KU LEUVEN

FACULTY OF SCIENCE

Peak Learning for Denoising Mass Spectrometry Imaging Data

Chris BUTCHER
Serkan SHENTYURK

Supervisor: Prof. Bart de Moor
Mentor: *Melanie Nijs*
Thomas Vanhemel

Thesis presented in
fulfillment of the requirements
for the degree of Master of Science
in Statistics and Data Science

Academic year 2023-2024

Preface

First and foremost, we are truly grateful to our thesis mentors, Melanie Nijs and Thomas Vanhemel. They did an amazing job by providing guidance, support, and constructive feedback throughout the project. We thank them for the great advice, feedback and support.

Lastly, we would like to thank the members of the jury for takın the time to read the manuscript.

Chris Butcher
Serkan Shentyurk

Contribution Statement

This thesis has been written in collaboration with our supervisor Prof. Bart de Moor and our daily mentors Melanie Nijs and Thomas Vanhemel. Below we will clarify our personal contributions to the thesis.

Chapter 3 and Chapter 4 are written by Serkan Shentyurk; Chapter 5, Section 6.1, and Section 6.2 are written by Chris Butcher. Other chapters are written and edited by both authors.

Summary

The distribution of molecules within tissue is crucial for proper functionality, and disruption in the composition of the molecules can lead to disease formation. Mass spectrometry imaging (MSI) is a technique used to investigate the spatial distribution of the molecules by scanning the tissue sample, ionising the molecules and determining their mass-to-charge (m/z) ratios. However, the method is not error-proof: unstable molecules can decompose, a phenomenon called fragmentation. In addition to fragmentation, ionized molecules may possess different kinetic energies and isotopes, resulting in a Gaussian distribution around the m/z ratio of the molecule. These 'additional peaks' that are related to the true molecules make it harder to interpret the data. Therefore, we aim to identify and remove these additional peaks to explore whether we can extract further analysis from the data.

In this dissertation, we first constructed a map of the islets of Langerhans, where insulin is produced and stored, by applying an edge detection algorithm. Then, we applied several techniques to identify the related peaks of insulin, which is abundant in the pancreatic tissue sample. Since insulin-related peaks exist where insulin is present, we initially used Pearson correlation analysis to identify the m/z ratios that are highly correlated with insulin. We then fitted a Gaussian Mixture Model to capture the Gaussian distribution around these related peaks.

In another method, we utilized the ion images of individual m/z values to group them based on their spatial distributions before using correlation methods to identify those correlated with insulin. We evaluated the candidate insulin-related peaks by first checking their spatial distributions using the map of the islets of Langerhans and then comparing the outputs of these two different methods. Finally, we investigated whether it is possible to gain further insights from the data when the related peaks are removed. We applied UMAP to the data both before and after removing the insulin-related peaks.

Within the results of both approaches, a common set of m/z values was identified, suggesting that these methods were able to identify at least some of the true insulin-related peaks within the data. Additionally, UMAP analysis showed that some pixels have similar m/z intensities to both the islets and the non-islets, indicating a different structure at the borders of the islet. UMAP analysis also revealed a distinguishable difference between the UMAP results on the original data and the data after the insulin-related peaks were removed. This implies that the application of these methods can lead to a greater understanding of the structures within pancreatic tissue.

List of Abbreviations and Symbols

δ-cells Somatostatin cells

AIC Akaike Information Criterion

AoI Area of Interest

BIC Bayesian Information Criterion

Da Dalton

DESI Desorption electrospray ionisation

ESI Electrospray Ionization

FPC Fuzzy Partition Coefficient

FT-ICR Fourier transform ion cyclotron resonance

GMM Gaussian Mixture Model

H&E Hematoxylin and Eosin

HDBSCAN hierarchical density-based spatial clustering of applications with noise

Ins1 insulin 1

Ins2 insulin 2

m/z mass-to-charge

MALDI Matrix Assisted Laser Desorption/Ionization

MS/MS tandem mass spectrometry

MSI Mass Spectrometry Imaging

MS Mass Spectrometry

SIMS Secondary ion mass spectrometry Ionization

TOF Time-of-Flight

WCSS Within-Cluster Sum of Square

Contents

Chapter 1

Introduction

This chapter is written by both authors.

1.1 Motivation and Objective

Organisms contain various components, ranging from small molecules to large proteins, which communicate with each other in order to function properly [3]. Disruptions in their composition have been shown to contribute to disease formation [24]. Therefore, understanding the distribution of small molecules, a field known as metabolomics, is crucial for uncovering insights about mechanisms that are at play during healthy and pathological tissue function.

Metabolomics plays a vital role in gaining insights into the mechanisms and events occurring within an organism. A widely used metabolomics approach, referred to as bulk metabolomics, operates as follows: metabolites from the sample of interest are extracted and then quantified collectively, typically using mass spectrometry (MS) [28, 40]. However, since this method quantifies all metabolites together, it cannot pinpoint the specific locations of these metabolites within the sample of interest. Spatial metabolomics represents an alternative approach in metabolomics designed to address the challenge of localizing metabolites, which is a limitation of bulk metabolism [42]. This methodology finds applications across various biomedical research domains [22, 32, 30, 48]. Several techniques are available for spatial metabolomics, with mass spectrometry imaging (MSI) standing out as the most widely adopted method [33].

The quality of MSI results is influenced by factors such as sample preparation [52], tissue thickness [50], the storage of the sample [58], and data processing [55]. However, despite its widespread use and favourable reputation, MSI can exhibit artefact-like peaks arising from a phenomenon known as fragmentation, which involves the decomposition of unstable molecular ions during the ionization phase [16]. Two main types of reactions lead to fragmentation: simple bond cleavage reactions and rearrangement reactions [16]. As a consequence of fragmentation, additional peaks are observed in mass spectra. Ionized molecules with the same mass-to-charge (m/z) ratio may possess different kinetic energies and molecules can have various isotopes, resulting in a Gaussian distribution around the exact m/z ratio of the molecule [57]. In addition to fragmentation and isotopes, additional ions can bind to the compounds, leading to an increase in the m/z ratio, a phenomenon

called adduct formation [51]. These complications make the interpretation of mass spectra more challenging. To accurately infer the mass spectra of a specific sample, these artefact-like peaks, including fragmented ions, molecules with varying kinetic energy, and adducts, must be identified.
We aim to identify the peaks of insulin, which are abundant in our MSI data obtained from healthy pancreatic tissue of mice, and its corresponding distribution, and subsequently remove them from the data. This approach allows us to test whether it is possible to uncover additional details that are masked due to the overwhelming abundance of insulin. We believe that the method we are developing will not only be applicable to insulin in mouse pancreatic tissue but also to the identification of peaks related to any compound of interest in a given mass spectrum. This has the potential to open new avenues for discovery.

1.2 Outline

The next chapter of this master's dissertation, Chapter 2, begins by describing pancreatic tissue and its structure, with a focus on the islets of Langerhans and insulin. Subsequently, the basic concepts of MS and MSI are explained, introducing the components of the MSI method, followed by the presentation of the MSI data used throughout the dissertation.

Following this, Chapter 3 presents our attempt to construct the map of the islets of Langerhans using edge detection and different clustering algorithms. This mapping will be utilized in the validation step to assess whether the candidate insulin-related peaks are significantly found within the islets of Langerhans.

Chapters 4 and 5 are dedicated to two different methods used to identify insulin-related peaks. Chapter 4 approaches the problem with the assumption that insulin-related m/z values must be correlated with insulin m/z ratios. Gaussian Mixture Models (GMMs) are then fitted to the significantly correlated m/z values, and the candidate insulin-related peaks are identified. Taking a different approach, the method in Chapter 5 attempts to utilize the spatial distribution of the m/z values to determine potential insulin-related peaks. A sliding window is used to group m/z values that are similar in spatial distribution, and the correlations between representative m/z values are subsequently assessed to identify candidate insulin-related peaks.

Chapter 6 focuses on validating our candidate m/z values by comparing the results of methods used in Chapters 4 and 5, examining their spatial distribution, and consulting literature. The MSI data is then analyzed after the insulin-related peaks are removed to explore whether new insights can be obtained from the data.

Finally, we conclude our work in Chapter 7.

Chapter 2

Pancreatic Tissue and MSI

This chapter is written by both authors.

The dissertation focuses on several methods to identify insulin-related peaks in MSI data obtained from healthy mouse pancreatic tissue. To ensure clarity, we begin by describing the pancreatic tissue and its pivotal component for our research, the islets of Langerhans. Understanding the structure and function of pancreatic tissue, particularly the islets of Langerhans, provides essential context for our investigation into insulin-related peaks.

Subsequently, we delve into the principles of MS and MSI methods. We explain how these techniques function, outline their main components, and elucidate the process of MSI data acquisition and analysis. This foundational knowledge is crucial for comprehending the methodologies employed in our research and the interpretation of our findings.

By providing a thorough understanding of pancreatic tissue, the islets of Langerhans, and the principles of MS and MSI, we establish a solid framework for exploring our methods to identify insulin-related peaks in MSI data.

2.1 Pancreatic Tissue

Before the advent of MSI, hematoxylin and eosin (H&E) staining facilitated the spatial analysis of biopsies. H&E staining imparts distinct colours to tissue, enhancing the contrast between various anatomical elements. Hematoxylin imparts purple colour to the nucleus, while eosin imparts pink colour to the rest of the cell. [13, 21].

Following staining, the tissue slice is examined under a microscope. Patterns in the colouring reveal cell structures, enabling the identification of the anatomical compositions. Figure 2.1 illustrates an example of a stained microscopic image of a healthy mouse pancreas.

2.1.1 The Islets of Langerhans

The islets of Langerhans play a crucial role in insulin production and serve as the endocrine component of the pancreas, regulating blood glucose levels [11]. Figure 2.2 illustrates a

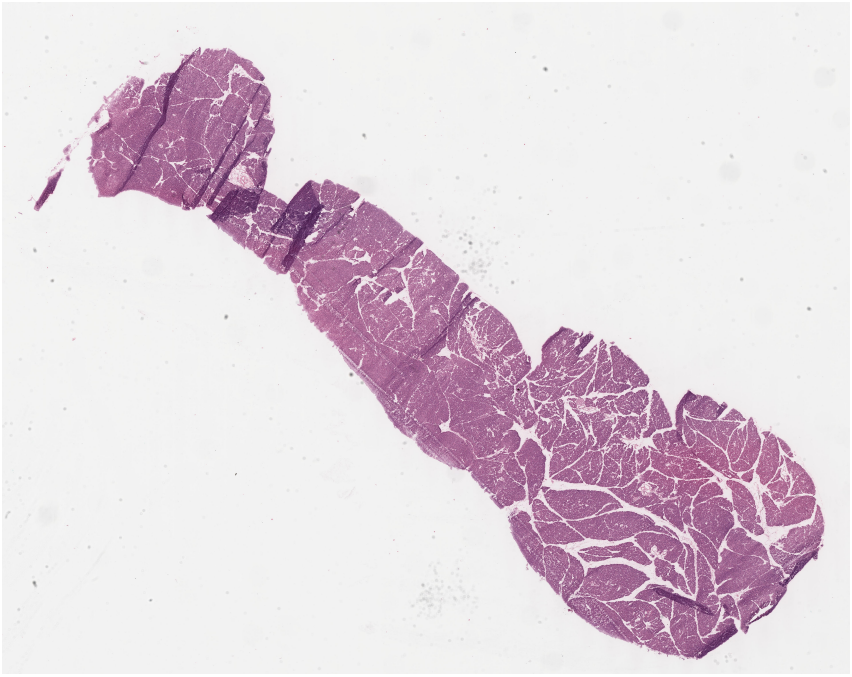


Figure 2.1: H&E Stained Mouse Pancreatic Tissue. Nucleus has purple colour, rest of the cell has pink colour.

magnified section of the same H&E staining, showcasing an islet. It highlights the characteristic circular arrangement of islets in rodents and the distinct cell shape compared to the surrounding tissue.

In mice, islets consist of an inner core housing insulin-producing β-cells and an outer mantle where glucagon-producing α-cells reside (Figure 2.2) [15, 54]. Besides insulin and glucagon, islets secrete other hormones such as somatostatin (δ-cells), ghrelin, and polypeptide P (PP-cells).

Pancreatic research predominantly focuses on rodent models. However, directly extrapolating knowledge gained from rodents to the human system may lead to inaccuracies. Several differences exist between human and rodent pancreas structure and function [18]. Firstly, humans lack a distinct outer periphery in their islet organization, unlike rodents [18]. Secondly, rodents possess two types of insulin encoded by separate genes, whereas humans have only one [49]. When interpreting findings from rodent pancreas tissue analysis for human relevance, it's crucial to consider both known and potential unknown similarities and differences between the two species.

2.1.2 Insulin

Insulin plays a critical role in regulating glucose metabolism, and its dysfunction is a defining characteristic of type 2 diabetes [25]. In mammals, insulin is primarily produced by β cells within the Islets of Langerhans [20]. In humans, the insulin hormone is encoded by the INS gene located on chromosome 11 (Figure 2.3a) [56]. However, rodents possess two distinct insulin genes: $Ins1$ (Figure 2.3b) and $Ins2$ (Figure 2.3c) [49].

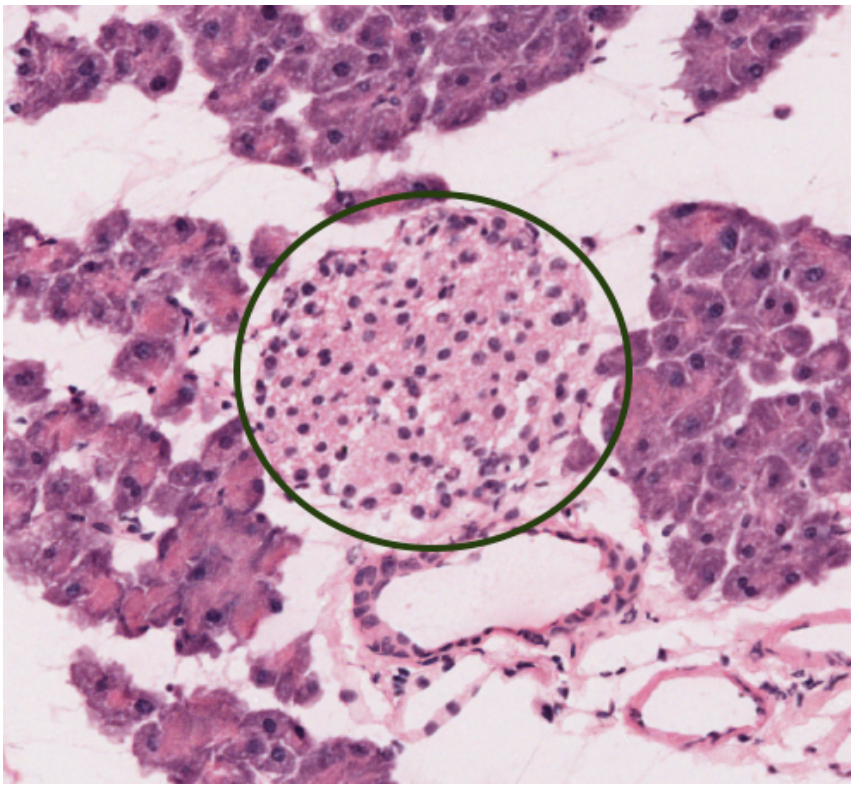
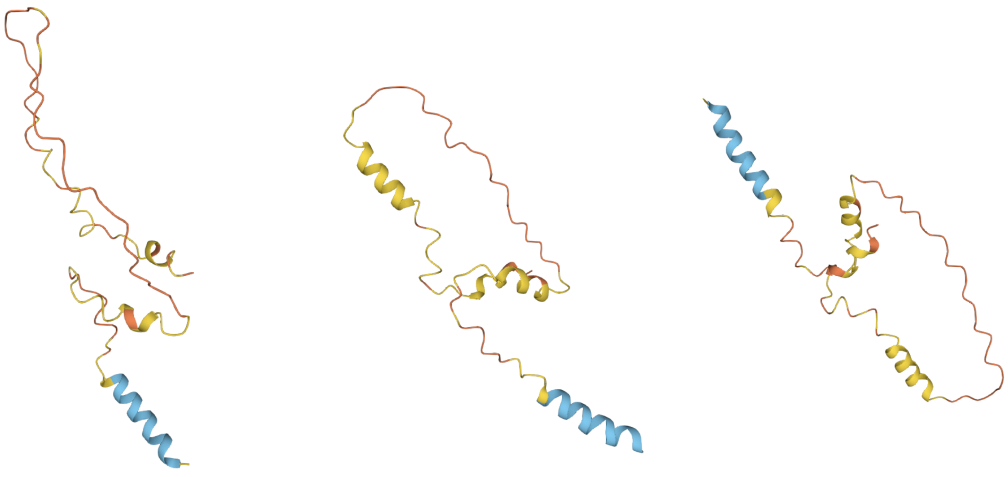


Figure 2.2: Hematoxylin and Eosin (H&E) Stained Mouse Pancreatic Tissue. Highlighting a close-up view of an islet of Langerhans, indicated by its characteristic circular shape.



(a) Human INS **(b)** Mouse - *Ins*1 **(c)** Mouse - *Ins*2

Figure 2.3: Structure of Insulin [7]. Different insulin structures are observed.

