firstly, as a direct consequence of the reduction in the analysis time, the measure of elemental distribution maps with a considerable quantity of pixels (1 pixel in the elemental map corresponds to 1 spectrum) becomes easily achievable, resulting in an increase in the complexity of the data treatment process since the number of spectra would be increased in, at least, one order of magnitude. To address this challenge, new and more efficient tools are required; previous work has proposed different solutions, such as the use of novel artificial intelligence techniques, including Facebook libraries for clustering [11], or new mask-creation operations via logical relationships for mineral phase identification [10]. Moreover, to avoid compromising the lateral resolution of the analysis, it is crucial to minimize the induced thermal or stress damage due to laser-matter interaction, especially for the biological specimen, as they are typically more fragile than mineral or metal samples. Thus, low-energy laser pulses (1 mJ or lower) are recommended. Unfortunately, this approach results in a weaker plasma emission. Additionally, detectors capable of achieving an analysis rate in the kHz range are necessary. For kHz LIBS, two-dimensional sensors, such as sCMOS, are typically employed. However, the effective area of the sensor (sensor pixels) must be reduced to achieve the desired acquisition rate, reducing the light amount recorded. The combination of these factors inevitably causes the worsening of the signal-to-noise ratio (SNR) of the acquired spectra and, therefore, a decrease in the elemental image quality generated from them. The noise could strongly impact the quality and quantitative analysis of minor compounds, affecting the limit of detection (LOD). In biological sample analysis, this could reduce the capability of detecting specific minor compounds in the tissue that can have a relevant role in some diseases. Losing this information plays a decisive (negative) role in the study of bioclinical samples. Then, identifying and reducing the noise contributions affecting the analysis is of paramount importance. Numerous methods are available for reducing the influence of noise on acquired signals and enhancing the signal-to-noise ratio (SNR) in analytical measurements. One option is to optimize the instrumental hardware, such as the detector cooling system or adequate gain level selection, which can minimize the noise generated by the detector reading process. We can indeed always make efforts at the instrumental level to reduce noise, but ultimately, this remains quite limited in this particular framework, and therefore, another approach is required: using chemometric methods to clean up the signal as much as possible and improve the quality of the analytical information [12]. Overall, selecting a specific method for reducing noise and enhancing SNR depends on the nature of the data and the specific analytical application. In this context, denoising can be applied either spectrum by spectrum or across the entire dataset. In the first case, denoising can be applied almost in real time; however, if optimization is needed, especially for a heterogeneous sample, a complete spectral dataset is required. Quite intuitively, we can imagine that a denoising method based on analyzing the entire dataset will better understand the structure of the signals, and therefore perform more effectively in this filtering task. In the LIBS domain, the most commonly used methods for denoising are Savitzky-Golay smoothing, Fast Fourier Transform, wavelet-based filtering, and the Whittaker Filter (Whitsm), all based on working on each spectrum individually [13]. To complement these four widely used techniques, we propose also exploring Principal Component Analysis (PCA) in this denoising framework. It's worth noting that PCA has already been utilized in combination with other methods to explore LIBS imaging datasets [14,15]; however, to our knowledge, this is the first time this approach has been applied for denoising purposes in the LIBS domain. In this paper, we compare different techniques for denoising with a focus on fast μLIBS-imaging for biomedical applications, particularly for analyzing endogenous and exogenous elements in tissue. Our approach is adapted to imaging applications: rather than seeking a global denoising method for the entire spectral range, we aim to optimize denoising parameters for specific elemental emission peaks. This targeted strategy allows us to enhance the Signal-to-Noise Ratio (SNR) for individual elements of interest, recognizing that different elements may require distinct denoising parameters or methods for optimal results.

2. Material and methods

2.1. LIBS experimental set-up

The LIBS experimental set-up comprises a kHz laser (Cobolt Tor XE, λ = 1064 nm) capable of achieving a shooting rate of 1000 Hz and the laser-focusing optics (x10 beam expander followed by an x5 objective). The sample is placed on a set of XYZ linear motorized stages to displace precisely the sample during the analysis. The elemental images presented in this work were recorded with a lateral resolution (i.e., shot-to-shot distance) of 10 μm, with a typical crater diameter around 7 μm. Each of them were composed of 1400 by 1500 pixels, derived from 2.1 million acquired spectra. The detection system is based on an Andor's iStar sCMOS sensor coupled to an Andor's Kymera Czerny-Turner spectrograph. The experimental conditions were fixed to 5 μs as integration time without delay time.

2.2. The sample of interest

Rat kidney was collected 1 h after intravenous injection of Au NP (size <10 μm). The sample was fixed in paraformaldehyde 4 % solution for 1 h, before epoxy-embedding following previously reported procedure [16]. The sample was split following the transversal axis. Then, the surface of the sample was polished before analysis.

2.3. Generation of chemical images from LIBS data

The typical μLIBS-imaging data processing workflow involves extracting the analytical signal from the spectra dataset. This analytical signal could be the emission line's maximum intensity, area, or net area for each element detected. Once this signal is extracted, it is used to create a chemical imaging map of the inspected sample. In this work, we use the net peak area as an analytical signal calculated as the signal region's mean minus the background region's mean.

In the sample presented in this work, three elements have been considered: P presents in the whole tissue; Fe, related to the blood vessels; and Au, an exogenous element due to the presence of gold nanoparticles. The purpose of this publication is, of course, not to provide an exhaustive exploration of this tissue but rather to demonstrate the validity of our approach on a set of elements with varying signal qualities within a single sample. Fig. 1a shows a visible image of the kidney analyzed in this study. On this same image, three specific positions (denoted as L1, L2, and L3) were randomly selected so that we could observe the associated LIBS spectra before and after applying denoising strategies. The raw spectra acquired at these three positions are shown in Fig. 1c. Fig. 1b, for its part, presents the integration images generated from the raw spectral data for four elements.

2.4. Denoising methods

As introduced before, we will compare 5 different methods to reduce noise from data while preserving important signal characteristics. The methods selected in this work reflect those commonly employed in various spectroscopic techniques, extending far beyond the scope of

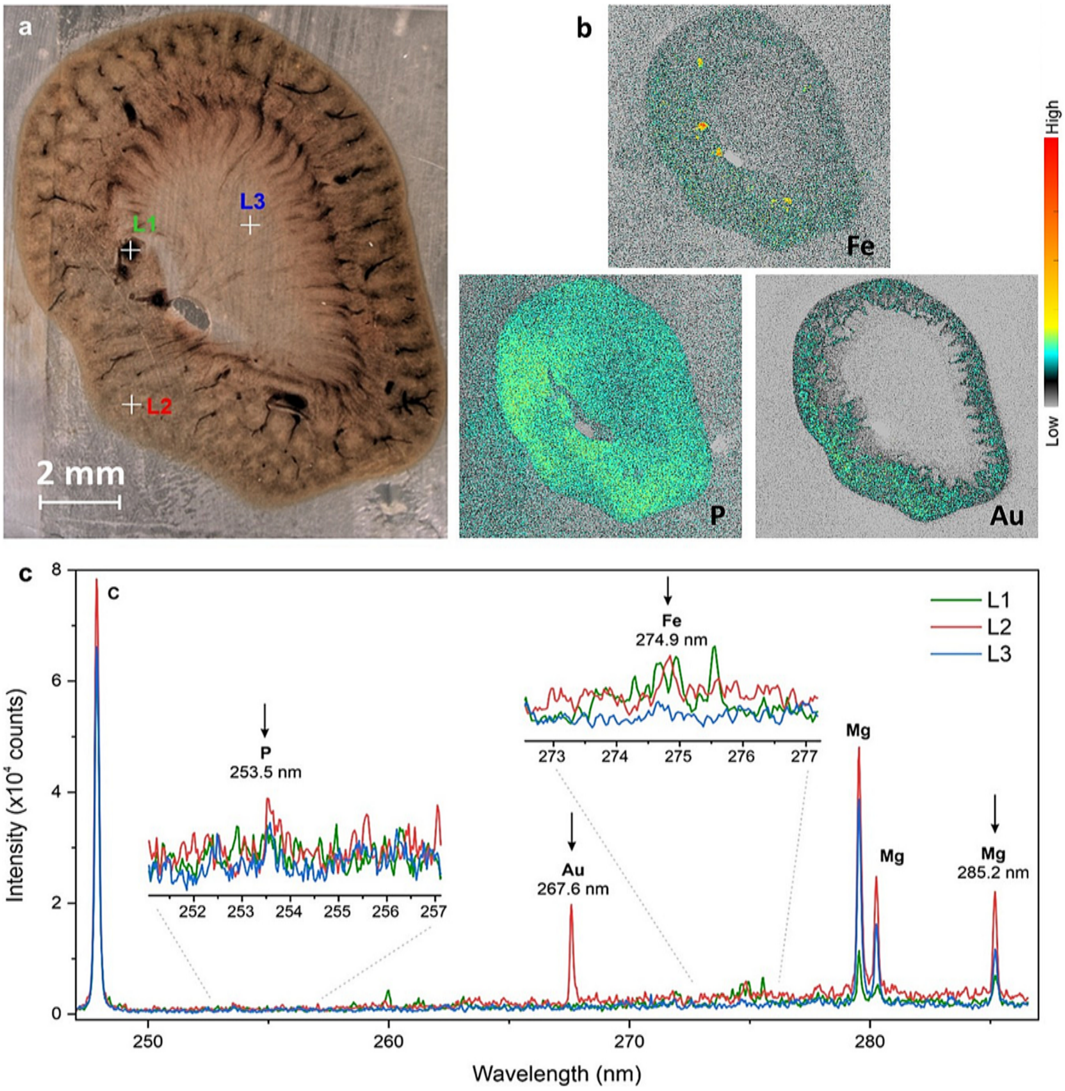


Fig. 1. a) Optical Image of the analyzed sample. b) Elemental distribution for different elements: P (253 nm), Au (267 nm), Fe (274 nm). c) Spectra obtained from the three selected positions, along with the specific spectral regions corresponding to the elements of interest.

LIBS. This is precisely why they are investigated here, as they have demonstrated good performance in denoising spectroscopic data.

2.4.1. *Principal component analysis (PCA)*

Principal Component Analysis [17] is a well-known method for the reduction of dimensionality, allowing a more straightforward representation and description of complex, multivariate datasets. The dataset X, applying PCA, is rewritten as a linear combination of the original variables to explain most of the variance, and it can be described as follows:

$$X = T.P^T + E$$

where **T** is the scores matrix, i.e., the projection of the original data into the low-dimensional space; **P** is the loadings matrix, and **E** is the residuals one. Because noise contributes to a small percentage of the total variance, reconstructing the original matrix after selecting a given number of the very first principal components (PCs) with the higher explained variance will potentially provide a denoised dataset. As can be seen, this denoising procedure applies to the entire dataset simultaneously. The entire optimization of such a denoising method lies in selecting an optimal number of principal components. Indeed, selecting too few components would risk losing part of the signal of interest, while selecting too many would only introduce additional noise.

2.4.2. *Savitzky-Golay smoothing*

This method performs a least-squares fit of a low-degree polynomial to a moving window of data points, effectively smoothing the signal while potentially preserving its features.

The parameters to optimize the Savitzky-Golay smoothing are the width of the window, and the degree of polynomial function used for signal interpolation. This procedure applies independently to each spectrum of the dataset [18].

2.4.3. *Fast Fourier Transform (FFT)*

The Fast Fourier Transform (FFT) decomposes the original signal into a series of frequency components. It can also be shown that this method, like PCA, performs a decomposition of the spectra into a basis of orthogonal components, which in this case are sinusoids of different frequencies. Spectral noise is usually related to small fluctuations with high frequencies in the signal. Relevant signals are usually associated with lower frequencies with higher contributions. By transforming the spectra to the frequency domain, if noise has frequencies different from the interesting ones, reducing them to zero and applying the inverse FFT, it is possible to rebuild the signal, reducing the noise. This procedure also applies independently to each spectrum of the dataset [12].

2.4.4. Wavelet Threshold Denoising (WTD)

Wavelets [19] enables spectra decomposition at various levels, capturing high-frequency noise, low-frequency baselines, and intermediate components at different frequency levels. This multi-scale representation allows for analyzing different signal features at different levels of detail. Wavelet decomposition involves recursively applying a matrix of wavelet filter coefficients to the signal [12]. After decomposition, the small coefficients are typically related to noise, while the large coefficients are associated with the significant features. Removing or reducing the smallest coefficients makes it possible to reduce noise without affecting the main signal features. WTD (Wavelet Threshold Denoising) method has already been used in the LIBS domain by Schlenke et al. [17] as a spectral denoising tool. The parameters to optimize with this approach are the type of wavelet function, the decomposition level, and the threshold choice. The decomposition level determines the number of times the signal is decomposed into approximate and detailed components. The selection of the threshold is crucial and challenging. However, we used in this work the universal threshold method proposed by Donoho et al. [20].