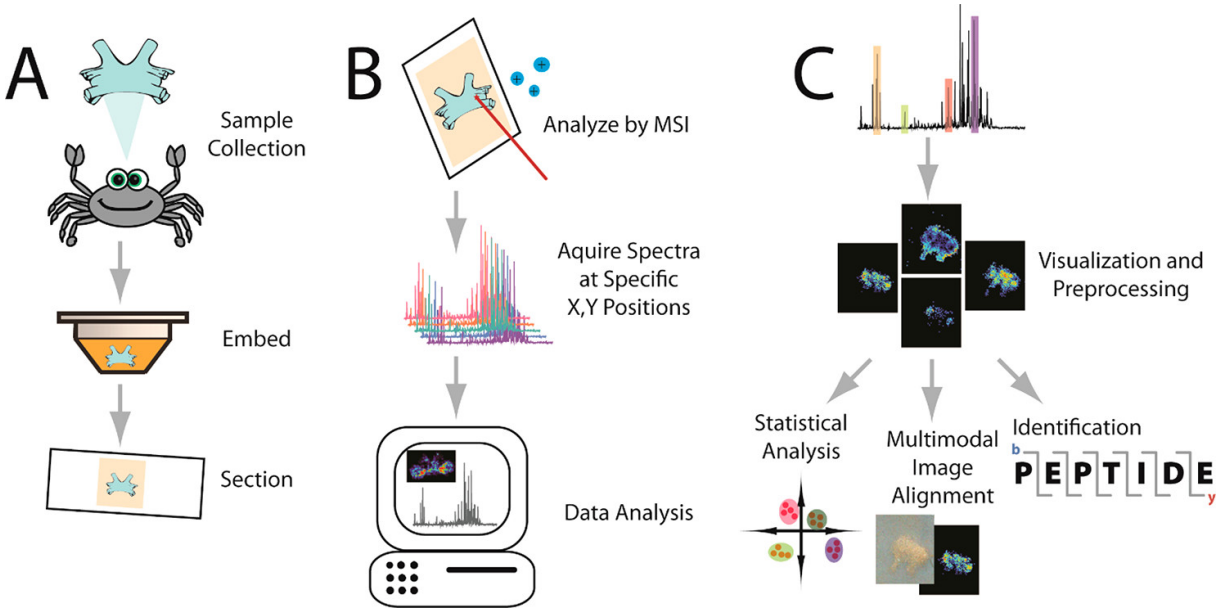


Figure 2.4: Visual Workflow for MSI Analysis. (A) Sample preparation involves embedding the tissue in a supporting medium or freezing it for sectioning onto slides. Additional processing, such as enzyme application or matrix treatment, may be performed depending on the molecular species of interest. (B) Sample analysis includes acquiring a spectrum at each (x,y) grid point on the tissue using laser or non-laser methods. (C) Data processing involves preprocessing (e.g., baseline correction) and visualizing the distribution of selected molecules. Identification of m/z values and statistical analysis can then be performed, along with image coregistration with other modalities. Reprinted with permission from [9], Copyright 2018, American Chemical Society.

2.2 Mass Spectrometry Imaging

MSI is a label-free technique utilized to detect the distribution of various molecules, including proteins [43], drugs [53], peptides [8], and other compounds [5]. It employs the fundamental principles of MS to generate a dataset resembling an image, consisting of a series of m/z vectors associated with specific (x, y) pixel coordinates, representing the relative intensity of a particular molecule [4].

While MSI is precise, certain aspects of the technique may introduce unwanted artefact-like peaks, resulting in multiple m/z ratios for a single molecule and a Gaussian-like distribution instead of a single ratio [16, 57]. An overview of MSI analysis is depicted in Figure 2.4.

2.2.1 The Basics of Mass Spectrometry

In MS [34, 44], a molecular profile of a sample is obtained through the ionization of molecules mixed with a solvent or matrix. Charged ions then traverse through an electromagnetic field where they are separated based on their mass-to-charge ratio (m/z), expressed in Dalton (Da). The most prevalent techniques for this purpose are matrix-assisted laser desorption/ionization (MALDI) [1, 2] and electrospray ionization (ESI). Ad-

vancements in speed, sensitivity, and chemical scope have enabled the analysis of larger biomolecules such as proteins, lipids, and peptides, rendering MS one of the primary technologies in many biomedical laboratories.

As preparation step, samples are mixed with an energy-absorbing chemical matrix, typically a small organic compound that crystallizes upon drying (hence the term 'matrix-assisted'). Then, the MALDI-TOF process consists of two phases. Firstly, the mixed sample undergoes laser irradiation, leading the matrix to absorb energy. Consequently, sample molecules vaporize and ionize simultaneously. The matrix plays a crucial role in this phase, absorbing laser energy at the appropriate wavelength and facilitating analyte ionization without molecular fragmentation. In the second phase, an electric field of known strength accelerates charged particles into a time-of-flight (TOF) tube. Molecules then reach a detector within a few nanoseconds after ionization. The flight time of ions to the detector depends on their mass and charge, enabling the differentiation of molecule types based on their m/z ratio. The resultant mass spectrum graph illustrates intensity, representing the number of detected ions, versus m/z ratio.

While MS offers excellent molecular coverage and sensitivity, it lacks spatial abundance information within the sample. Understanding the localization of different molecules in a biopsy is crucial for comprehending the behaviour of tumours and other cells within their microenvironment.

2.2.2 Mass Spectrometry Imaging (MSI)

In a standard mass spectrometry (MS) setting, spatial context is absent because samples are homogenized before analysis. MSI extends standard MS by measuring both spatial and molecular distributions, applying MS at multiple locations throughout the sample without the need for sample homogenization [4].

Prior to MSI analysis, rigorous sample preparation is crucial (Figure 2.4). To minimize batch effects and other artefact-like peaks, samples must be collected, frozen, and sectioned according to protocol, enabling comparable research between different experiments and across labs [23, 59].

Next, the tissue is analyzed by performing MS across the sample in a grid-like fashion (Figure 2.4). The distance between subsequent pixels in the grid determines the spatial resolution of the experiment. At each pixel, a laser beam or another ablation technique ionizes molecules and measures their mass-to-charge ratio (m/z). This results in a mass spectrum per pixel, depicting the intensities of the full m/z range measured.

Using the m/z spectra obtained for each pixel, heat maps or ion images can be generated, displaying the spatial distribution of a specific m/z value throughout the tissue.

To further identify a molecule, its detected mass is compared to an atlas of known molecules or a subsequent MS (tandem MS or MS/MS) experiment is performed [26, 31]. By breaking down a protein into its amino acid building blocks, its sequence can be derived. The protein's identification is then determined by matching this sequence to

existing libraries, such as UniProt[7].

2.2.3 MSI Components

The fundamental components of a mass spectrometer include the ion source, the mass analyzer, and the detector. Various technologies exist for each component, resulting in a wide range of possible MSI equipment configurations, each offering slightly different properties in terms of speed, sensitivity, and chemical scope.

Ion Source

The ion source converts a portion of the molecules present in the sample into ions, which are then directed through the mass analyzer. A diverse array of ionization techniques exists, each suitable for different applications. The choice of ionization technique depends heavily on priorities such as spatial resolution, mass range, or sensitivity. Dominant types of ion sources include MALDI [39], secondary ion mass spectrometry (SIMS) [19], and desorption electrospray ionization (DESI) [29].

Mass Analyzer

The mass analyzer, situated between the ion source and the detector, is responsible for separating different molecules based on their mass-to-charge ratio (m/z). Various technologies are available for mass analyzers, with common types including TOF analyzers [45], quadrupole ion traps [35], and Fourier transform ion cyclotron resonance (FT-ICR) [37].

Detector

The detector is responsible for detecting ions sent through the mass analyzer. To obtain spatial information, the detector records the exact (x, y) coordinates of the detected ions [39].

2.2.4 MSI Data

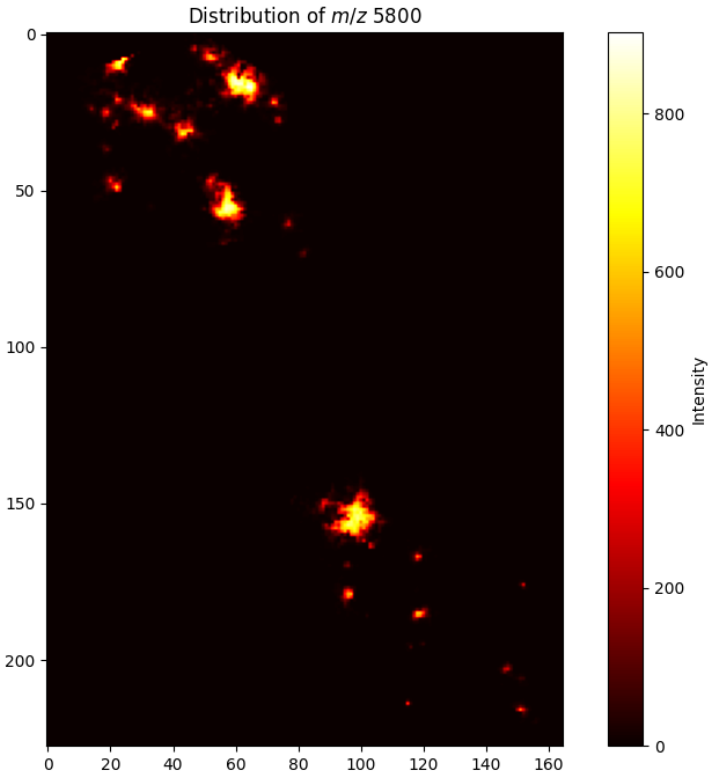
As previously mentioned, an MSI experiment involves acquiring mass spectra at various positions in a tissue section. These positions are determined by a virtual grid laid out over the tissue, corresponding to the pixels in the final image. Depending on the experimental goals, a trade-off is often made between spatial resolution and mass resolution to ensure feasible measurement times. One can opt for high spatial resolution by measuring a large number of pixels, thereby limiting the mass spectrum range at each pixel and reducing the measurement time per pixel. Alternatively, one can choose to measure a lower number of pixels, resulting in a broader mass spectrum per pixel and a longer measurement time per pixel.

Due to the nature of an MSI experiment, MSI data sets have three key axes: the spatial axes (x and y), indicating the pixel's position, and the spectral axis (m/z), containing the pixel's associated mass spectrum. Thus, an MSI experiment yields an array with the shape $[mz \times N \times M]$, where mz represents the number of m/z ratios, N represents the

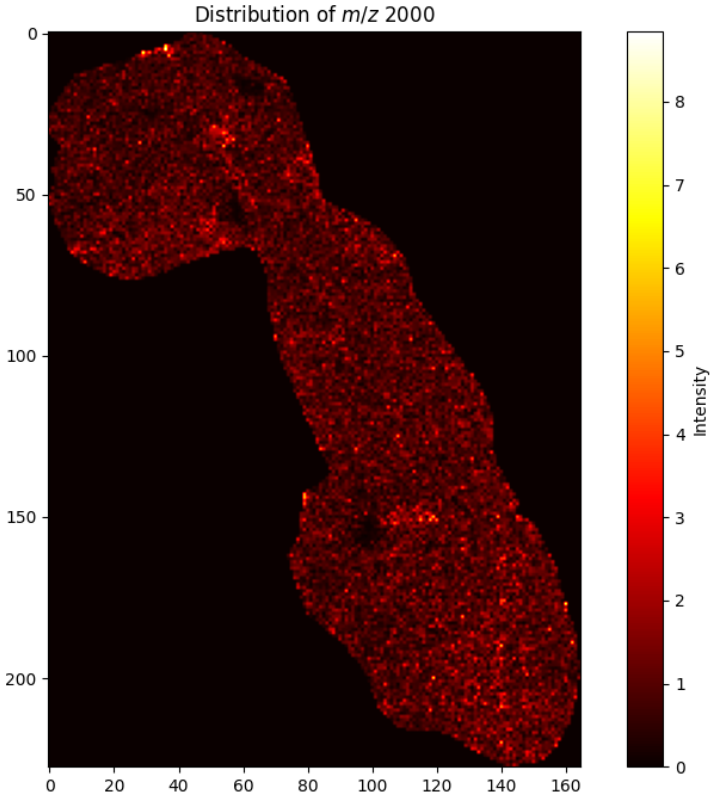
x-coordinates, and M represents the y-coordinates.

For our specific data, D, we have 14,000 different m/z ratios, 165 x-coordinates, and 228 y-coordinates, represented as $D = [mz \times N \times M] = [14000 \times 165 \times 228]$. Ion images can be produced for different m/z ratios to visualize the data (see Figure 2.5).

$$IonImage_k = D_k = [mz = k \times N \times M] \tag{2.1}$$



(a) Ion Image of $m/z = 5800$



(b) Ion Image of $m/z = 2000$

Figure 2.5: Ion Images of Different m/z Ratios. (a) $m/z = 5800$, (b) $m/z = 2000$.

Chapter 3

Constructing the Map of the Islets of Langerhans

This chapter is written by Serkan Shentyurk.

There is a structural organisation within biological tissues [3] and pancreatic tissue is no exception. There are different compartments responsible for different tasks, but the most important compartment for this dissertation is the Islets of Langerhans which produces and stores insulin [15]. Therefore, insulin, and the insulin-related peaks, can only be found within these islets in pancreatic tissue.

In this dissertation, we aim to develop a method to identify the insulin-related peaks. Since insulin is mainly found within the islets, we anticipate that insulin-related peaks will be significantly more prevalent within the islets compared to the rest of the tissue. Therefore, it is essential for us to map the Islets of Langerhans to identify the insulin-related peaks spatially.

This chapter initiates the identification of the islet map using an edge detection method. We convolve the insulin ion image with a kernel to obtain rough borders of the islets. Given the circular organization of the Islets of Langerhans (Figure 2.2) [15], we employ various algorithms to transform the output of edge detection into an initial map of the Islets. Subsequently, we apply different clustering algorithms to refine and finalize the map of the Islets.

3.1 Detection of the Edges of the Islets of Langerhans

Insulin is primarily expressed and stored within the Islets of Langerhans [15], making its distribution useful for identifying the map of these islets. Therefore, we hypothesise that if a rough mapping of insulin distribution is identified, it can be used to construct the map of the islets. We start with applying an edge detection algorithm to the insulin ion image D_{insulin}:

$$D_{\text{insulin}} = [Z = m/z_{\text{insulin}} \times N \times M]. \tag{3.1}$$

Firstly, we determine the value of m/z_{insulin}. By manually selecting a pixel within the islets using a microscope image (Figures 2.1 and 2.2), we obtain a vector V_{x_1,y_1} corresponding

to the intensities of different m/z ratios at a particular coordinate (x_1, y_1).

$$V_{x_1,y_1} = [mz \times N = x_1 \times M = y_1]. \tag{3.2}$$

Plotting the intensities of m/z ratios from vector V_{x_1,y_1}, we observe distinct peaks corresponding to the two insulin genes present in mice [49] and we select the values $m/z = 5800$ for $Ins2$ and $m/z = 5804$ for $Ins1$. Despite intensity variations, these peaks are co-localised, thus picking only one of them is sufficient to construct the map of the islets. We have chosen $m/z = 5800$ as the insulin of interest and obtain specific ion image data D_{insulin} (Figure 3.1).

$$D_{\text{insulin}} = [mz = 5800 \times N \times M]. \tag{3.3}$$

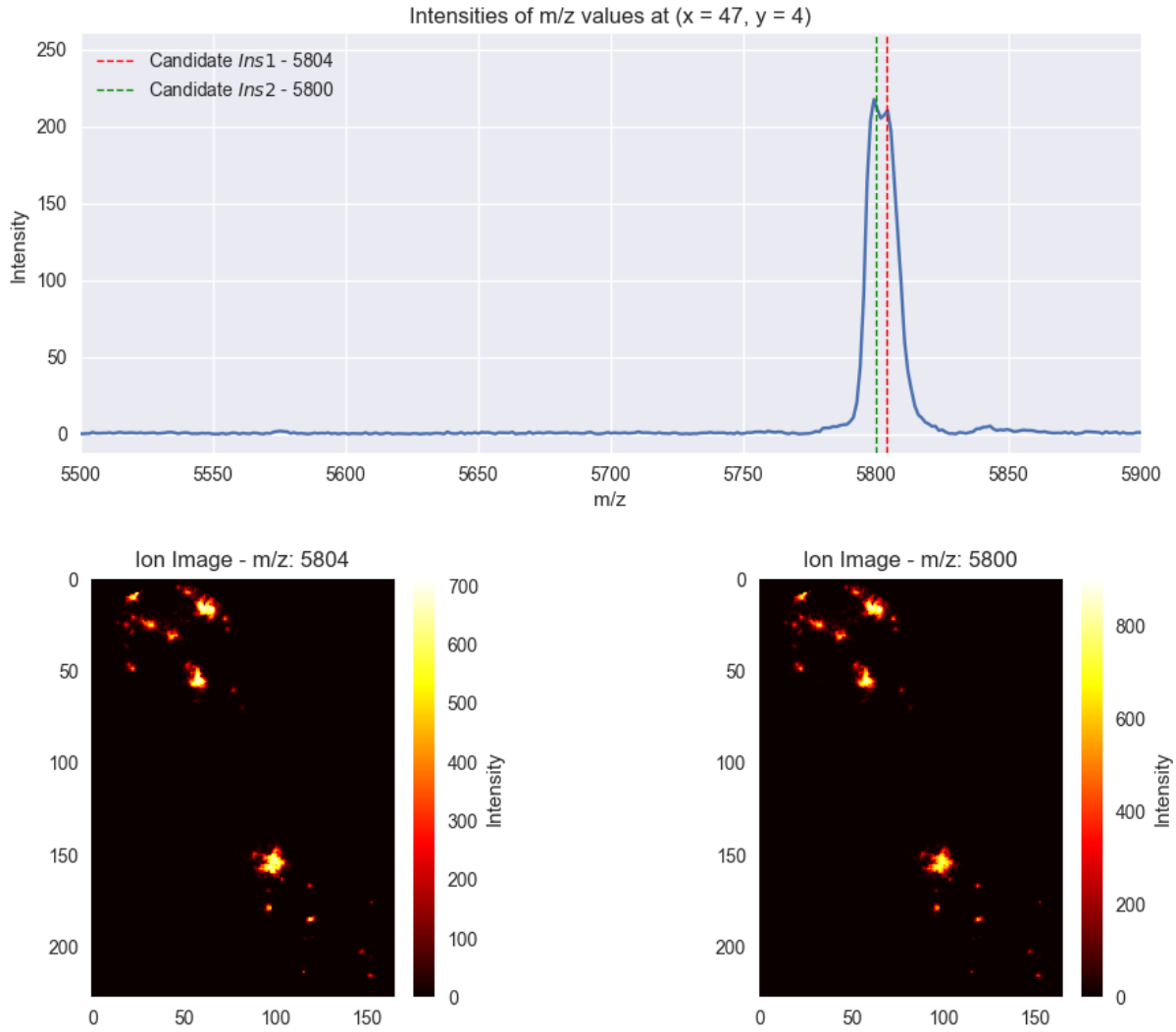


Figure 3.1: Determination of m/z ratios of $Ins1$ and $Ins2$. m/z intensities of a pixel within the Islet of Langerhans (top) and the corresponding ion image of the candidate insulin m/z ratios 5800 (bottom left), and 5804 (bottom right).