The method involves three steps:

1. Apply the Wavelet transform with a decomposition level chosen in advance.
2. Select a proper threshold for each level.
3. Filter the coefficients below the threshold and reconstruct the signal with the inverse wavelet transform.

There are two main types of filtering, both evaluated in this work: hard and soft. With hard filtering, all the coefficients below the threshold are forced to be zero. Soft filtering reduces the coefficients by a given threshold [20]. Regarding the type of wavelet, we limited ourselves to Daubechies wavelets, which have proven effective in many spectroscopies beyond LIBS.

2.4.5. Whittaker smoothing

Considering the original signal as y and the smoothed one as z, Whittaker Smoothing [21] modifies the signal, balancing two conflicting objectives: (1) data fidelity and (2) roughness of z. This method is based on the use of the penalized least squares method to seek the series z that minimizes Q:

$$Q = S + \lambda R$$

where “S” is the discrepancy from the data, measured as the sum of squared differences $\mathrm{S} = \sum_i (y_i - z_i)^2$; lambda, λ, is a user-chosen parameter; and R, roughness, defined as $\mathrm{R} = \sum_i (z_i - z_{i-1})^2$. A larger λ increases the weight of R in Q, resulting in a smoother z but potentially a poorer fit to the data, as the smoother z is, the more it will diverge from y. This method is particularly used in vibrational spectroscopy on data that are often especially noisy.

2.5. Figures of merit

One of the critical points for better selecting and optimizing the denoising method is the definition of a signal-to-noise ratio (SNR) criterion. Different methods for the calculations of the SNR are available to evaluate the quality of a spectral denoising procedure. The approach used in this paper is the calculation of SNR based on the spectra. It will be denoted SNR_{sp} in this work. It is calculated in three steps considering a given element:

1. Estimation of the noise level: a specific spectral range is selected for the estimation. The 269.2–270.8 nm area is chosen in our case because it does not exhibit significant spectral contributions. Thus, for each spectrum in the dataset, the standard deviation of the measured values within this spectral range is calculated. This dispersion provides an estimate of the noise level for a given spectrum. Then the mean of all these standard deviations is calculated, as a global estimation of noise in the dataset.
2. Estimation of the signal level: a spectral region around an emission line of an element of interest is selected. We calculate the average of these integration values for all the spectra in the dataset, retaining only those within the 99.99th percentile beforehand. This allows us to avoid biasing the estimation due to potentially outlier values while ensuring the detection of weak signals.
3. The SNR_{sp} value is then calculated by taking the ratio of the two previously estimated values.

Beyond these SNR_{sp} values, spectra and images can also be examined before and after correction to better understand the impact of such a method on both spectral and spatial levels.

3. Results and discussions

3.1. Denoising methods optimization

The following section will present the optimization for the different denoising methods. We want to show here that beyond selecting a given method, each one uses parameters that must be optimized to achieve the best denoising. Fig. 2 represents a schema summarizing the data treatment process used in this work. Typically, in the case of imaging results based on spectroscopic techniques, we generally carry out the denoising, before or after extracting the analytical signal. The denoising after extracting the analytical signal is based on the use of image treatment methods; however, in this work, we will only focus on denoising the spectral information before extracting the analytical signal.

As a reference point, the SNR_{sp} values evaluated from the raw dataset are given in Table 1. In fact, we can see that the highest values naturally come from the elements with the highest average emission level, i.e., Au, across the entire dataset. These values will enable us to assess the effectiveness of the different denoising methods.

The proposed methods require an optimization process to obtain the optimum SNR_{sp} in order to obtain the best image reconstruction possible, i.e., we have some a priori knowledge of the elements to be optimized, allowing us to evaluate the SNR improvement for each of the element’s analytical signals separately. This optimization step would need to be redone if we manage spectral data from a new sample with a different composition.

3.2. Principal Component Analysis (PCA)

The PCA was applied to the entire dataset also considering the whole spectral domain. As previously stated, the denoised approach based on PCA is based on the fact that the first principal components (PCs) extracted reflect the chemical variances, and beyond a specific component, only noise is captured. Consequently, selecting an appropriate number of PCs is an essential optimization step. The number of PCs selected is optimized for each element based on the highest SNR improvement (ratio between the calculated SNR and the raw spectra SNR), which varies as a function of the number of retained PCs, as shown in Fig. 3a. We thus observe that the signal-to-noise ratio is improved by a factor of approximately 5 for the three elements when PCA is used, which is quite remarkable. Another way to look at it would be to say that, theoretically, the signal-to-noise ratio observed after this denoising could potentially be achieved with instrumentation 25 times slower, operating not in the kHz range but at 40 Hz. Beyond this initial observation, we must also highlight that an optimal number of different components is obtained for each element. As a result, optimal denoising is achieved for the elements P, Au, and Fe with a total number of principal components equal to 10, 9, and 23, respectively. This result is quite logical, as each one exhibits varying degrees of variance within the spectral dataset and is, by definition, described by a different number of

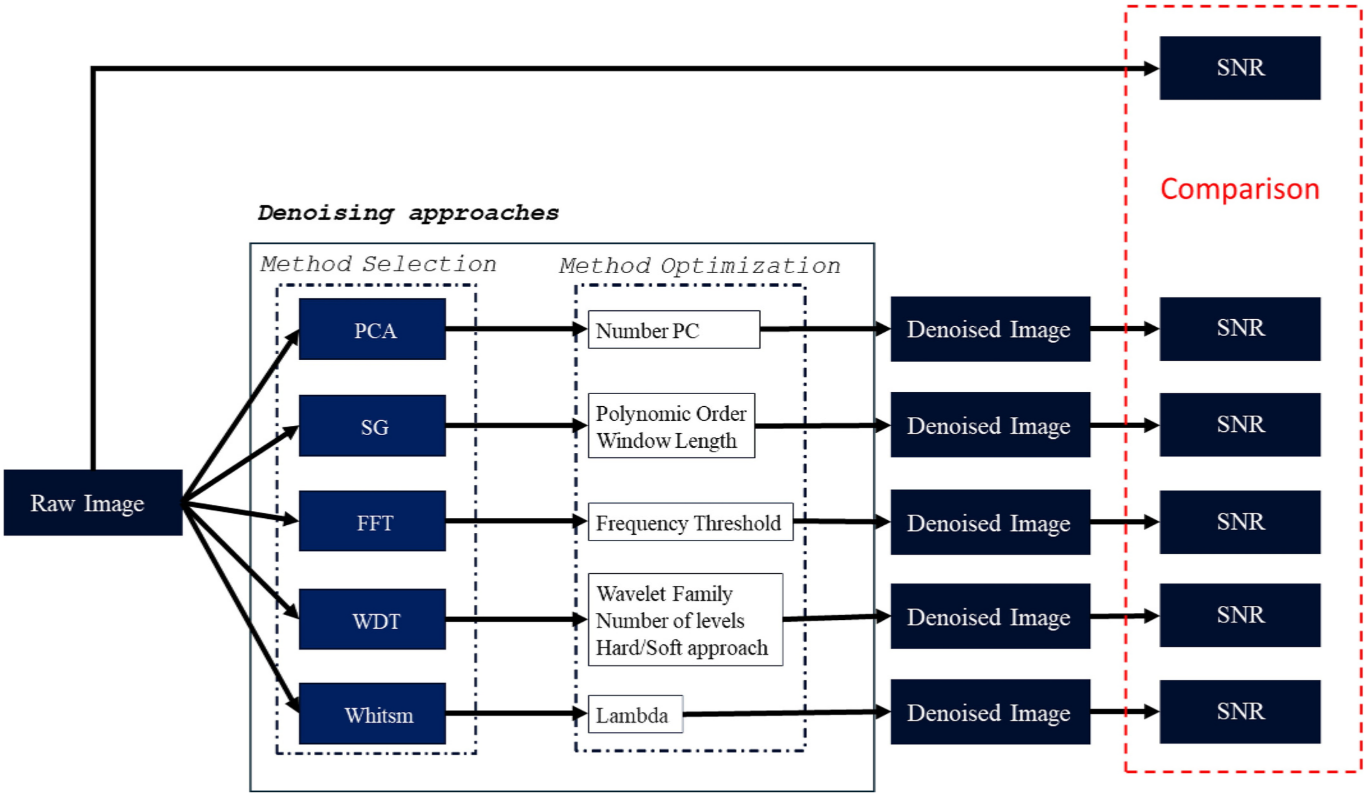


Fig. 2. The data analysis strategy.

Table 1
SNR obtained from the raw spectra for the 3 elements into consideration.

	SNR (x 10^2)
P I $_{253.5\ nm}$	1.19
Au I $_{267.6\ nm}$	4.35
Fe II $_{274.9\ nm}$	1.03

principal components. Beyond these signal-to-noise ratio values measured on the raw data and then on the potentially denoised data, it also seemed important to us to assess the relevance of a denoising strategy in terms of preserving the spectral information contained in the spectra. Indeed, such processing could, in extreme cases, distort or even eliminate an emission line of interest. Fig. 4 shows the evolution of the L1, L2, and L3 spectra, whose locations were presented in Fig. 1, before and after the application of a given denoising method. To be more precise, only the spectral regions of interest are considered for each element. The visual comparison of the spectra enables us to observe the denoising efficiency of PCA for all elements while preserving the associated emission lines, ensuring no distortion. At first glance, some readers might argue that, on the contrary, distortions are indeed observed when PCA is applied for denoising. We are not actually referring to distortions across the entire spectrum but only to the emission line related to the element of interest, which is quite different. Thus, if we refer to the emission lines of P, Au, and Fe at 253.56 nm, 267.6 nm, and 274.9 nm respectively, we observe that, compared to the raw data, their profiles are preserved after PCA denoising. Regarding the emission line observed around 253.5 nm in the raw data, it is not inconsistent for it to disappear after PCA denoising. Indeed, this line is not associated with phosphorus but with another element, which is likely represented by principal components beyond the optimal number chosen for denoising in this specific context of generating a phosphorus image. Building on these results, it is worth pausing to consider the applicability of this denoising approach to a dataset of this scale, comprising 2.1 million spectra. The computations, performed in a MATLAB environment, required in our case 23 GB of RAM, which remains entirely feasible within the computational frameworks used in spectroscopic imaging. Should memory constraints arise for certain users, adopting an HDF5 data format is recommended, as it allows for sequential access to smaller subsets of the spectral dataset as needed. Under such conditions, an alternative implementation of the PCA algorithm, such as incremental PCA, should be considered to accommodate this segmented data exploration [22].

3.3. *Savitzky-Golay smoothing*

The parameters to optimize for the Savitzky-Golay smoothing are the polynomial order and the window size. The first and second orders were tested, while the window lengths tested were between 3 and 31 (with a step of 2). All combinations of these two parameters have thus been studied. When examining the results in Fig. 3b and c, it is noticeable that the signal-to-noise ratio tends to stabilize when the window size reaches approximately 20, regardless of the polynomial order for Au. The improvement in the signal-to-noise ratio is, however, relatively limited in this case to a value of approximately 2, whereas it was 5 for PCA. The signal-to-noise ratio appears to increase consistently for the elements P and Fe, but this trend does not reflect a true improvement in spectral quality in terms of noise. Indeed, in Fig. 4, we observe completely distorted emission lines for the elements of interest following such correction, which is unacceptable. Therefore, Savitzky-Golay smoothing is not optimal for our LIBS imaging data. This approach is indeed very effective for vibrational spectroscopies, for example, but the bands observed are much broader than the noise structure. This is absolutely not the case in LIBS, where we observe very sharp emission lines.