The quality of the detected edges can affect the subsequent analysis. It is a common practice to apply a Gaussian filter F to reduce the noise and high-frequency components by introducing a controlled blur as a pre-processing step [6]. In addition, since the edge

detection algorithms generally rely on estimating image gradients, it is important to have a stabilised gradient to have a robust edge detection process.

$$F(x,y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right). \tag{3.4}$$

Therefore, we start our analysis by applying a Gaussian filter with different standard deviations F_σ to the insulin ion image D_{insulin}. We apply different σ values of 1, 2, and 4, however, the final map of the islets remains unchanged. Therefore, we select the default σ value: $sigma = 2$.

$$D^{\text{smooth}}_{\text{insulin}} = F_{\sigma=2} * D_{\text{insulin}}. \tag{3.5}$$

Subsequently, we apply a Sobel filter by convolving Sobel kernels S_x and S_y over the image in the x and y directions to detect the edges, yielding edge magnitude matrices G_x and G_y.

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \tag{3.6}$$

$$S_y = \begin{bmatrix} -1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}, \tag{3.7}$$

$$G_x = S_x * D^{\text{smooth}}_{\text{insulin}}, \tag{3.8}$$

$$G_y = S_y * D^{\text{smooth}}_{\text{insulin}}. \tag{3.9}$$

Combining these matrices provides the edge magnitude matrix $G_{\text{edges}} = [N \times M]$, which will be used in constructing the map of the Islets of Langerhans (Figure 3.2).

$$G_{\text{edges}} = \sqrt{G_x^2 + G_y^2}. \tag{3.10}$$

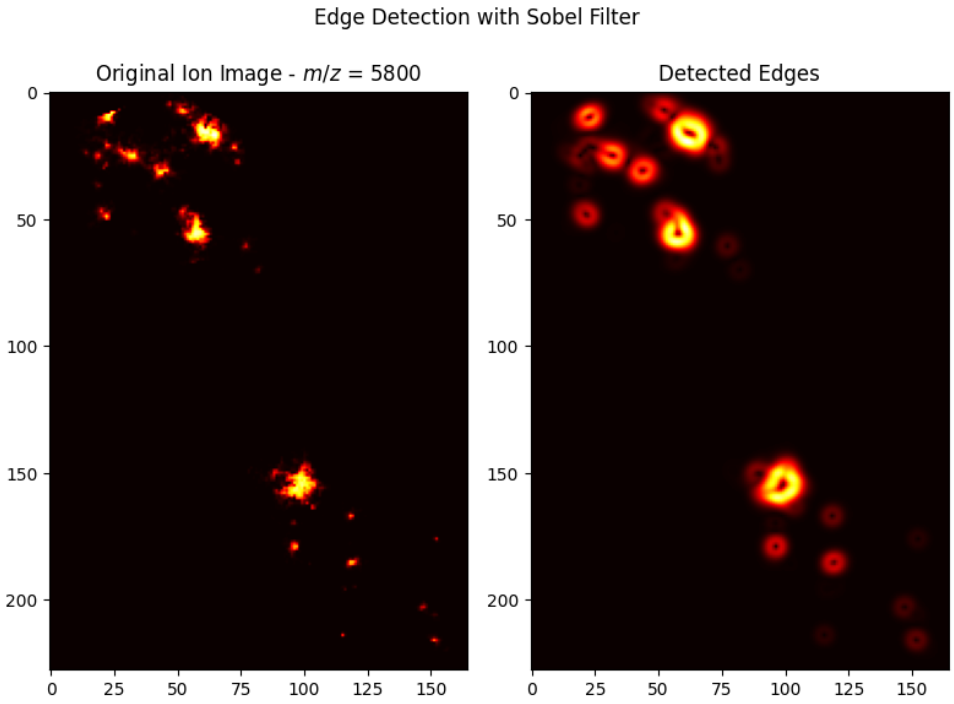


Figure 3.2: Edge Detection with Sobel Filter. The distribution of $Ins2$ and the edge magnitudes detected by Sobel Filter.

3.2 Initial Map of the Islets of Langerhans

The circular shape of the Islets of Langerhans suggests that relying solely on the edges of insulin (Chapter 3.1) may be overly restrictive. In this section, we aim to create a broader and less restrictive map by applying binary dilation to the magnitude matrix G_{edges}.

Binary Dilation

Binary dilation, an important technique in image processing, enhances or modifies shapes within binary images. It convolves a binary image with a kernel or structuring element and for each pixel in the input image, the pixel is set to 'on' if the kernel overlaps with any part of the 'on' in the input image. Otherwise, the output pixel remains 'off' [47]. In our case, the detected edges are treated as 'on' pixels and others as 'off' pixels. As a result, binary dilation expands the boundaries of the edges and makes the region larger.

Let A be the binary image and B be the structuring element. The dilation of A by B, denoted as $A \oplus B$, is defined as:

$$(A \oplus B)(x, y) = \bigcup_{(i,j) \in B} A(x - i, y - j). \tag{3.11}$$

where (x, y) are the coordinates of a pixel in the dilated image, and function $\bigcup$ is the union function which combines all sets into one set.

We begin by binarising the edge magnitude matrix G_{edges} to obtain $G_{\text{edges}}^{\text{binary}}$:

$$G_{edges}^{binary}(x,y) = \begin{cases} 1 & \text{for } G_{edges}(x,y) > 0, \\ 0 & \text{else .} \end{cases} \tag{3.12}$$

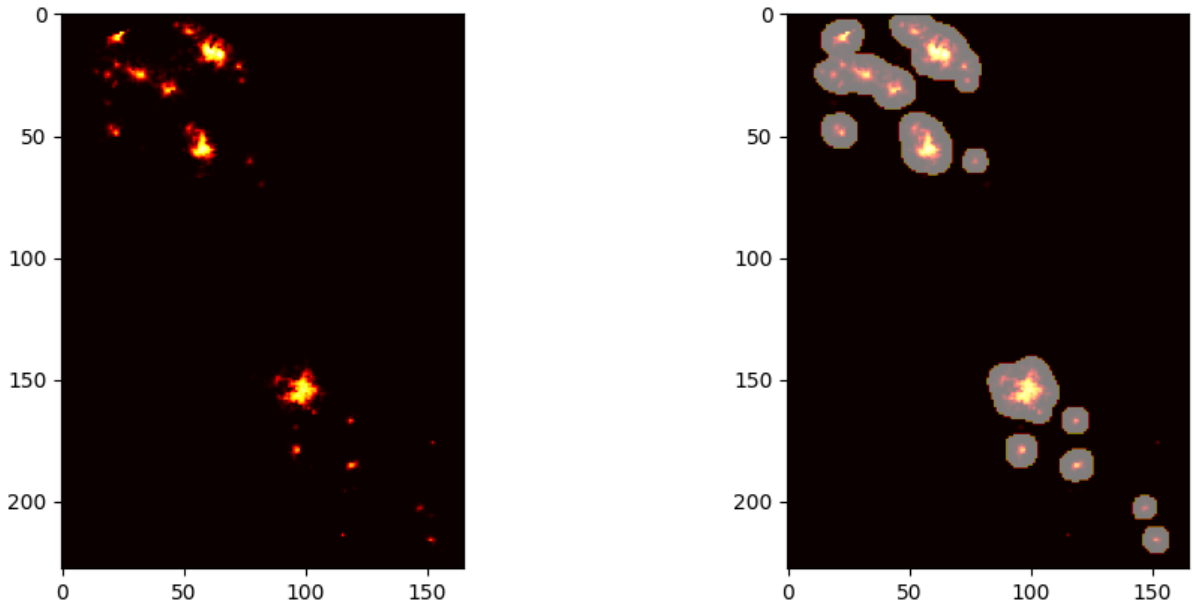
We use a default 3x3 square as the structuring element:

$$B = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{3.13}$$

The resulting equation is:

$$G = (G_{edges}^{binary} \oplus B)(x,y). \tag{3.14}$$

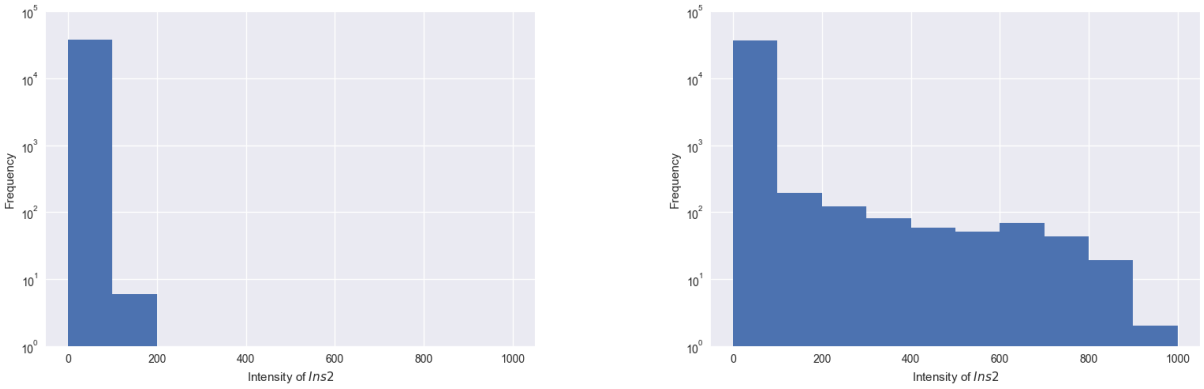
Matrix G is of shape $[N \times M]$, containing 1s and 0s, where 1 corresponds to the Islets of Langerhans and 0 corresponds to areas outside the islets (Figure 3.3).



(a) Original Ion Image - $m/z = 5800$ **(b)** Binary Dilation Results

Figure 3.3: Initial Mapping of the Islets of Langerhans. (a) The ion image of $Ins2$ - $m/z = 5800$, (b) The initial mapping detected by the binary dilation algorithm.

Binary dilation produces a relatively tolerant mapping (Figure 3.3b): the candidate area of the islets contains not only the pixels that have high insulin intensity but also the ones with low insulin intensity. To address this issue, we analyse the intensity distribution of $Ins2$ in two areas of the binary dilation result (Figure 3.4).



(a) $Ins2$ Intensity in Non-Islets

(b) $Ins2$ Intensity in Islets

Figure 3.4: Intensities of $Ins2$ in the Candidate Areas. (a) $Ins2$ distribution within the non-Islets, (b) $Ins2$ distribution within the Islets.

Quantitative analysis reproduces similar results to those of our visual inspection. In non-islet areas (Figure 3.4a), $Ins2$ is not found in high intensities, whereas the candidate area of the islets (Figure 3.4b) exhibits high intensities. However, the candidate area of the islets also contains a high frequency of low $Ins2$ intensity pixels, indicating the presence of non-islet pixels. These results show that our algorithm is able to group the islets of Langerhans; however, the group also contains pixels that do not correspond to the islets. Further analysis of the candidate area is needed to cluster different pixels within the area and isolate the islets of Langerhans, hence allowing for the construction of the map of the Islets of Langerhans.

3.3 Final Map of the Islets of Langerhans

The candidate areas of the islets determined in Section 3.2 (Figure 3.4 and Figure 3.3b) encompass the islets and non-islet cells. In this section, we aim to classify the candidate areas of the islets to establish a well-defined mapping of the Islets of Langerhans. We plan to tackle this problem by applying a clustering algorithm to not only the ion image data but also using the intensities of other m/z ratios. We explore three different clustering algorithms: HDBSCAN [12], Fuzzy K [46], and K-Means [27].

We conceptualize our data D as an $N \times M$ image containing Z channels, m/z ratios in our case. We want to cluster the data based on the intensities of the m/z ratios and the pixel location. We employ Algorithm 1 to prepare the data for clustering algorithms and create 2 new columns that store x and y coordinates. This algorithm reshapes the data into a 2D matrix $D_{\text{clustering}}$ with N x M rows and Z + 2 columns to include the pixel positions (x, y) and the mass spectra per pixel.

$$D_{\text{clustering}} = [Z = 5800 + 2 \times (N \times M)]. \tag{3.15}$$

Algorithm 1: Prepare Data for Clustering Algorithms

Function `prepareClustering(`D`):`
- $shape_D \leftarrow$ Shape of D
- $D_{clustering} \leftarrow$ 2D array of zeros with dimensions $(shape_D[1] * shape_D[2], shape_D[0] + 2)$
- **for** $mz \leftarrow 0$ **to** $shape_D[0] - 1$ **do**
  - **for** $x \leftarrow 0$ **to** $shape_D[1] - 1$ **do**
    - **for** $y \leftarrow 0$ **to** $shape_D[2] - 1$ **do**
      - $D_{clustering}[x * (shape_D[2]) + y, 0] \leftarrow x$
      - $D_{clustering}[x * (shape_D[2]) + y, 1] \leftarrow y$
      - $D_{clustering}[x * (shape_D[2]) + y, mz + 2] \leftarrow D[mz, x, y]$
    - **end**
  - **end**
- **end**
- **return** $D_{clustering}$

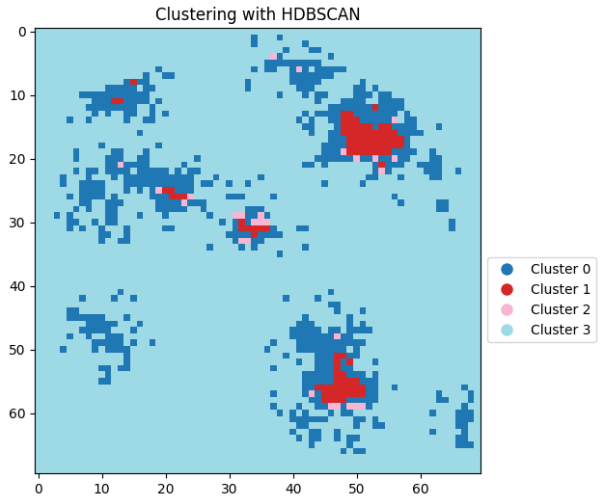
3.3.1 Clustering with HDBSCAN

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is an important unsupervised learning algorithm that employs the minimum spanning tree approach to construct a hierarchical representation of the data density [12]. Self-determination of the number of clusters makes HDBSCAN easier to use, however, in the case of an incorrect number of clusters, there is no intuitive way to change the outcome.

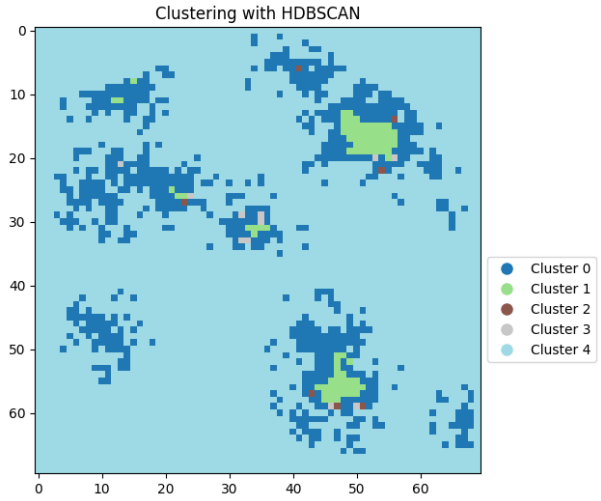
Algorithm 2: HDBSCAN Clustering Algorithm

Function `HDBSCAN_Clustering(`$D, min\_cluster\_size$`):`
- Construct the mutual reachability graph
- Compute the minimum spanning tree
- Construct the cluster hierarchy
- Select clusters from the hierarchy with given $min\_cluster\_size$
- Condense the selected clusters
- Label the data points with cluster labels
- **return** the cluster labels

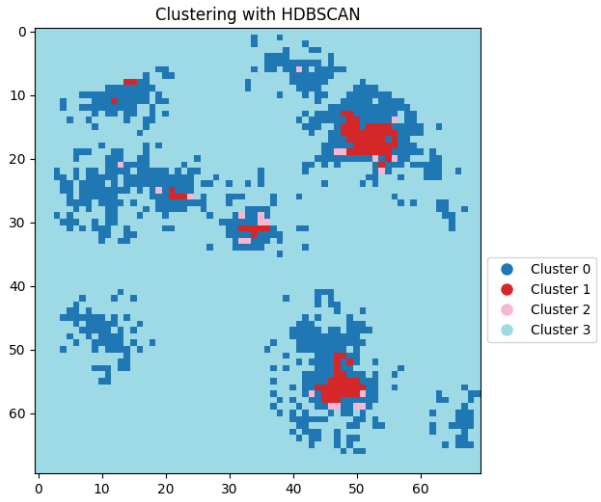
We apply the HDBSCAN algorithm, whose steps are briefly explained in Algorithm 2, with different minimum cluster size parameters (sizes 5, 6, and 7), and plot an image of the resulting clustering (Figure 3.5). Each colour represents a cluster that HDBSCAN predicted. Regardless of the minimum cluster size, the outputs of the algorithm do not align with the domain knowledge [15] - the islets are circular shaped (Figure 2.2), however, the algorithms are unable to capture this expected circular shape (Figure 3.5).



(a) HDBSCAN Clustering with $min\_cluster\_size = 5$



(b) HDBSCAN Clustering with $min\_cluster\_size = 6$



(c) HDBSCAN Clustering with $min\_cluster\_size = 7$

Figure 3.5: HDBSCAN Clustering Results. Different $min\_cluster\_size$ parameter is selected and applied: 5 (a), 6 (b), and 7 (c).

3.3.2 Clustering with Fuzzy K

Fuzzy K-Means offers a flexible approach to clustering by assigning membership values to data points which reflects the degree of association with a particular cluster [46]. The algorithm iteratively updates cluster centroids and membership values (Algorithm 3).

Algorithm 3: Fuzzy K-Means Clustering Algorithm

```
Procedure Fuzzy_K_Means_Clustering(D, k):
    Initialize cluster centroids randomly
    repeat
        for each data point d in D do
            for each cluster centroid c do
                Calculate membership degree μ(d, c)
            end
        end
        for each cluster centroid c do
            Calculate the new centroid as the weighted average of data points,
             considering their membership values
        end
    until Convergence or maximum iterations reached
    for each data point d in D do
        Assign d to the cluster with the highest membership degree
    end
    return cluster centroids and membership values
```

The fuzzy K-Means algorithm requires the number of clusters as a parameter. To determine the optimal number of clusters, the fuzzy partition coefficient (FPC) - which measures the degree of fuzziness within the clusters - is determined for different numbers of clusters. It is important to keep the domain knowledge in mind while evaluating the FPC.

To determine the number of clusters K, we apply Algorithm 3 with varying K and calculate the FPC (Figure 3.6). We expect to have 4 different clusters: The inner parts of the islets, the periphery of the islets, just outside of the islets, and the rest of the tissue. The FPC scores align with our expectations with an elbow at $K = 4$ (Figure 3.6). FPC further decreases where $K = 7$ and $K = 10$, but since our aim is not to determine the different structures within the pancreatic tissue but to locate the islets of Langerhans, $K = 4$ is picked as the number of clusters.

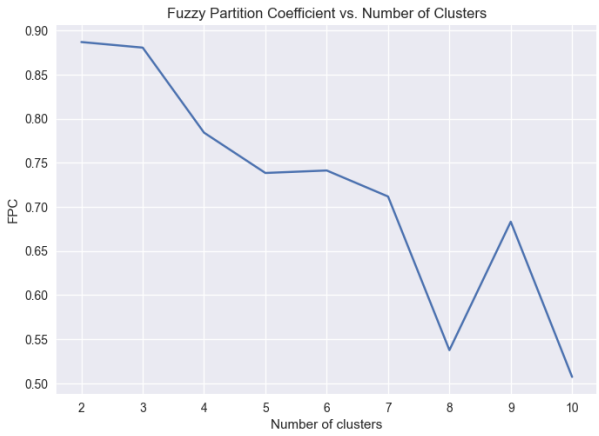


Figure 3.6: FPC Scores with Different Number of Clusters. Scores changes step by step.

We apply the Fuzzy K algorithm 3 with $K = 4$ and plot the class representation of the tissue where colours represent the clusters assigned by the algorithm and the distribution of $Ins2$ within the determined clusters (Figure 3.7). Cluster 0 represents the rest of the tissue and Cluster 1 represents the centre of the islets (Figure 3.7a). However, the difference between Cluster 2 and Cluster 3 is not apparent. In Figure 3.7a, they are generally co-located and the $Ins2$ has a similar distribution in both clusters (Figure 3.7b).

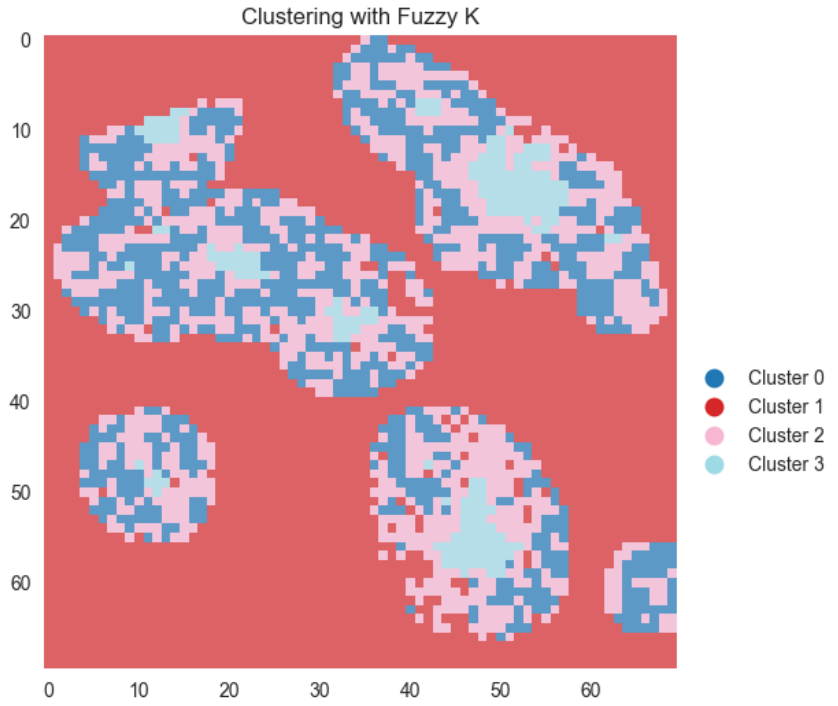
3.3.3 Clustering with K-Means

K-means clustering is a fundamental unsupervised learning method, widely used to cluster data into K different clusters by iteratively optimising the cluster centroids [27]. There are two main steps: assignment and update. In the assignment step, each data point is assigned to the nearest centroid based on distance metrics. Subsequently, in the update step, cluster centroids are recalculated as the mean of all data points assigned to the cluster. The algorithm is summarised in Algorithm 4.

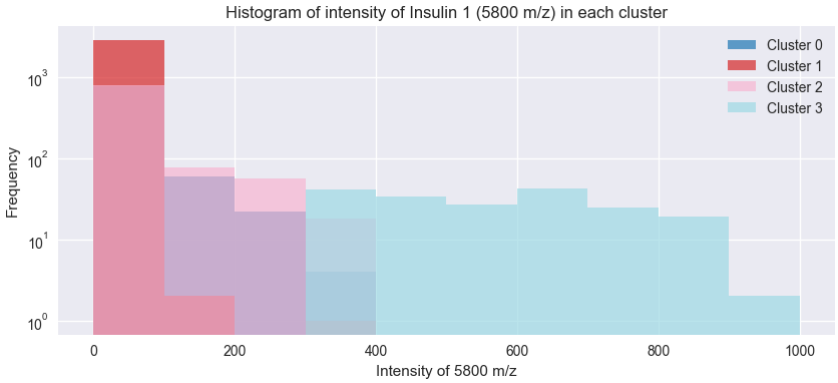
Algorithm 4: K-Means Clustering Algorithm

Procedure `KMeans`(D, k)**:**
- Initialize k cluster centroids randomly
- **repeat**
  - Assign each data point to the nearest centroid
  - Update cluster centroids as the mean of assigned data points
- **until** *convergence*
- **return** cluster centroids and labels

One common challenge in K-Means clustering is determining the optimal number of clusters K for a given dataset. The elbow method is used to overcome this challenge by running K-Means algorithm with different k values and plotting the within-cluster sum of squares (WCSS), which is a representation of the fit of the cluster, against the number of clusters [27]. As k increases, the WCSS typically decreases. However, there is a limited



(a) Clusters assigned by Fuzzy K



(b) Histogram of $Ins2$ within the clusters

Figure 3.7: Fuzzy K-Means Clustering Results. (a) The distribution of the clusters, (b) $Ins2$ distribution within the clusters.

decrease after a certain point, which is characterised by an elbow in the WCSS-k graph and the corresponding k value is selected as the optimal number of clusters.

To determine the number of clusters k, we apply Algorithm 4 with varying k values and calculate the WCSS (Figure 3.8). There is a clear elbow structure at $K = 3$ (Figure 3.8), however, since our previous knowledge of the structure of the tissue suggests that there are 4 different structures and FPC also indicates that there are 4 different clusters within the data (Figure 3.6), we pick $K = 4$.

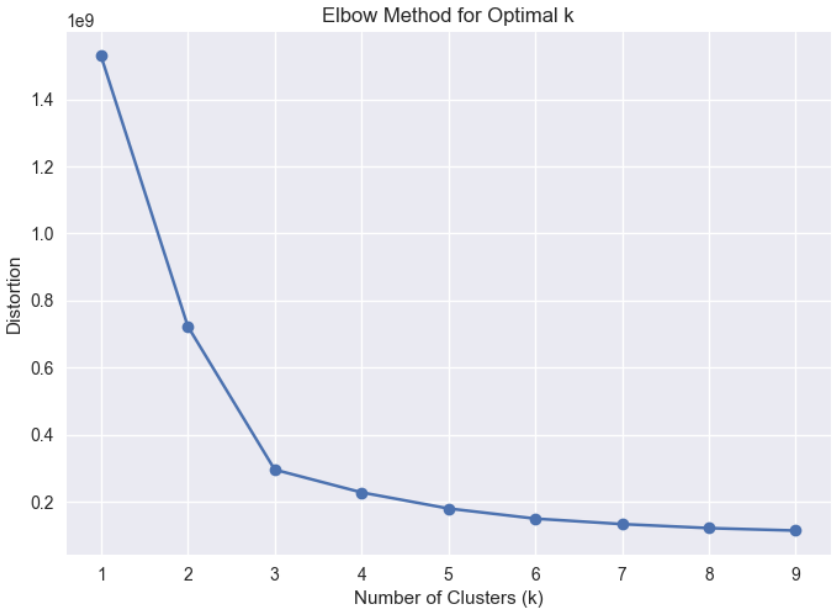
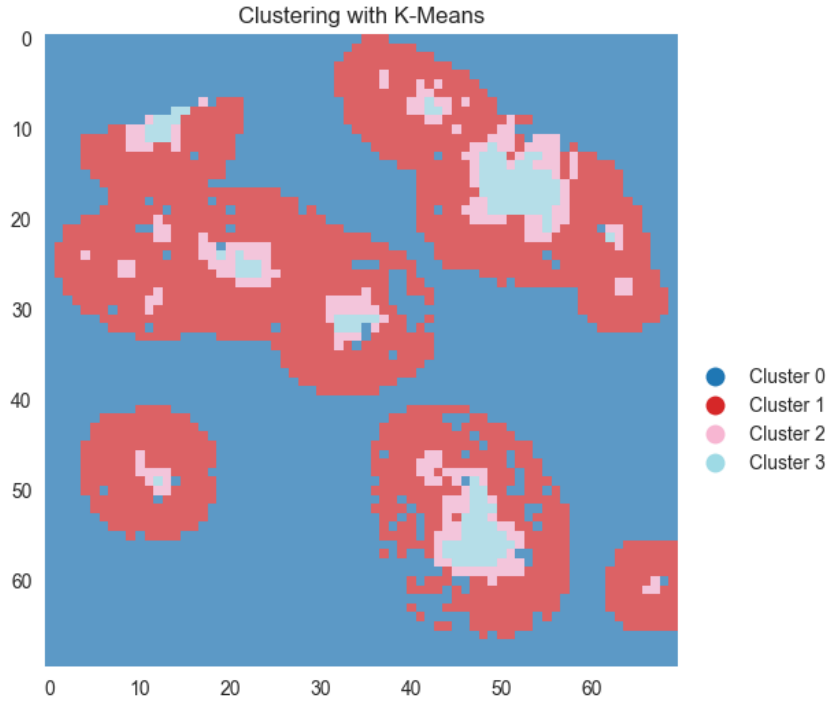
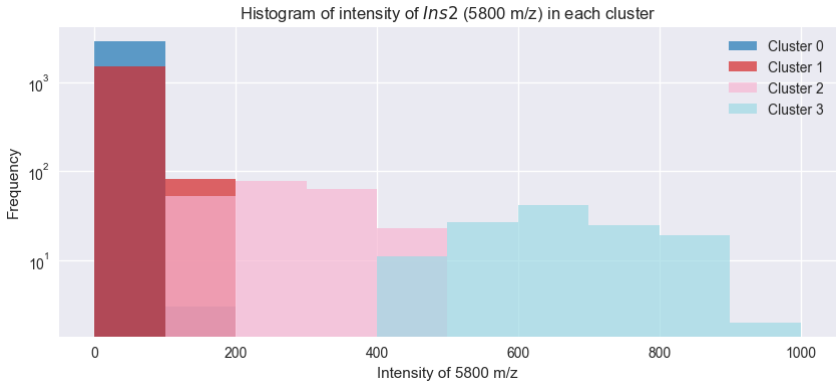


Figure 3.8: WCSS Scores with Different Number of Clusters k. The elbow point is seen at $k = 3$ or $k = 4$.

We apply the Algorithm 4 with $K = 4$ and plot the image where colours represent the clusters assigned by the algorithm and the distribution of $Ins2$ within the determined clusters (Figure 3.9). Cluster 0 represents the rest of the tissue, Cluster 1 represents the surrounding of the islets, Cluster 2 represents the outer periphery of the islets, and Cluster 3 represents the centre of the islets (Figure 3.9a). In other words, Cluster 2 and Cluster 3 can be used as a map of the islet of Langerhans. This structure can also be seen when the distribution of the $Ins2$ is checked (Figure 3.9b). $Ins2$ is mainly located at Cluster 3 and Cluster 2, there is almost no $Ins2$ in Cluster 0 and there is little to no $Ins2$ in Cluster 1 (Figure 3.9b). Since we have a good distribution of $Ins2$ and it also matches our prior knowledge, we will use the output of the K-means approach as our map of the Islets of Langerhans.



(a) Clusters assigned by K-Means



(b) Histogram of $Ins2$ within the clusters

Figure 3.9: K-Means Clustering Results. (a) The distribution of the clusters, (b) $Ins2$ distribution within the clusters.

3.3.4 Comparison of the Clustering Methods

In Chapter 3, we aimed to find a mapping of the islets of Langerhans. We started with an edge detection algorithm in Section 3.1, then constructed the initial mapping in Section 3.2. In Section 3.3, we applied different clustering algorithms to obtain the final mapping of the islets of Langerhans. Different algorithms produce different mappings, so it is important to choose the appropriate method for the final map of the islets of Langerhans.

We began by aligning the cluster numbers across methods: Cluster 3 contains the inner part of the islet, Cluster 2 contains the islets, Cluster 1 contains the periphery of the islets, and Cluster 0 contains the rest of the tissue (Figure 3.9a). Since insulin is produced

and stored in the islets [57], we expect Clusters 2 and 3 to have high insulin intensity, with peaks around an m/z ratio of 5800, while other clusters should have very low insulin intensity. To compare the intensities, we calculated the mean intensities of m/z ratios within each cluster for different methods (Figure 3.10). Each method performs similarly in Cluster 0 and Cluster 1. However, there is a noticeable change in the clustering of Cluster 2 and Cluster 3: clusters defined by K-means clustering show high insulin intensity in both the inner parts of the islets (Cluster 3) and the outer parts of the islets (Cluster 2).

The underlying difference in mean intensities between Cluster 2 and Cluster 3 might be due to the fact that the Fuzzy K-means and HDBSCAN algorithms classify non-islet pixels as islets as well, resulting in a decrease in the mean intensity of insulin. This can explain the mean intensity similarity between methods in Cluster 0 and Cluster 1.

The differences between the methods can be seen in the visualization of the clusters (Figures 3.5c, 3.7a, and 3.9a): the K-means algorithm produces results similar to anatomical findings, whereas the HDBSCAN and Fuzzy K-means algorithms fail to do so. The histograms of $Ins2$ within the clusters also indicate similar findings (Figures 3.7b and 3.9b). Therefore, we select the K-means algorithm to determine the final mapping and construct the map of the islets of Langerhans (Figure 3.11).

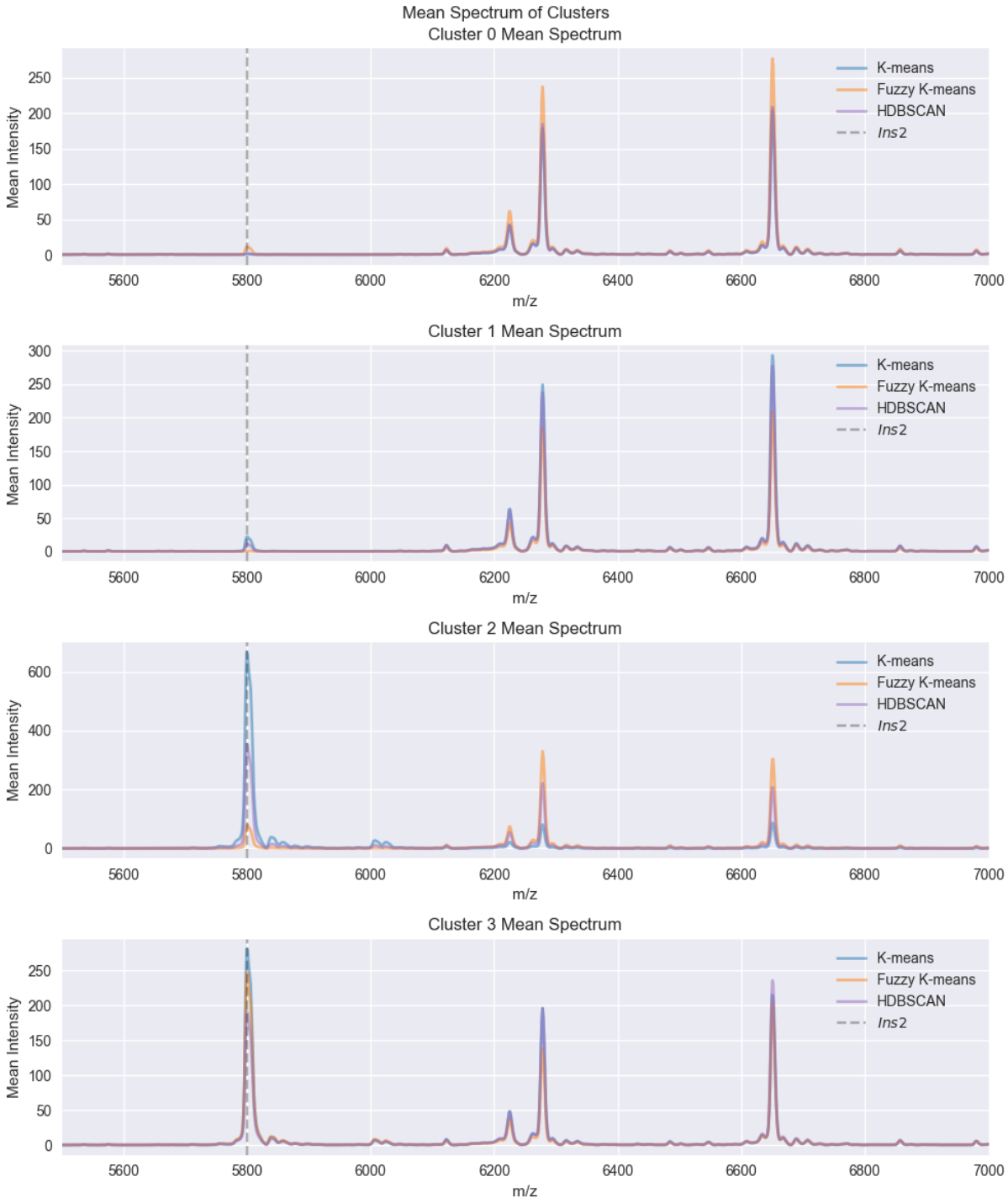


Figure 3.10: The Mean Intensities of m/z Ratios Across Clusters. Each graph shows the mean intensities of m/z ratios and colours represent different clustering algorithms.