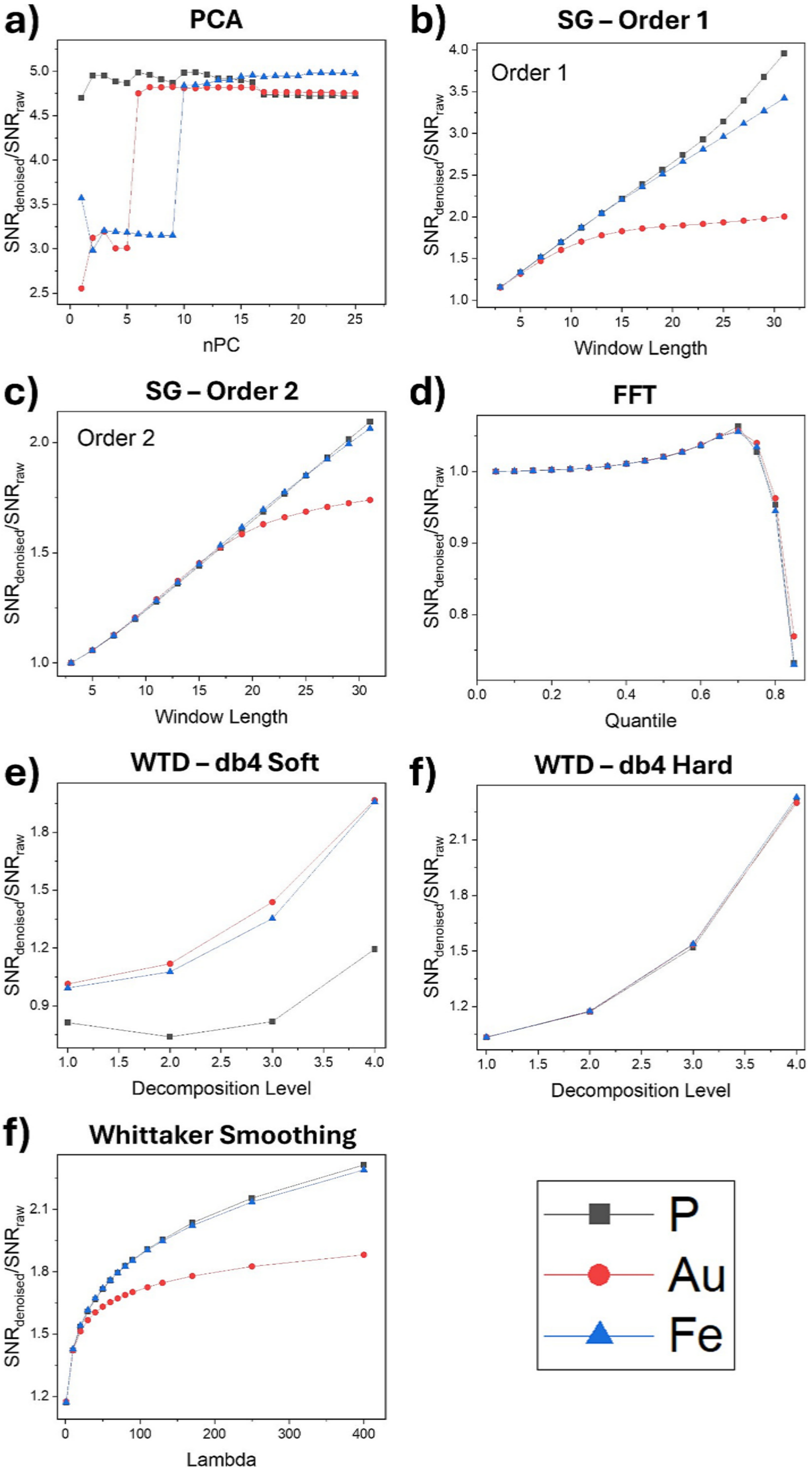


Fig. 3. Evolution of the SNR$_{\text{denoised dataset}}$ /SNR$_{\text{raw dataset}}$ for the different methods.

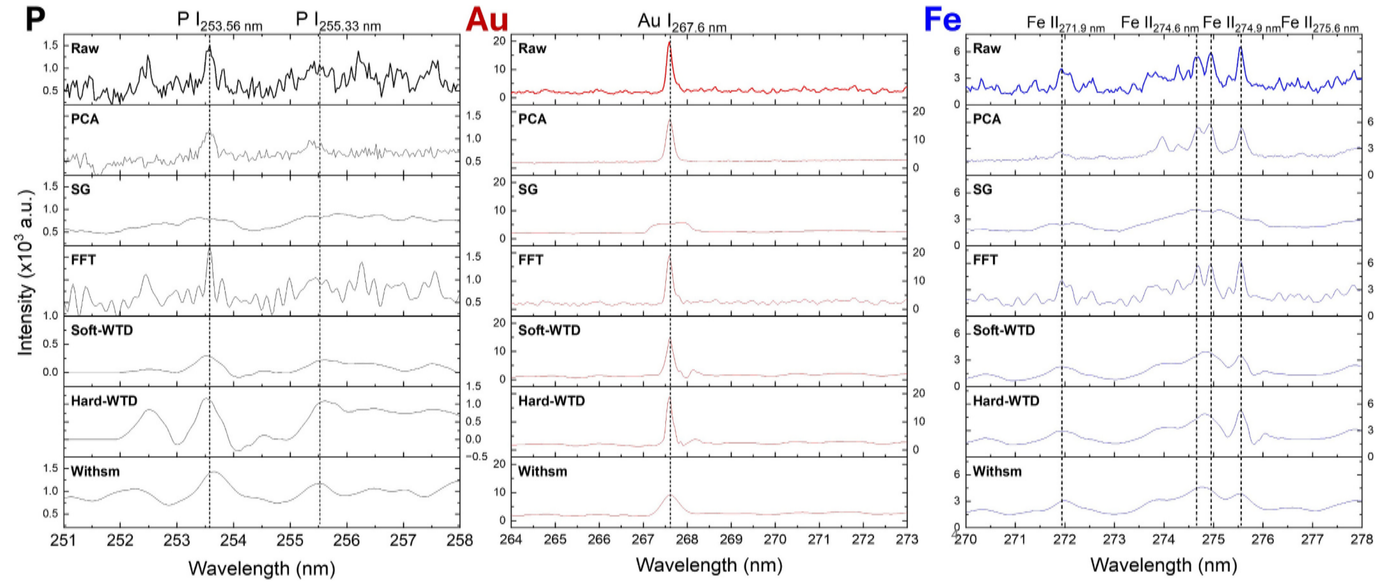


Fig. 4. Denoised Spectra for the SNR optimal values with all the studied denoising methods. Fe spectrum corresponds to the point-of-interest L1, Au to L2, and P to the point-of-interest L3, denoted in Fig. 1.

3.4. *Fast Fourier Transform (FFT)*

Once the signal is decomposed into the frequency domain with the FFT approach, the frequencies that are not related to the signal of interest are set to zero. To achieve this, we calculate the Power Spectrum Density (PSD) and set all frequencies below a predetermined threshold value to zero to eliminate all frequencies that do not significantly contribute to the signal. The different values used as thresholds are quantiles of the PSD. Various quantiles, ranging from 0.05 to 0.85 with a step of 0.05, were evaluated. Fig. 3d shows an optimal SNR$_{sp}$ value of 1.06 for a 0.70 quantile threshold regardless of the element considered. Consequently, the improvement in the signal-to-noise ratio is very minimal, if not negligible, which is also evident in the spectra shown in Fig. 4.

3.5. *Wavelet Threshold Denoising (WTD)*

The wavelet family used in this case was db4 (i.e., Daubechies 4), with a hard and soft denoising approach. We selected this wavelet family as it has demonstrated its effectiveness in numerous spectroscopic techniques [23,24]. We evaluated up to 4 decomposition levels for both hard and soft denoising because going further in the decomposition introduces artifacts in the spectra. Fig. 3e and f present a situation quite comparable to the previously introduced method. For both strategies, the signal-to-noise ratio consistently increases with the level of decomposition, though the improvement remains modest, barely reaching 2.5. Even worse, Fig. 4 shows that applying such a denoising method removes nearly all of the chemical information originally present in the spectra.

3.6. *The Whittaker smoothing*

The lambda value was optimized in this case, ranging from 1 to 400. Fig. 3g demonstrates that we quickly reached a plateau in the improvement of the signal-to-noise ratio for all three elements of interest. This improvement remains limited, as it averages around 2 times. It can be observed in Fig. 4 that, even though the denoised spectra using this approach appear significantly better than those obtained for SG, FFT, and WTD, the emission lines are noticeably broadened compared to the raw data.

Based on these results, we can unequivocally state that the PCA-based denoising procedure is undoubtedly more suited for LIBS imaging. This is a particularly interesting finding, as the other techniques studied here generally perform quite well in the context of other spectroscopic methods. This can, of course, be explained by the specific nature of LIBS data, especially the presence of particularly narrow emission lines. We also emphasize the importance of optimizing the number of components for each element. The final step of this work now involves observing the effects of the denoising procedure using PCA on the integration images. Fig. 5 thus presents a comparison of the integration images obtained from the raw spectra and the spectra denoised using PCA for the three elements considered. Starting with a global observation of this figure, we can see that the largest differences are observed for Fe and P. The differences are indeed less noticeable for Au. This is naturally explained by the fact that the weakest signals are associated with Fe and P, and it is precisely under these conditions that an increase in the signal-to-noise ratio has a significant impact. Arrows have also been added to these images to highlight details or areas that differ significantly between them and will therefore be discussed in greater detail. If we first examine the iron maps, we can immediately observe an overestimation of concentrations when raw spectra are used, as seen, for example, in positions a vs. a' and b vs. b'. This iron originates from blood and is particularly present in the vascularized areas of the organ. Therefore, it cannot be found outside the organ or at its center, which is much more consistent in the image obtained after PCA denoising. A low dynamic range and very low contrast are also observed in the iron image derived from the raw spectra—values that are significantly improved after denoising, allowing the observation of previously unseen details (d vs. d', e vs. e', and f vs. f'). Fairly similar observations can be made for P. However, these differences are less striking, as we know that phosphorus is present in numerous cells and is distributed throughout the organ. As a result, overestimations of phosphorus concentrations are once again observed outside outside the organ and at its center (g vs. g', i vs. i', and l vs. l') when raw spectra are used. Imperceptible details from the raw data are also revealed when the spectra are denoised (h vs. h', j vs. j' and k vs. k'). Regarding gold, our biological understanding of the issue allows us to state that it can only be present at the periphery of the organ. As a result, an overestimation of concentrations is observed both outside the organ and at its center, which is much more consistent when denoised spectra are used. Through these three elements with varying concentration ranges and locations within this complex organ, we were able to demonstrate that applying PCA as a