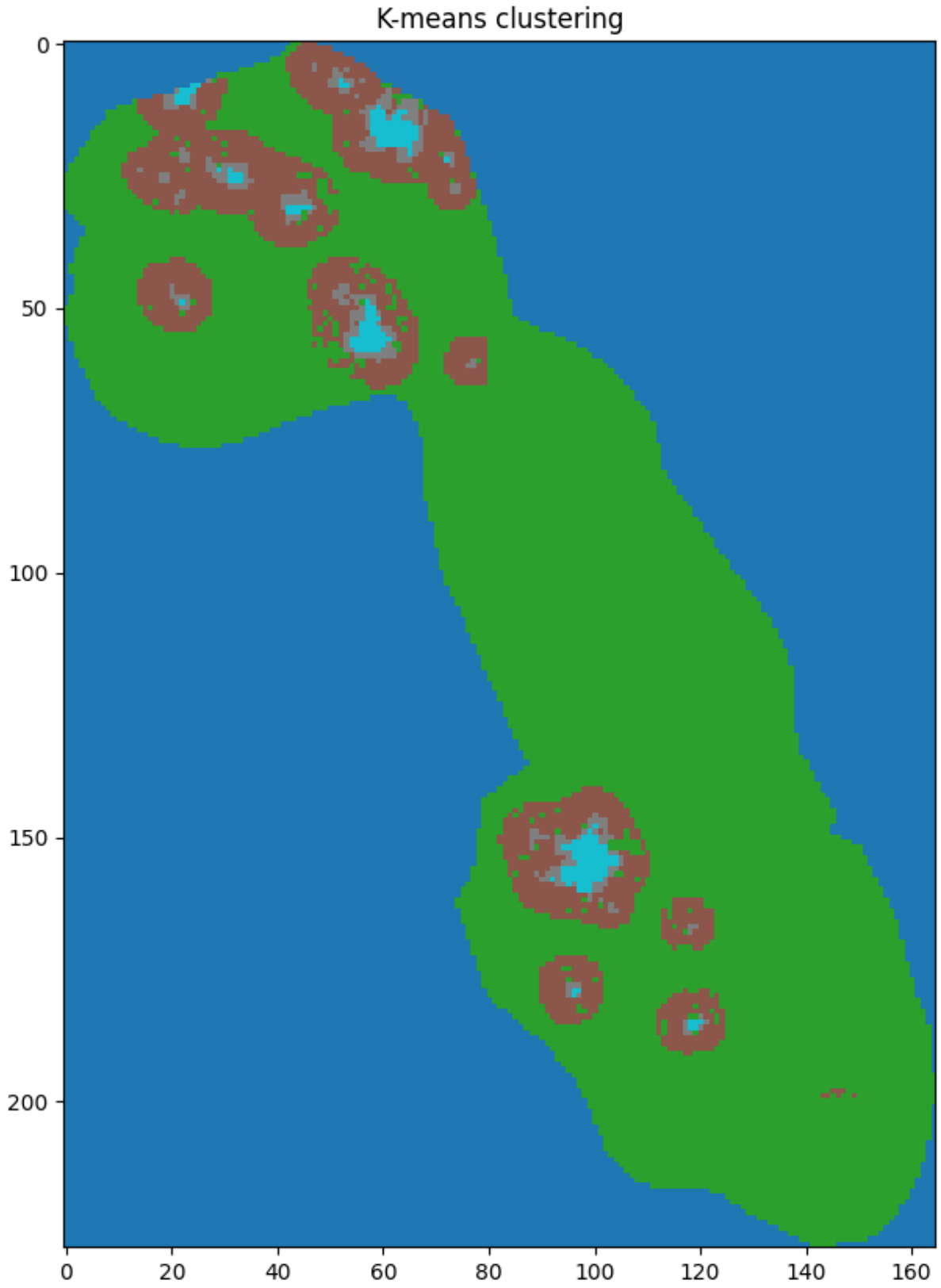


Figure 3.11: The Map of Islets of Langerhans. Cyan corresponds to the inner parts of islets, grey is the outer parts of the islets, brown is the surrounding the islets, and green is the rest of the tissue.

Chapter 4

Identifying the Insulin-related Peaks: Pearson Correlation Approach

This chapter is written by Serkan Shentyurk.

MSI, explained in Chapter 2.2, is a label-free technique utilized to detect the distribution of various molecules, including proteins [43], drugs [53], peptides [8], and other compounds [5].

However, MSI can exhibit artefact-like peaks due to the decomposition of unstable molecular ions during the ionization phase [16]. Consequently, additional peaks are observed in mass spectra that correspond to the same molecule or compound. Furthermore, ionized molecules with the same mass-to-charge (m/z) ratio may have different kinetic energies, and molecules can have various isotopes resulting in a Gaussian distribution around the exact m/z ratio of the molecule [57].

Since artefact-like peaks are unwanted byproducts of ionization, it is reasonable to assume that they should coexist with the compound of interest, insulin, in our case. Therefore, artefact-like peaks should exhibit a significant correlation with the source, insulin. In the first section of this chapter, we apply a Pearson correlation analysis to detect the m/z ratios that are highly correlated with insulin. Since the distribution of the compounds follows a Gaussian distribution [57], we expect the correlation values of m/z ratios with insulin to follow a Gaussian distribution as well. Therefore, in the second section of this chapter, we attempt to identify the mean of these Gaussian-like distributions using a peak detection algorithm and Gaussian Mixture Models (GMMs). We label the peaks of the GMMs as the candidate insulin-related peaks.

4.1 Correlation Analysis

Pearson correlation measures the strength and the direction of the linear relationship between the components [41]. The Pearson correlation coefficient, r, is calculated using the following formula:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}, \tag{4.1}$$

where:

- X_i and Y_i are individual data points,
- $\bar{X}$ and $\bar{Y}$ are the means of the two variables,
- n is the number of data points.

To apply Pearson correlation, we transform our dataset D, shaped $[mz \times x \times y]$, into the D_{corr} matrix shaped $[mz = 14000 \times x * y = 165 * 228 = 37620]$. We treat each pixel as a sample and each m/z ratio as a variable. Thus, when we apply Pearson correlation analysis, we obtain a matrix r shaped $[14000 \times 14000]$. Since we are only interested in the correlation of m/z ratios with $Ins2$, the correlation of m/z ratios with $Ins2$ is selected, resulting in r_{Ins2} shaped $[1 \times 14000]$.

To determine whether the correlation between a particular m/z ratio and $Ins2$ is significant, we hypothesize that the correlation between $Ins2$ and the m/z ratio with the least variance can be used as a baseline. Our rationale is that the m/z ratio with the least variance is found in all of the tissue and its presence is not affected by the presence of insulin. Our null and alternative hypotheses are:

H_0: The correlation between the m/z ratio X and $Ins2$ is the same with the correlation between the ratio that has the least variance and $Ins2$.
H_1: The correlation between the m/z ratio X and $Ins2$ is not the same with the correlation between $Ins2$ and the ratio m/z_{least} that has the least variance.

To determine the least varying m/z_{least}, we compare the variances of the m/z ratios and selected the m/z ratio of 20005 as the reference m/z ratio since it has the least variance, 0.073 (Figure 4.1), and the correlation between the m/z ratio of 20005 and $Ins2$ is 0.0263. Therefore, our hypothesis becomes:

H_0: The correlation between the m/z ratio X and $Ins2 = 0.0263$.
H_1: The correlation between the m/z ratio X and $Ins2 > 0.0263$.

We apply a one-tailed t-test with a significance level of $\alpha = 0.05$, select the m/z ratios that are significantly correlated, and form the 2-D matrix $D_{corr}^{significant}$, which contains 432 different m/z ratios and their correlation values with $Ins2$. The detected correlated values exhibit a Gaussian-like distribution as expected (Figure 4.2). Although this approach cannot entirely solve the Gaussian distribution problem, it significantly reduces the number of possible candidate insulin-related peaks from 14000 to 432.

4.2 Detection of the Candidate Peaks

Ionization of compounds may result in different kinetic energies, leading to a Gaussian distribution around the exact m/z ratio of the molecule [57]. We aim to identify the exact m/z ratios of insulin-related peaks and their corresponding Gaussian distribution.

In section 4.1, we discovered 432 m/z ratios that are significantly correlated with $Ins2$.

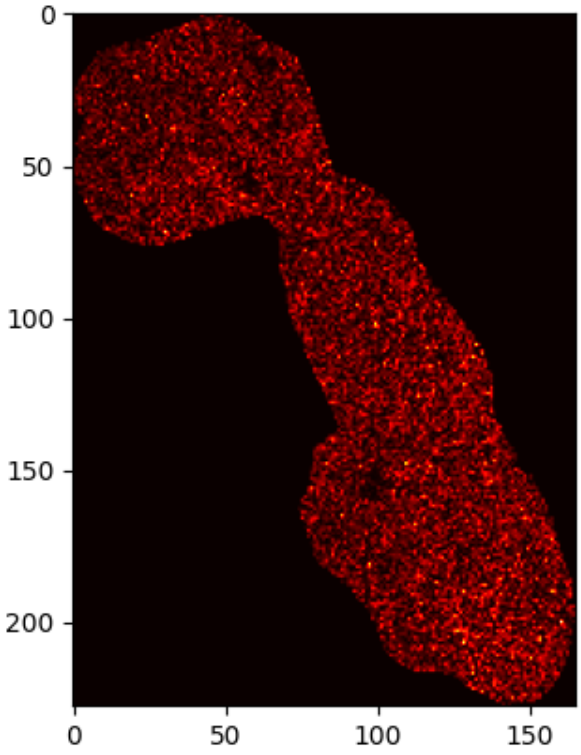


Figure 4.1: Ion Omage of $m/z = 20005$. The m/z ratio of 20005 has the least variance $\sigma^2 = 0.073$.

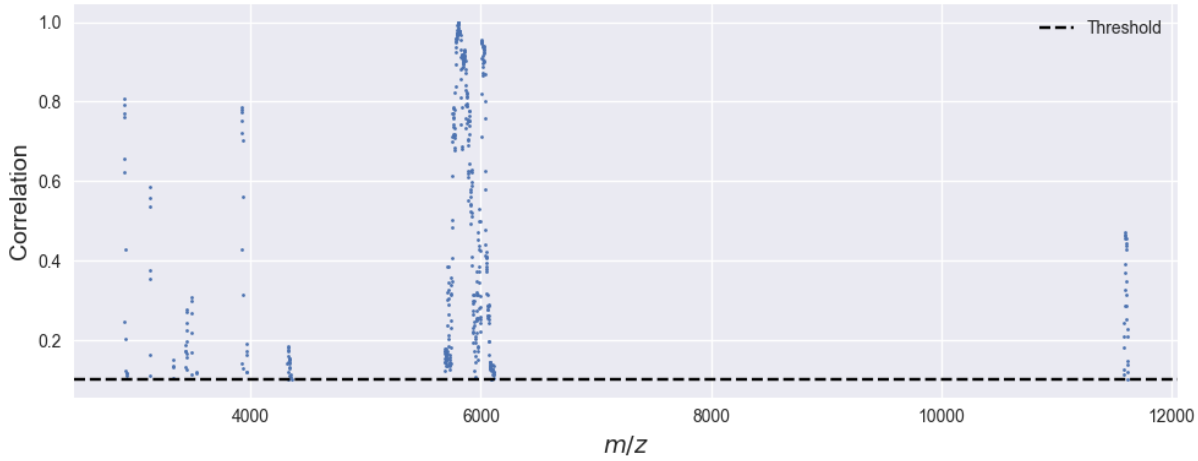


Figure 4.2: m/z Ratios Significantly Correlated with $Ins2$. Correlated m/z ratios are plotted against the correlation coefficient.

However, they still exhibit a Gaussian distribution (Figure 4.2). In this section, we attempt to identify the exact m/z ratios, or in other words, the candidate insulin-related peaks, by directly applying Gaussian Mixture Models (GMMs).

We start the analysis by dividing the correlated m/z ratio matrix $D_{corr}^{significant}$ into 3 clusters (Figure 4.3), as upon inspection, it was evident that there are 3 clusters present (Figure 4.2). This step is done to make the visualisation of the analysis easier.

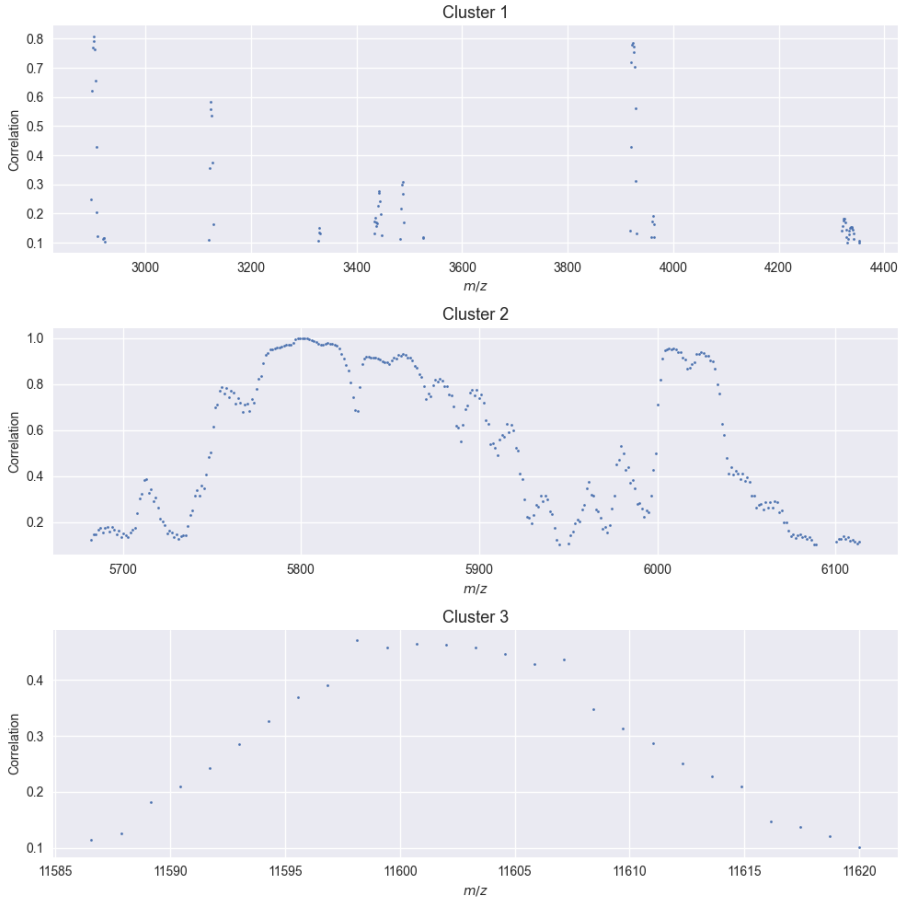


Figure 4.3: The Correlated m/z ratios with $Ins2$. m/z ratios are clustered into 3 groups.

4.2.1 Determining the Number of GMM Components

Gaussian Mixture Models (GMMs) are probabilistic models used to represent populations by fitting multiple Gaussian distributions [17]. They are particularly useful when the underlying data cannot be represented by a single Gaussian distribution but is coming from multiple sources.

The overall probability distribution is modeled as a weighted sum of multiple Gaussian distributions, also known as components or clusters. Each component represents a

subpopulation within the data. The overall probability distribution is represented as:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k), \tag{4.2}$$

where:

- $p(x)$ is the probability density function of the mixture model,
- K is the number of components,
- π_k is the mixing coefficient for the k-th component ($\sum_{k=1}^{K} \pi_k = 1$),
- $\mathcal{N}(x|\mu_k, \Sigma_k)$ is the Gaussian distribution with mean μ_k and covariance matrix Σ_k.

To apply the GMM, the number of components K has to be determined. The Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) are commonly used for determining the K by assessing model complexity and goodness of fit [10].

The BIC is a criterion for model selection among a finite set of models. It balances the goodness of fit of the model with the complexity of the model. It penalizes models with more parameters, favouring simpler models that can explain the data. For GMMs, the BIC is typically calculated as:

$$\text{BIC} = -2\log(L) + k\log(n). \tag{4.3}$$

Similar to the BIC, the AIC is also a criterion for model selection based on the trade-off between goodness of fit and model complexity. The AIC penalizes complex models but with a less severe penalty compared to the BIC. It is calculated as:

$$\text{AIC} = -2\log(L) + 2k. \tag{4.4}$$

where the terms have the same meanings as in the BIC formula.

1. Determining the number of components: AIC and BIC scores

To determine the number of components using BIC or AIC scores, the number of components k is varied from 1 to k_{max} and fitted to the data. Therefore, we start by fitting different numbers of components with diagonal covariance matrices to the 3 clusters and calculate the AIC and BIC scores (Figure 4.4). AIC and BIC are known to yield different results [10], hence we decide to observe both scores and pick the optimum number of components based on two scores. As we aim for smaller AIC or BIC values, we identify the elbow point for AIC and local minimum point for BIC: we select 5 components for Cluster 1, 4 components for Cluster 2, and 2 components for Cluster 3 (Figure 4.4).

Subsequently, we fit the specified number of GMMs to our data and label the data points according to the GMM components they belong to (Figure 4.5). Upon visual inspection, it is evident that, particularly for Cluster 1 and Cluster 2, the AIC and BIC scores do not provide an optimal number of components. While the elbow point could be determined differently, it would not resolve the issue since there is no clear boundary and the

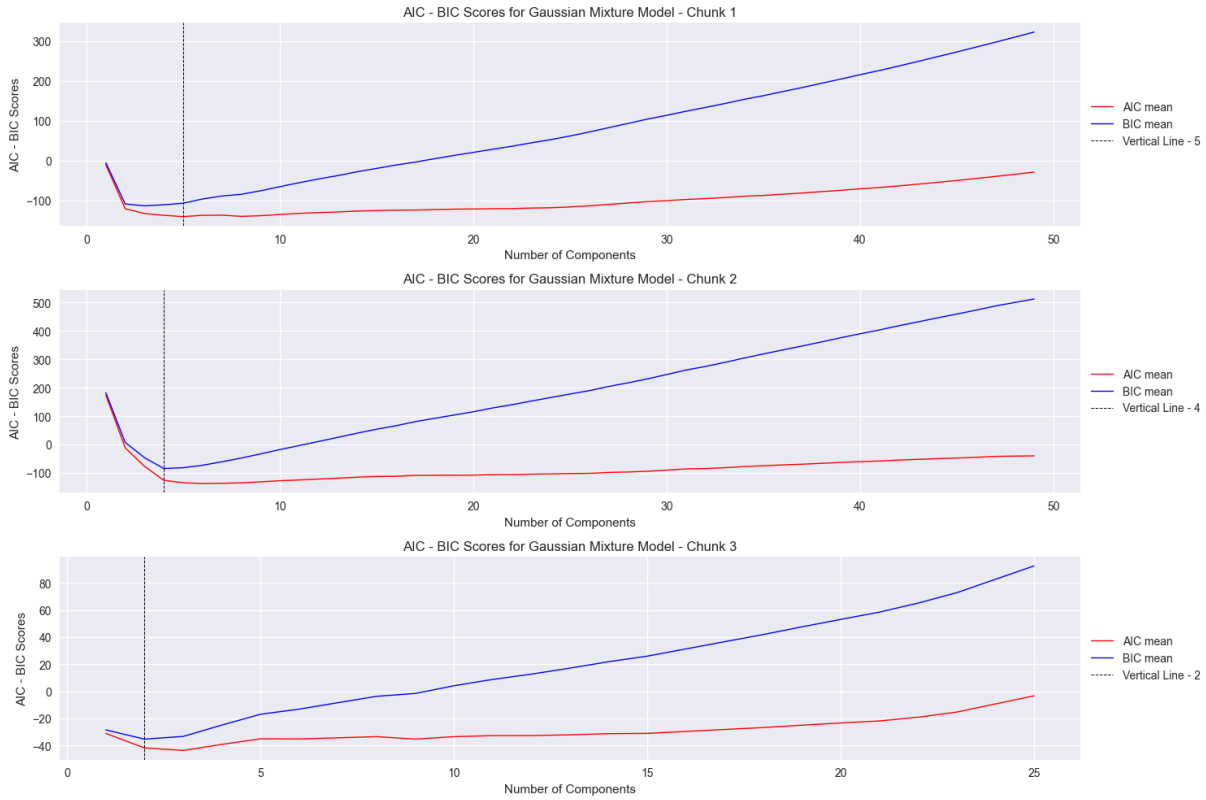


Figure 4.4: AIC - BIC Scores for Gaussian Mixture Model for all Clusters. The predicted number of the number of components (indicated by a vertical dashed line) for each cluster are 5, 4, and 2.

number of components that is observed by visual inspection do not correspond to the elbow points or local minimum of the AIC and BIC scores (Figure 4.4 and Figure 4.5). For example, visual inspection indicates that Cluster 1 might encompass 8 - 12 different components (Figure 4.5), yet AIC or BIC scores do not exhibit a clear elbow around those values. Consequently, we conclude that directly applying GMMs would not be sufficient to determine the number of components.

2. Determining the number of components: Peak Detection

Given the Gaussian-like distribution of compounds in Mass Spectrometry (MS) [57], insulin-related peaks are expected to exhibit similar characteristics. Thus, we anticipate insulin-related peaks to manifest as peaks representing true m/z ratios. Hence, analyzing the peaks of the data may provide insight into this distribution. Consequently, we employ a peak detection algorithm to identify these peaks.

Before applying peak detection, we want to smooth the data and reduce the risk of overshooting of the number of peaks. Therefore, we fit a cubic function to our data and then apply the Savitsky Golay filter which is shown to improve the quality of the peak detection algorithm by reducing the noise and preserving the important features [14, 38]. Subsequently, the peak detection algorithm is applied to identify peaks.

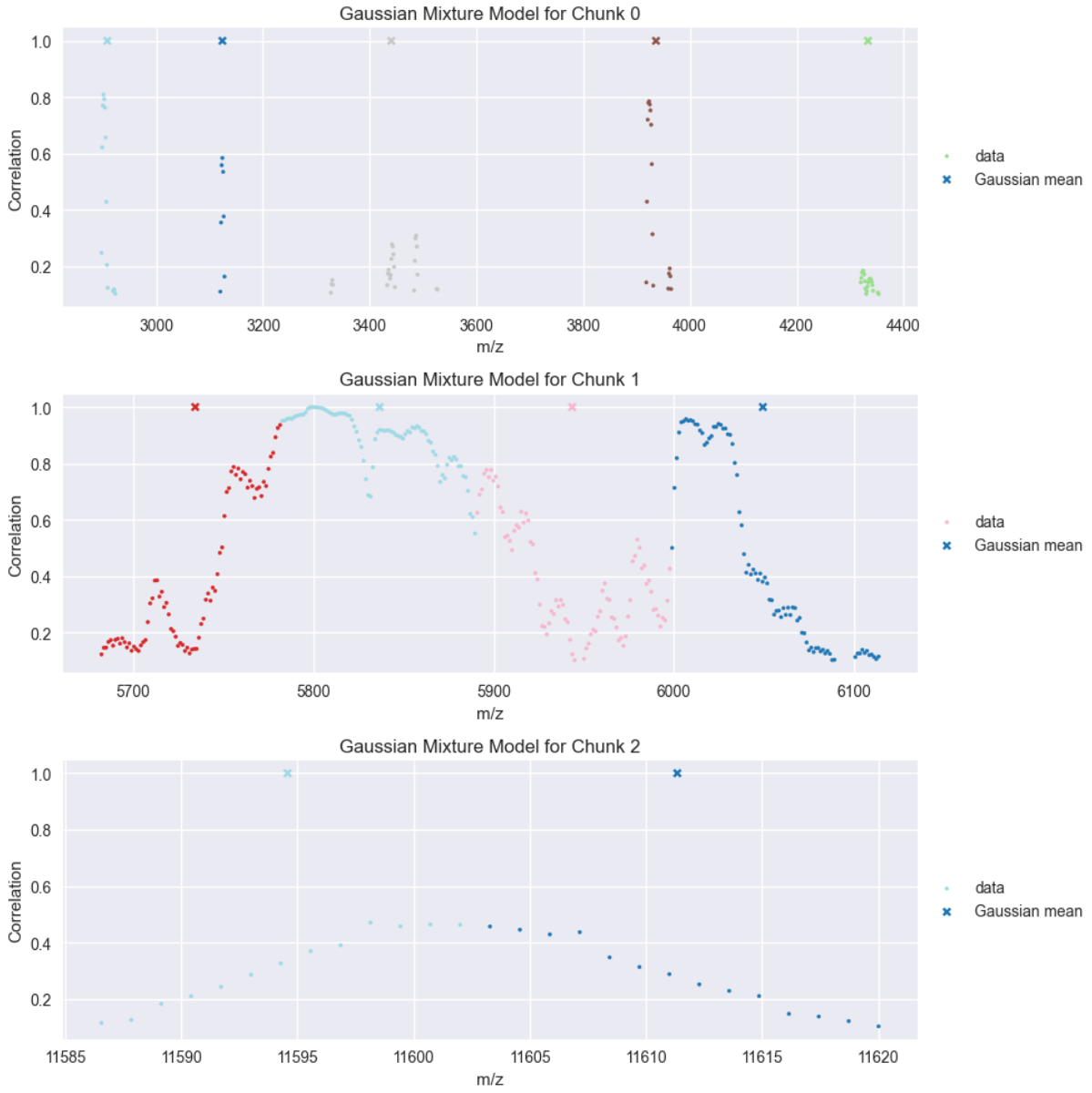


Figure 4.5: GMMs Gitted into the Clusters with the Given Number of Components. The means of the components are labelled with an 'x' and the colour represents the component.

A data point is designated as a peak if it surpasses its neighbouring data points by 10^{-5}. The rationale behind this threshold selection is that when the threshold is set to 10^{-4}, the algorithm fails to detect any peaks. Therefore, we opt for the smallest power of 10 capable of peak detection. This analysis identifies 13 peaks in Cluster 1, 37 peaks in Cluster 2, and 2 peaks in Cluster 3 (Figure 4.6).

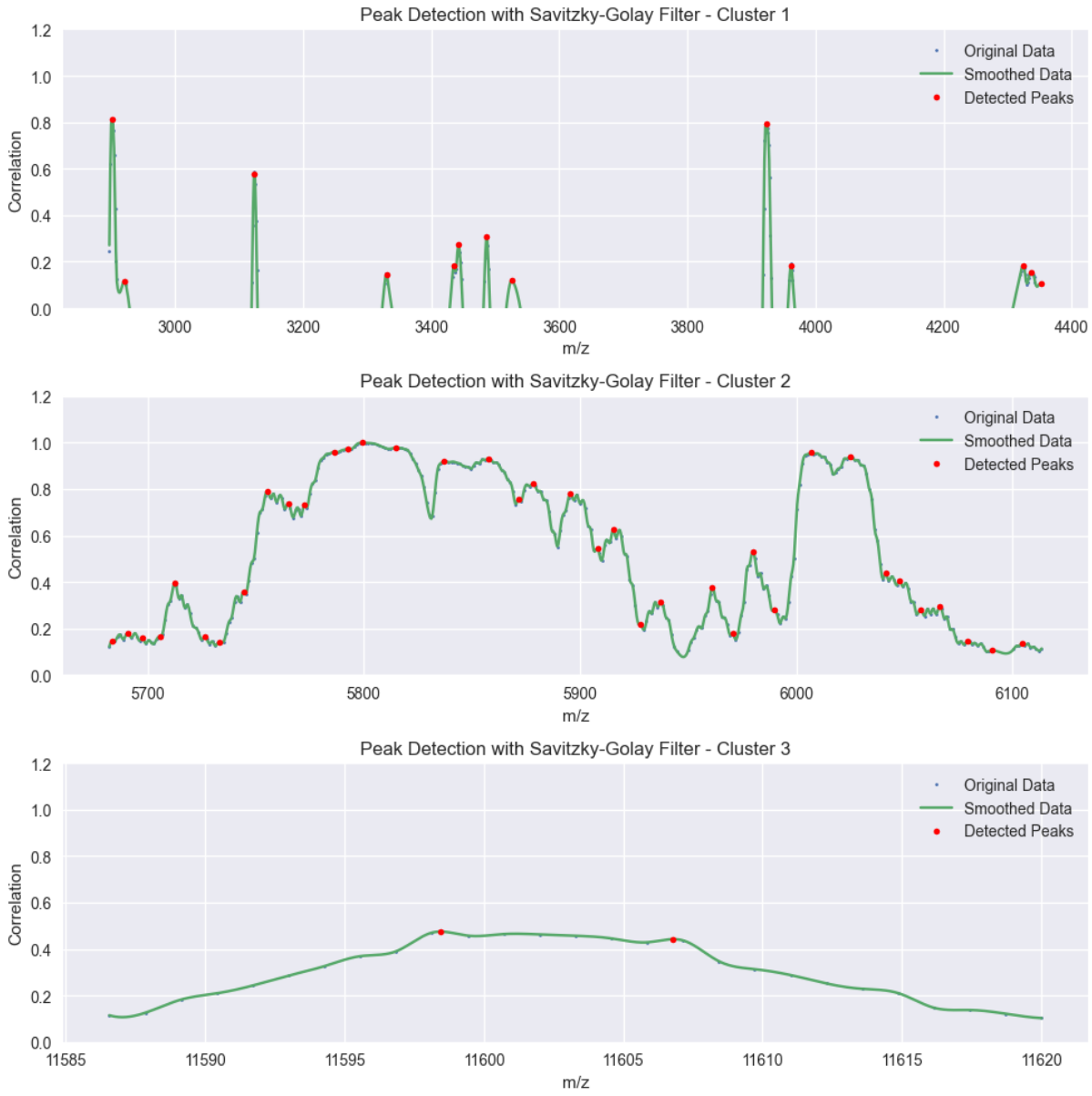


Figure 4.6: Peak Detection Using Savistky Golay filter. Cubic extrapolation is applied and peak threshold is 10^{-5}.

4.2.2 Fitting the GMM: Conclusion

After identifying candidate components (number of Gaussian distributions) within each cluster in subsection 4.2.1, this section aims to determine the corresponding Gaussian distributions and their means.

The detected peaks can serve directly as the means of the components, and such Gaussian Mixture Models (GMMs) can be fitted into the data. Alternatively, only the number of detected peaks can be utilized, allowing the GMM algorithm to determine the means itself. Given the uncertainty regarding which method might be superior, both approaches are employed, and their outcomes are compared.

1. GMM with Fixed Number of Components

By applying the peak detection algorithm, we determined the number of peaks in Cluster 1, Cluster 2, and Cluster 3 to be 13, 37, and 2, respectively. We select these numbers as the components, fit GMM models with a diagonal covariance structure for each cluster, and visualise the results (Figure 4.7). Upon visual inspection, it is evident that this method outperforms the approach of minimizing AIC/BIC scores. However, it is also notable that some of the allocated means are situated at local minima instead of peaks.

2. GMM with Fixed Number of Components and Means

The peak detection algorithm not only provided the number of peaks but also their locations. Thus, we fit GMM models with the given number of components - 13, 37, and 2 - employing a diagonal covariance structure and utilising the identified means of the Gaussian distributions (Figure 4.8). Upon visual inspection, similar results are observed for Cluster 1 and Cluster 3, whereas slight differences are noticeable in Cluster 2. To further illustrate this, we plot the means of the components for both methods (Figure 4.9). It is also apparent here that, particularly for Cluster 2, the two methods yield different outcomes.

Since the outcome is slightly different, we list the candidate means for both methods (Table 4.1) and we validate both outcomes in the following chapters.

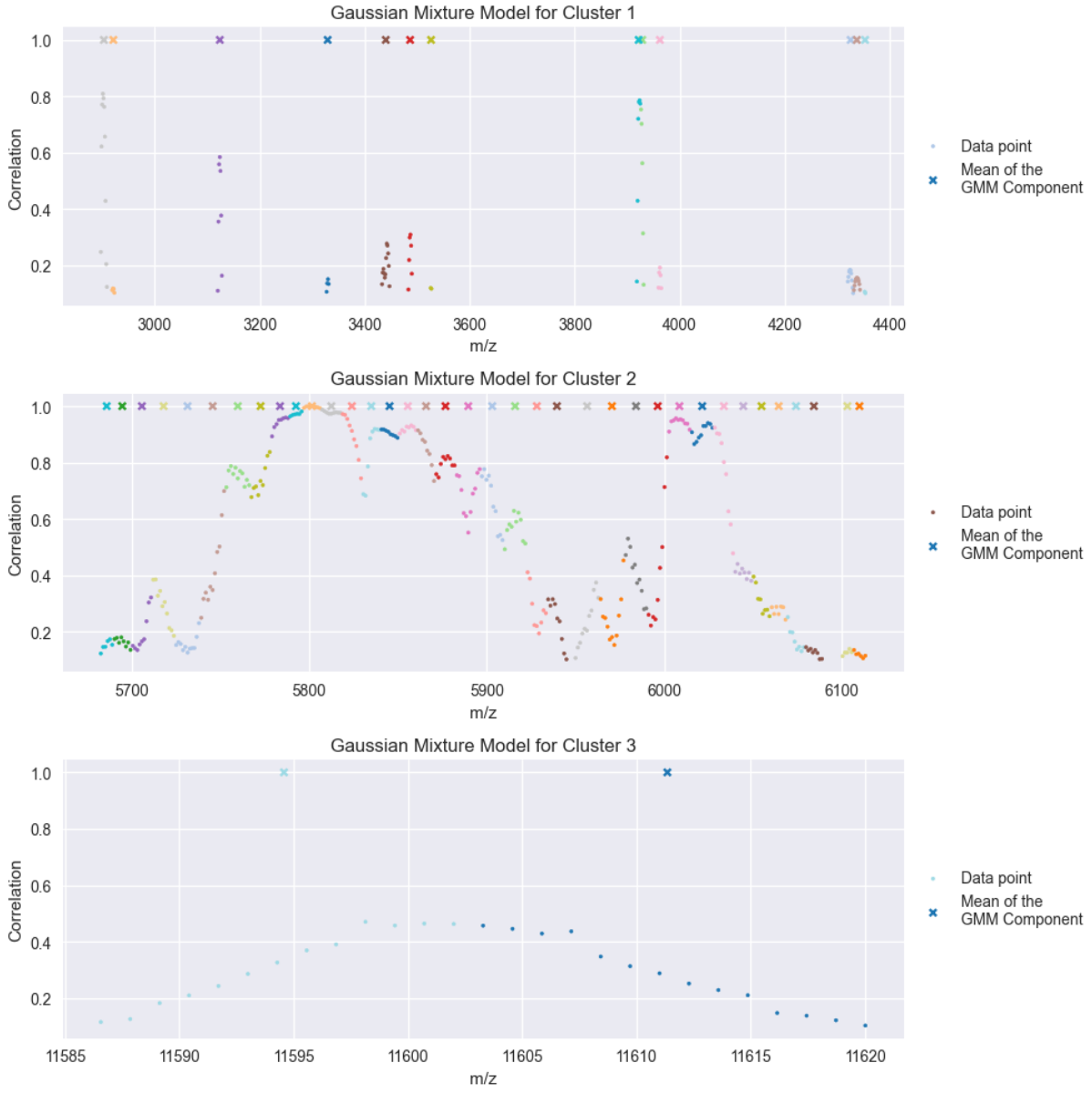


Figure 4.7: GMM Model with the Given Number of Components for each Cluster. Colour represents the component which the data is drawn from.