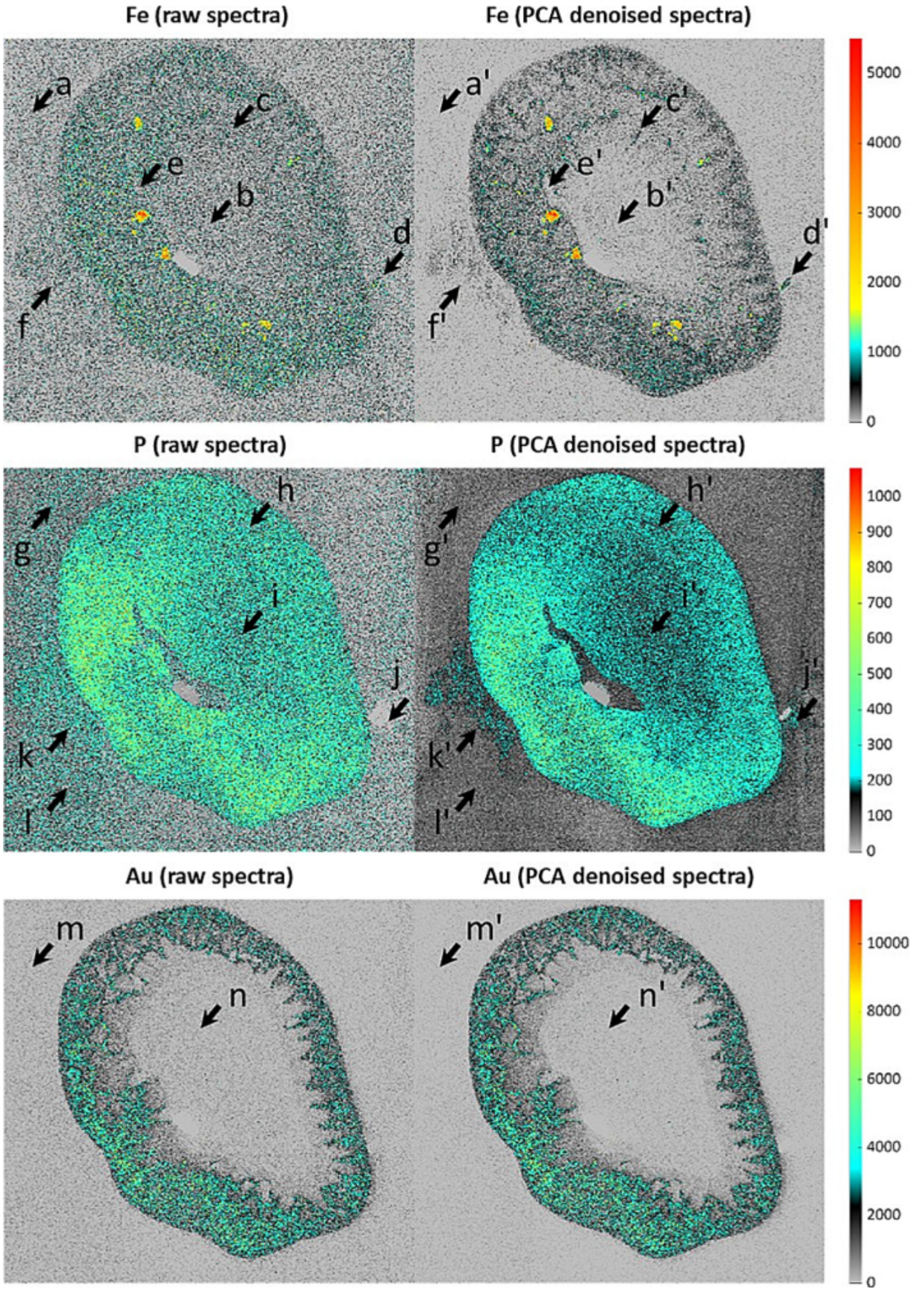


Fig. 5. Comparison between the original and denoised elemental distribution maps for Fe (274.9 nm), P (253.5 nm) and Au (267.6 nm).

denoising method generates elemental images with improved contrast, greater dynamic range, and, most importantly, reduced bias, allowing us to better highlight the biological reality of the sample after processing.

4. Conclusions

The use of new experimental approaches that enable higher throughput and higher resolution analysis makes the parallel development of new chemometric tools mandatory. As the complexity and size of spectral data increases - with hypercubes that can hold millions of spectra - it is necessary to develop algorithms and workflows for spectral processing to handle, analyze, and extract analytical information from these data. In this work, we highlight the application of kHz μLIBS-imaging for the analysis of samples of bio-clinical interest, with a focus on a comparative evaluation of 5 different denoising methods. Furthermore, to our knowledge, this research applies principal component analysis (PCA) and Whittaker Smoothing to LIBS data for the first time, opening new ways to improve the accuracy of such analyses. The results shows that PCA is by far the most effective method in this specific LIBS framework, offering a better enhancement than the other methods. Specifically, PCA provides an important SNR enhancement of approximately 5 times for the three elements under study compared to the raw data; moreover, no distortion of the emission line of a given element has been found, unlike the other denoising methods studied. In this work, PCA was applied to the entire available spectral range, but in more delicate cases, we could consider using the restricted range around the emission line of interest for the PCA calculation. We certainly evaluated this option, but we did not observe any significant improvement, at least for this particular dataset. In conclusion, this enhancement in the quality of the kHz μLIBS-imaging highlights the PCA value for the data treatment of LIBS-based applications, particularly where the experimental conditions limit the quality of spectral data.

CRediT authorship contribution statement

Ruggero Guerrini: Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Cesar Alvarez-Llamas:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Lucie Sancey:** Writing – review & editing, Resources, Methodology, Conceptualization. **Vincent Motto-Ros:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Funding acquisition. **Ludovic Duponchel:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Cesar Alvarez Llamas reports financial support was provided by Horizon Europe. Vincent Motto-Ros reports financial support was provided by French National Research Agency. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The project leading to this application was supported by the European Union's Horizon Europe Research and Innovation Program under the MSCA Postdoctoral Fellowships 2022, with agreement number 101106090; by the French Government "France 2030" initiative, under the DIADEM program managed by the "Agence Nationale de la Recherche" (ANR-22-PEXD-0014, "Libelul") and the "French Agence Nationale de la Recherche" (ANR-20-CE17-0021 "dIAg-EM", and "Micro-Q-Li" ANR-20-CE05-0015)

Data availability

Data will be made available on request.

References

[1] R.C. Wiens, A. Udry, O. Beyssac, C. Quantin-Nataf, N. Mangold, A. Cousin, L. Mandon, T. Bosak, O. Forni, S.M. McLennan, V. Sautter, A. Brown, K. Benzerara, J.R. Johnson, L. Mayhew, S. Maurice, R.B. Anderson, S.M. Clegg, L. Crumpler, T.S. J. Gabriel, P. Gasda, J. Hall, B.H.N. Horgan, L. Kah, C. Legett, J.M. Madariaga, P.-Y. Meslin, A.M. Ollila, F. Poulet, C. Royer, S.K. Sharma, S. Siljeström, J.I. Simon, T. E. Acosta-Maeda, C. Alvarez-Llamas, S.M. Angel, G. Arana, P. Beck, S. Bernard, T. Bertrand, B. Bousquet, K. Castro, B. Chide, E. Clavé, E. Cloutis, S. Connell, E. Dehouck, G. Dromart, W. Fischer, T. Fouchet, R. Francis, J. Frydenvang, O. Gasnault, E. Gibbons, S. Gupta, E.M. Hausrath, X. Jacob, H. Kalucha, E. Kelly, E. Knutsen, N. Lanza, J. Laserna, J. Lasue, S. Le Mouélic, R. Leveille, G. Lopez Reyes, R. Lorenz, J.A. Manrique, J. Martinez-Frias, T. McConnochie, N. Melikechi, D. Mimoun, F. Montmessin, J. Moros, N. Murdoch, P. Pilleri, C. Pilorget, P. Pinet, W. Rapin, F. Rull, S. Schröder, D.L. Shuster, R.J. Smith, A.E. Stott, J. Tarnas, N. Turenne, M. Veneranda, D.S. Vogt, B.P. Weiss, P. Willis, K.M. Stack, K. H. Williford, K.A. Farley, The SuperCam Team, Compositionally and density stratified igneous terrain in Jezero crater, Mars, Sci. Adv. 8 (2022) eabo3399, https://doi.org/10.1126/sciadv.abo3399.

[2] C. Fabre, N.E. Ourti, C. Ballouard, J. Mercadier, J. Cauzid, Handheld LIBS analysis for in situ quantification of Li and detection of the trace elements (Be, Rb and Cs), J. Geochem. Explor. 236 (2022) 106979, https://doi.org/10.1016/j.gexplo.2022.106979.