Table 4.1: The Candidate Outputs of GMM Methods. Method 1 is the GMM with the means determined by peak detection and Method 2 is the GMM with given number of components.

Method 1	Method 2	Method 1 (cont.)	Method 2 (cont.)
2903.38	2903.38	5842.60	5801.42
2921.39	2921.39	5854.19	5815.81
3124.08	3124.08	5867.70	5837.01
3329.33	3329.33	5880.55	5857.41
3436.28	3435.25	5890.84	5869.88
3443.73	3442.61	5900.53	5881.05
3486.33	3486.33	5912.09	5894.68
3526.22	3526.22	5925.67	5906.41
3924.51	3924.51	5939.01	5916.88
3961.83	3961.83	5954.71	5927.31
4324.94	4323.68	5964.15	5938.87
4336.74	4335.37	5973.86	5958.09
4352.40	4352.40	5983.50	5970.50
5685.69	5684.98	5991.87	5979.96
5694.48	5691.51	6000.26	5992.30
5705.54	5698.77	6010.52	6007.79
5718.41	5705.08	6021.46	6024.51
5733.21	5713.55	6032.42	6038.62
5748.66	5724.87	6044.64	6048.40
5763.45	5734.24	6058.16	6057.57
5776.97	5744.14	6071.79	6067.68
5789.85	5755.31	6083.78	6078.66
5802.07	5765.09	6103.32	6086.82
5812.34	5774.69	6110.13	6107.05
5821.39	5784.48	11594.58	11594.46
5831.64	5792.71	11611.35	11611.15

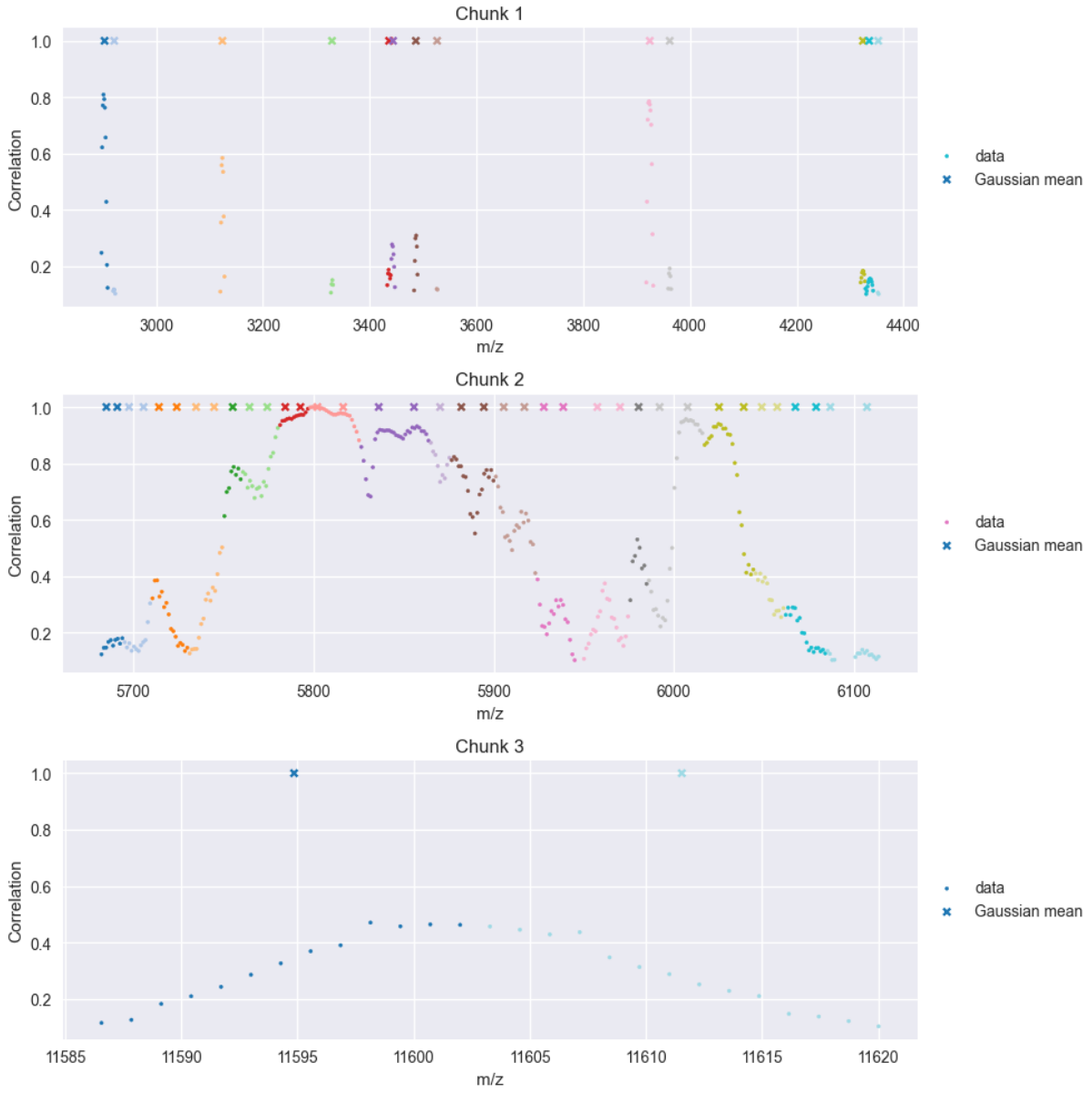


Figure 4.8: GMM Model with the Given Number of Components and Given Means for each Cluster. Colour represents the component which the data is drawn from.

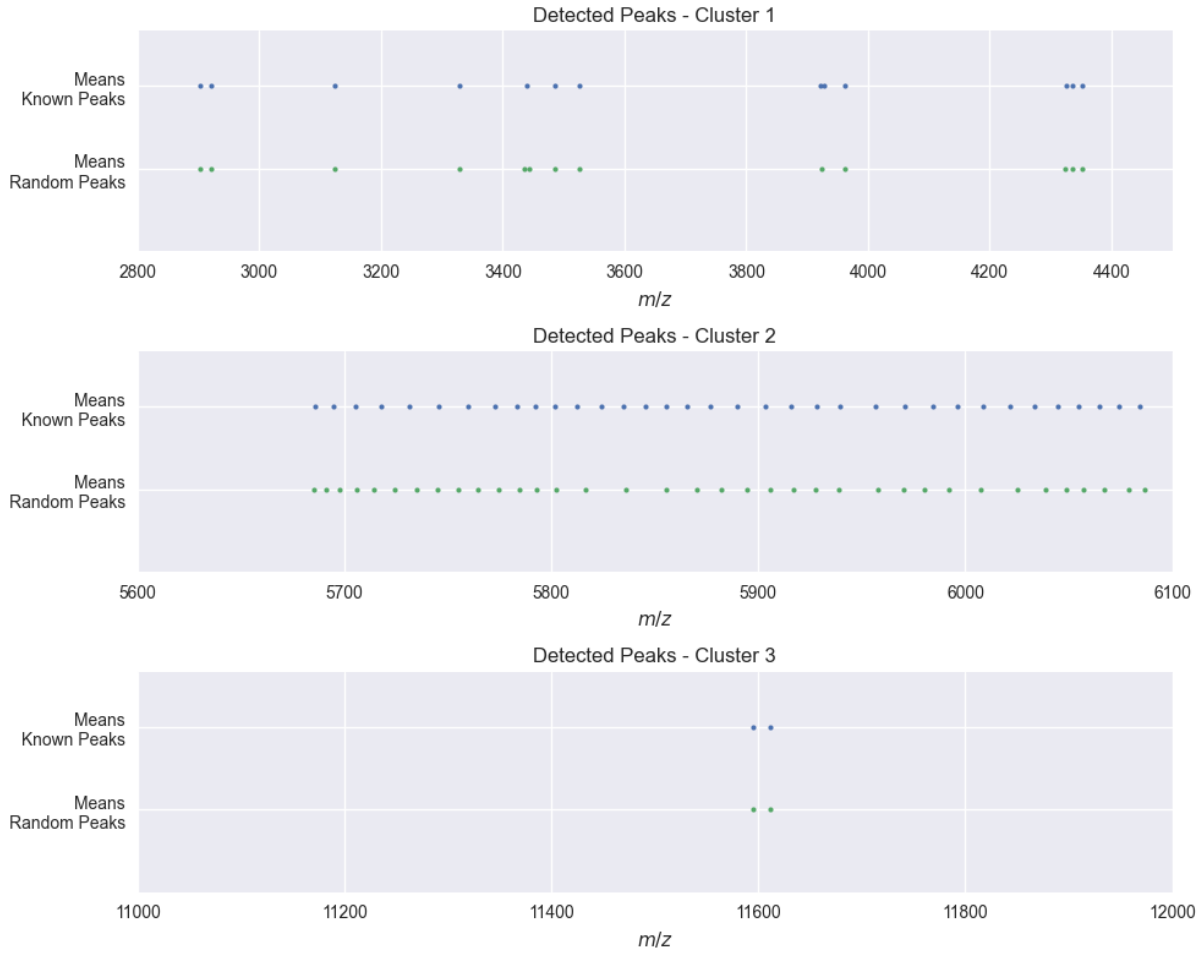


Figure 4.9: The Difference between GMM Methods. The means of the GMM model with pre-determined means (orange) and calculated means (blue).

Chapter 5

Identifying the Insulin-related Peaks: Sliding Window Approach

This chapter is written by Chris Butcher.

Previously, in Chapter 4, we used Pearson correlation to determine the mass-to-charge (m/z) values that are correlated with insulin. Subsequently, Gaussian Mixture Models were applied to extract the candidate insulin-related peaks. In this chapter, we start by utilising a sliding window approach to group m/z values based on their spatial distributions. In order to do this, we make use of the Frobenius norm. Following this, we find the groups of m/z values that are correlated with the group containing $Ins2$. As a final step, we obtain the candidate insulin-related peaks, again through use of the Frobenius norm.

5.1 Grouping m/z Values

As previously mentioned, the MSI process can result in multiple peaks that correspond to the same compound. On top of this, isotopes of the same compounds lead to Gaussian distributions around the true m/z value of a molecule [57]. Since artefact-like peaks are the result of the data acquisition process, we expect that the insulin-related peaks will have the same spatial distribution as the true insulin molecule. Hence, we aim to utilise the spatial distributions of the m/z values to identify those that correspond to insulin.

To achieve this, we first group the m/z values based on the similarity of their spatial distributions. Because we expect a Gaussian distribution around the true m/z value of a molecule [57], we group the m/z values sequentially, in order to create segments of m/z values that correspond to the same peaks in the spectra. To determine whether m/z values belong to the same group or not, we go through the ordered set of 14,000 m/z values and compare the distribution of each one with those of the 5 previous m/z values. For each m/z value, we have a matrix of dimension 165 * 228 containing its intensity in each pixel of the sample. We can refer to this matrix as the ion matrix for a particular m/z value. To compare the distributions of two given m/z values, we take the difference in their ion matrices, and then compute the Frobenius norm of the resultant matrix. The Frobenius norm of a matrix, A, can be defined as:

$$\|A\| = [\sum_{i,j} abs(a_{i,j})^2]^{1/2}$$

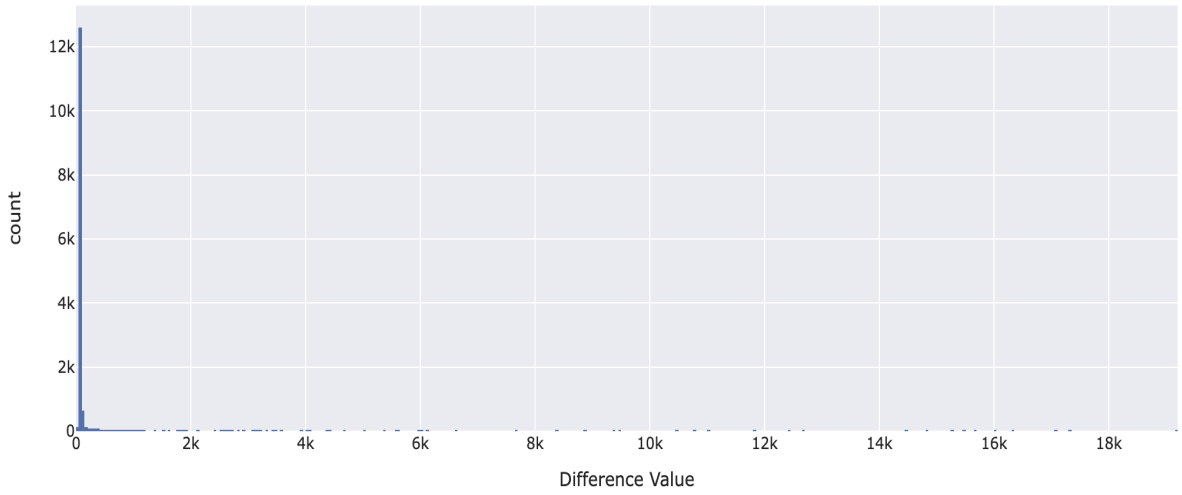
We calculate this difference measure between the current m/z value and the 5 previous values, before calculating the mean of these 5 values. This is done to ensure that the current m/z value is sufficiently different to the preceding values to justify creating a new group. We can visualise the distribution of all the difference measures that we calculate (Figure 5.1). The maximum difference measure observed is 19183.94, which results in a sparse, stretched histogram. However, from closer inspection we can see that the majority of the difference measures are smaller than 150 in size. Indeed, we find that roughly 95% of the difference measures are less than 150. Hence, in order to determine whether we should consider an m/z value as a so called 'change value', we take 150 as a cutoff value. So, if the average difference measure of an m/z value with the 5 values before it is larger than 150, we consider it as belonging to a new grouping of m/z values, and save it as a 'change value'. We can then use these saved values to determine the groups of m/z values that we expect to have similar spatial distributions, and that correspond to the same molecule or molecule artefact-like peak. Applying this approach results in the range of m/z values being sorted into 137 separate groups of m/z values. This grouping of the m/z range is visualised in Figure 5.2.

5.2 Correlating m/z Groups

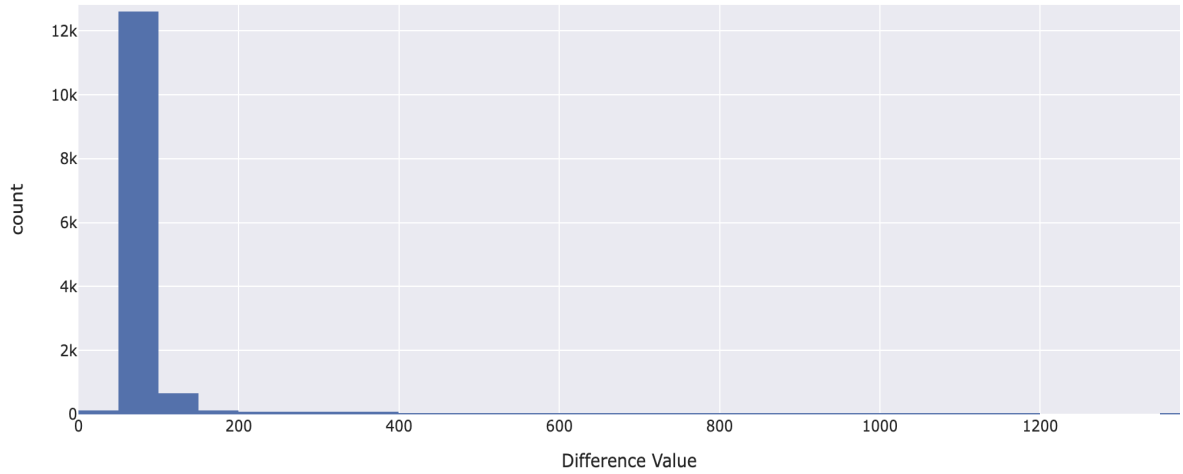
In the previous section 5.1, our set of 14000 m/z values was separated into 137 separate groups. Each of these groups contains a range of m/z values that have a similar spatial distribution. In this section, we compute the correlation between these groups. However, since our aim is to identify which of these groups are correlated with the group containing $Ins2$, we only select the correlations that are relevant.

Through manual inspection, we find that $Ins2$ is contained in group 31, which is the m/z range [5795.63, 5800.77]. Therefore, to determine which ranges of m/z values are correlated to insulin, we calculate the correlation between all groups and group 31.

To apply any form of correlation between the groups, we need to obtain a matrix that represents each group. In order to do this, we calculate the element-wise mean of the ion matrices in each group. Per group, this gives us the average intensity of the m/z values in each of the pixels of our sample. The ion images for each of the first two groups, along with the range of m/z values they include, are visualised (Figure 5.3).