[3] S. Grünberger, G. Watzl, N. Huber, S. Eschlböck-Fuchs, J. Hofstadler, A. Pissenberger, H. Duchaczek, S. Trautner, J.D. Pedarnig, Chemical imaging with laser ablation – spark discharge – optical emission spectroscopy (LA-SD-OES) and laser-induced breakdown spectroscopy (LIBS), Opt. Laser Technol. 123 (2020) 105944, https://doi.org/10.1016/j.optlastec.2019.105944.

[4] C. Fabre, Advances in laser-induced breakdown spectroscopy analysis for geology: a critical review, Spectrochim. Acta Part B At. Spectrosc. 166 (2020) 105799, https://doi.org/10.1016/j.sab.2020.105799.

[5] M. López-López, C. Alvarez-Llamas, J. Pisonero, C. García-Ruiz, N. Bordel, An exploratory study of the potential of LIBS for visualizing gunshot residue patterns, Forensic Sci. Int. 273 (2017) 124–131, https://doi.org/10.1016/j.forsciint.2017.02.012.

[6] V. Gardette, V. Motto-Ros, C. Alvarez-Llamas, L. Sancey, L. Duponchel, B. Busser, Laser-induced breakdown spectroscopy imaging for material and biomedical applications: recent advances and future perspectives, Anal. Chem. 95 (2023) 49–69, https://doi.org/10.1021/acs.analchem.2c04910.

[7] V.H.C. Ferreira, V. Gardette, B. Busser, L. Sancey, S. Ronsmans, V. Bonneterre, V. Motto-Ros, L. Duponchel, Enhancing diagnostic capabilities for occupational lung diseases using LIBS imaging on biopsy tissue, Anal. Chem. 96 (2024) 7038–7046, https://doi.org/10.1021/acs.analchem.4c00237.

[8] R. Noll, H. Bette, A. Brysch, M. Kraushaar, I. Monch, L. Peter, V. Sturm, Laser-Induced Breakdown Spectrometry – Applications for Production Control and Quality Assurance in the Steel Industry ℭ, At. Spectrosc, 2001.

[9] F. Boué-Bigne, Analysis of oxide inclusions in steel by fast laser-induced breakdown spectroscopy scanning: an approach to quantification, Appl. Spectrosc. 61 (2007) 333–337, https://doi.org/10.1366/000370207780220895.

[10] C. Alvarez-Llamas, A. Tercier, C. Ballouard, C. Fabre, S. Hermelin, J. Margueritat, L. Duponchel, C. Dujardin, V. Motto-Ros, Ultrafast μLIBS imaging for the multiscale mineralogical characterization of pegmatite rocks, J. Anal. At. Spectrom 39 (2024) 1077–1086, https://doi.org/10.1039/D3JA00438D.

[11] L. Duponchel, R. Guerrini, V.H.C. Ferreira, C.A. Llamas, C. Dujardin, V. Motto-Ros, When social media empowers analytical chemists to explore millions of spectra derived from a complex sample, Anal. Chem. 96 (2024) 3994–3998, https://doi.org/10.1021/acs.analchem.3c05724.

[12] J.-M. Roger, J.-C. Boulet, M. Zeaiter, D.N. Rutledge, Pre-processing methods, in: Compr. Chemom., Elsevier, 2020, pp. 1–75, https://doi.org/10.1016/B978-0-12-409547-2.14878-4.

[13] B. Zhang, L. Sun, H. Yu, Y. Xin, Z. Cong, A method for improving wavelet threshold denoising in laser-induced breakdown spectroscopy, Spectrochim, Acta Part B At. Spectrosc. 107 (2015) 32–44, https://doi.org/10.1016/j.sab.2015.02.015.

[14] R. Finotello, M. Tamaazousti, J.-B. Sirven, HyperPCA: a powerful tool to extract elemental maps from noisy data obtained in LIBS mapping of materials, Spectrochim. Acta Part B At. Spectrosc. 192 (2022) 106418, https://doi.org/10.1016/j.sab.2022.106418.

[15] P. Pořízka, J. Klus, E. Képeš, D. Prochazka, D.W. Hahn, J. Kaiser, On the utilization of principal component analysis in laser-induced breakdown spectroscopy data analysis, a review, Spectrochim, Acta Part B At. Spectrosc. 148 (2018) 65–82, https://doi.org/10.1016/j.sab.2018.05.030.

[16] L. Sancey, S. Kotb, C. Truillet, F. Appaix, A. Marais, E. Thomas, B. van der Sanden, J.-P. Klein, B. Laurent, M. Cottier, R. Antoine, P. Dugourd, G. Panczer, F. Lux, P. Perriat, V. Motto-Ros, O. Tillement, Long-term *in vivo* clearance of gadolinium-based AGuIX nanoparticles and their biocompatibility after systemic injection, ACS Nano 9 (2015) 2477–2488, https://doi.org/10.1021/acsnano.5b00552.

[17] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (2014) 2812–2831, https://doi.org/10.1039/C3AY41907J.

[18] Abraham Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639, https://doi.org/10.1021/ac60214a047.

[19] K. Jetter, U. Depczynski, K. Molt, A. Niemöller, Principles and applications of wavelet transformation to chemometrics, Anal. Chim. Acta 420 (2000) 169–180, https://doi.org/10.1016/S0003-2670(00)00889-8.

[20] D.L. Donoho, I.M. Johnstone, Ideal Spatial Adaptation by Wavelet Shrinkage, 2024.

[21] P.H.C. Eilers, A perfect smoother, Anal. Chem. 75 (2003) 3631–3636, https://doi.org/10.1021/ac034173t.

[22] D.A. Ross, J. Lim, R.-S. Lin, M.-H. Yang, Incremental learning for robust visual tracking, Int. J. Comput. Vis. 77 (2008) 125–141, https://doi.org/10.1007/s11263-007-0075-7.

[23] D. Chen, X. Shao, B. Hu, Q. Su, A background and noise elimination method for quantitative calibration of near infrared spectra, Anal. Chim. Acta 511 (2004) 37–45, https://doi.org/10.1016/j.aca.2004.01.042.

[24] X.-G. Ma, Z.-X. Zhang, Application of wavelet transform to background correction in inductively coupled plasma atomic emission spectrometry, Anal. Chim. Acta 485 (2003) 233–239, https://doi.org/10.1016/S0003-2670(03)00395-7.