(a) Full Histogram



(b) Zoomed Histogram

Figure 5.1: Histogram of Difference Values

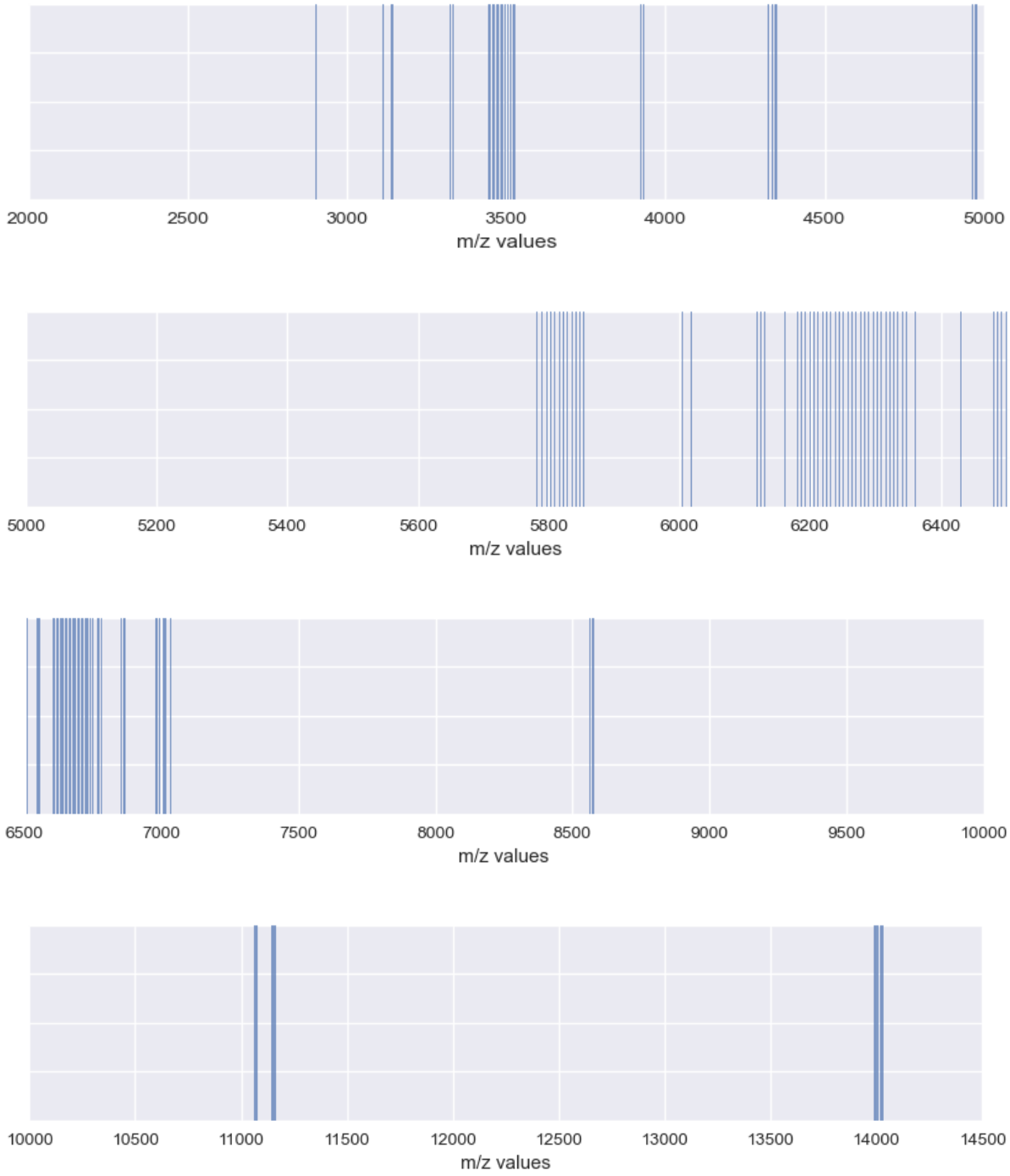
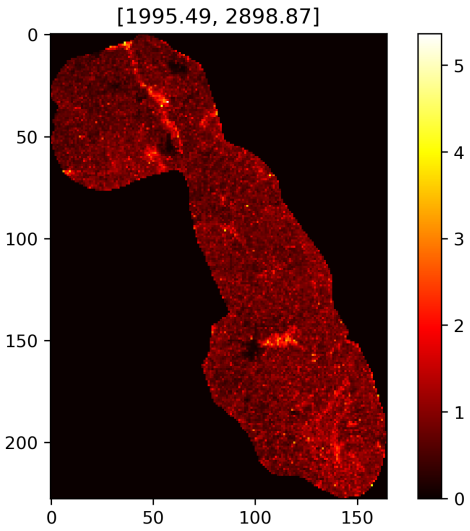
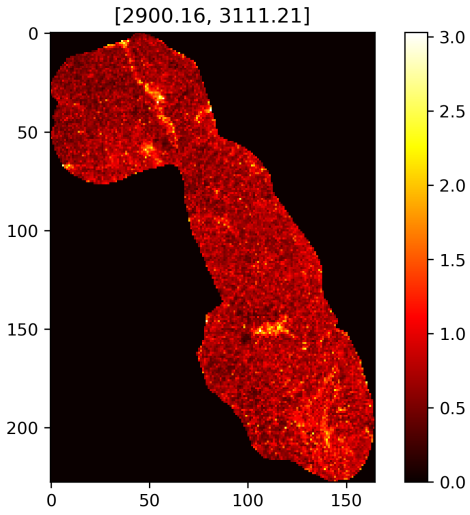


Figure 5.2: Grouping of m/z Range



(a) Group 0 Ion Image and m/z Range



(b) Group 1 Ion Image and m/z Range

Figure 5.3: Ion Images of the First Two Groups. (a) The ion image of Group 1, (b) The ion image of group 2.

5.2.1 Pearson Correlation

We initially use Pearson correlation to compare the relationship between the groups. The formula for Pearson correlation was previously defined in Section 4.1. For each of the 137 groups, we have an ion matrix of size [165, 228]. In a similar way to our approach in Section 4.1, we transform our data by treating each group as a sample, and each pixel as a variable. So, when we apply Pearson correlation, a matrix r of size [136, 136] is obtained. We are only interested in the correlation of groups with group 31. We select the relevant correlations, leaving us with the vector r_{corr} of size [1, 136].

We can visualise the obtained Pearson coefficients in a histogram (Figure 5.4). The Pearson coefficient measures the strength of the linear relationship between components [41], and so we can see that the majority of groups do not have a strong correlation with $Ins2$. There is a clear separation in the distribution of the Pearson coefficients. Hence, use a cut-off value of 0.5 to obtain the groups that have a notable correlation with group 31. This results in 15 groups of m/z values that are related to $Ins2$. In Section 5.2.2 we will select the candidate insulin-related peak values from these groups.

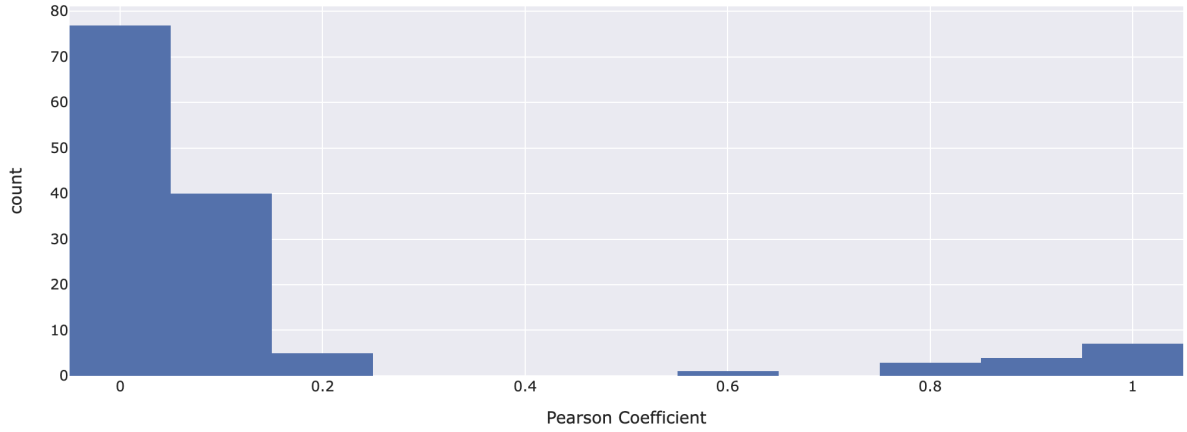


Figure 5.4: Histogram of the Pearson Coefficients.

5.2.2 Spearman Correlation

We decided to also explore a non-parametric measure of correlation between the groups. In order to do this, we selected the Spearman rank correlation coefficient to use. Spearman correlation measures the strength and direction of monotonic association between two components. The Spearman correlation coefficient is calculated using the following formula:

$$r_s = \frac{cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}} \tag{5.1}$$

where:

- raw data X_i and Y_i are converted into rank variables R(X_i) and R(Y_i),
- Cov(R(X), R(Y)) is the covariance of the ranks variables,
- $\sigma_{R(X)}$ and $\sigma_{R(Y)}$ are standard deviations of the rank variables.

As in the previous section 5.2.1, we transform the ion matrix for each group in order to apply Spearman correlation. This results in a matrix of Spearman correlation coefficients of size [136, 136]. Again, as before we are only interested in the correlation between group 31 and all other groups, so we select the relevant correlations, leaving us a vector of size [1, 136].

The obtained Spearman correlation coefficients are visualised in figure 5.5. There is not as clear of a separation in these correlation coefficients. However, there appears to be a drop off in frequency after the Spearman coefficient value 0.932, so we will take this as the cut-off point. We retain the groups with a Spearman correlation coefficient that is greater than 0.932 with group 31. This results in 19 candidate groups. The candidate insulin-related peaks will be obtained in the next section.

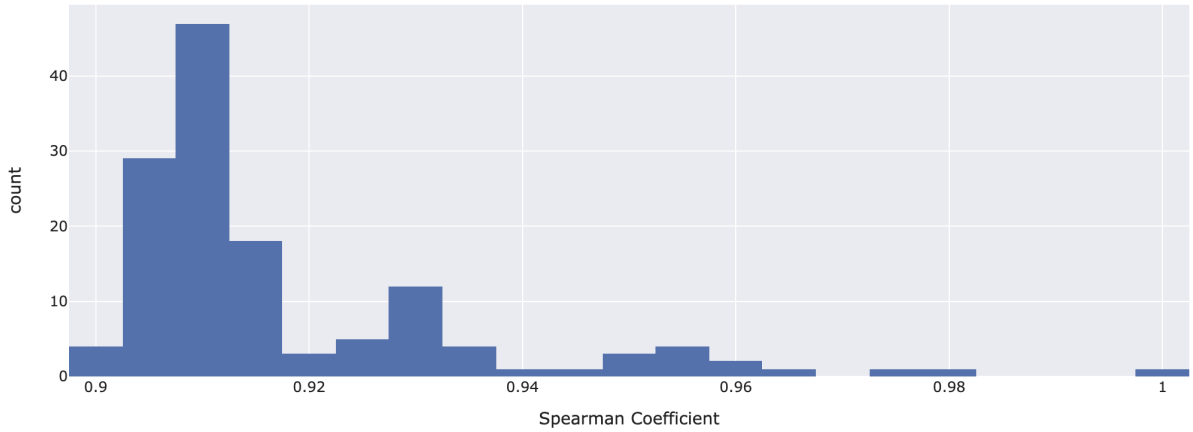


Figure 5.5: Histogram of the Spearman Coefficients.

5.3 Selecting Candidate Values

After applying Pearson and Spearman correlation, we are left with 15 and 19 groups that are correlated with group 31, respectively. Due to the presence of isotopes, molecules exhibit a Gaussian distribution around their true m/z value [57]. Therefore, the true m/z ratios of the insulin-related peaks will correspond to the peak value.

To obtain the candidate m/z values from each of the groups that are correlated with $Ins2$, we want to find the m/z value that has maximum intensity. To achieve this, we again make use of the Frobenius norm, which was defined in Section 5.1. In each group, we calculate the Frobenius norm of every m/z value. The m/z value with the maximum norm is then selected as the candidate insulin-related peaks. The obtained candidate m/z values resulting from both correlation approaches are illustrated in figures 5.6 and 5.7. Initially, both approaches appear to provide similar results for potential insulin-related peak values. Indeed, upon inspection we find that the values returned from using Pearson correlations are a subset of those obtained from using Spearman correlation. Along with the 15 candidate values from the Pearson correlation approach, the Spearman approach detects four extra possible insulin-related peak.

The results from both approaches will be validated, along with those from Chapter 4, in the following Chapter 6. The exact values obtained from each approach can be seen in Table 5.1.

Table 5.1: Candidate m/z values obtained from the Pearson and Spearman approaches.

Pearson	Spearman
	2901.45
	3113.78
3923.23	3923.23
	4336.31
	5778.90
5787.91	5787.91
5794.34	5794.34
5800.77	5800.77
5802.06	5802.62
5808.50	5808.50
5814.93	5814.93
5821.36	5821.36
5827.80	5827.80
5839.38	5839.38
5840.67	5840.67
5847.10	5847.10
5859.97	5859.97
6006.67	6006.67
6024.69	6024.69

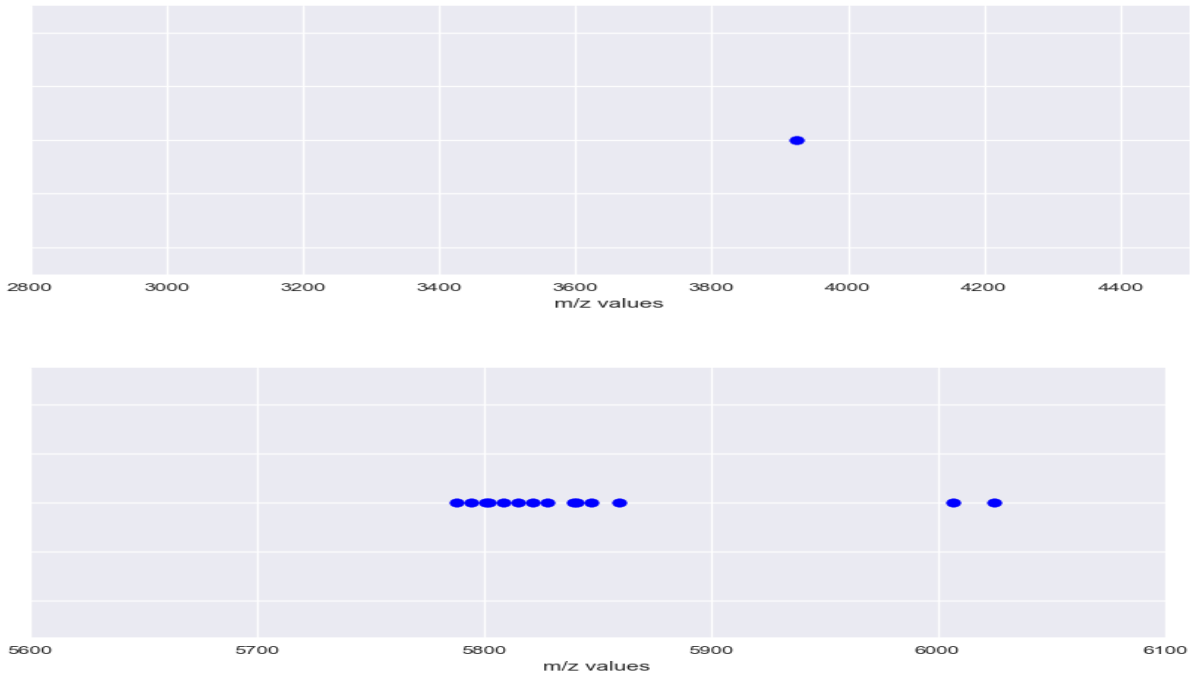


Figure 5.6: Candidate Insulin-Related Peak m/z values for Pearson Correlation Approach.

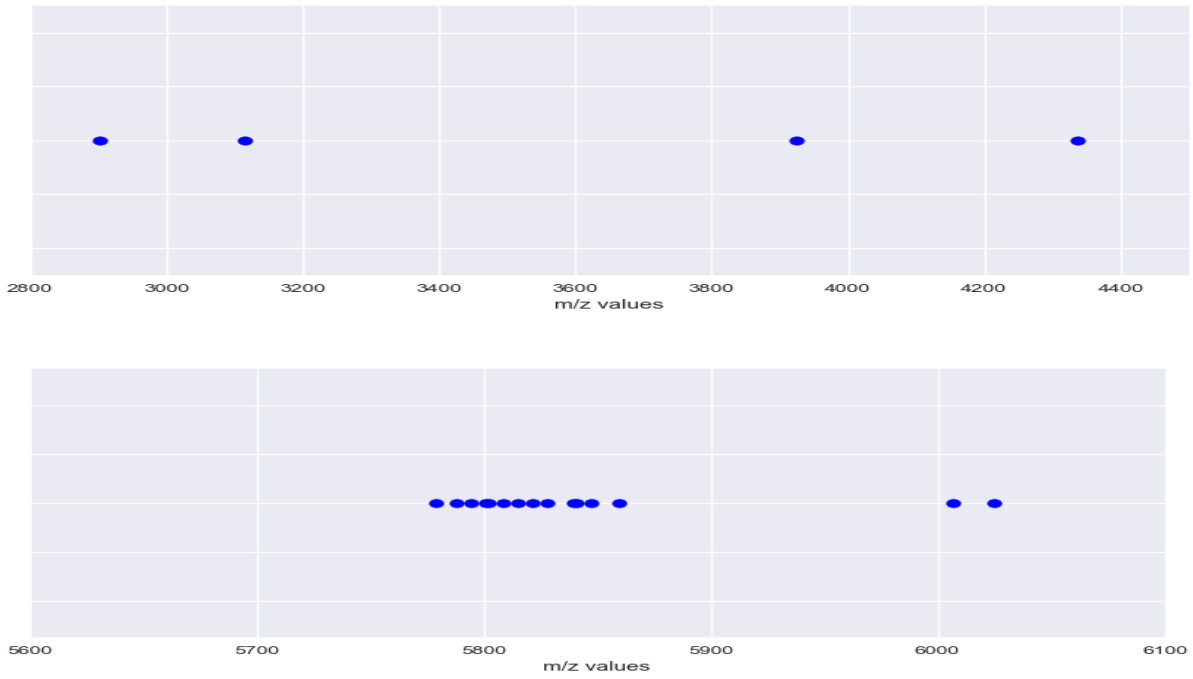


Figure 5.7: Candidate Insulin-related Peak m/z values for Spearman Correlation Approach.

Chapter 6

Validation and Evaluation of the Candidate Peaks

Section 6.1 and section 6.2 are written by Chris Butcher. Other sections are written by both authors.

In Chapter 3 we constructed the map of the islets of Langerhans. In parallel, we applied 2 different approaches (see Chapter 4 and Chapter 5) to identify insulin-related peaks. Now we have the insulin-related peak candidate values but validation and evaluation are necessary. Since insulin-related peaks are co-located with insulin, and insulin is synthesised and stored in the islets of Langerhans [57], we expect insulin-related peaks to be found within the islets. We first test the distribution of the candidate peaks and check whether they exist within the islets significantly more than the non-islets. Then we compare the significant candidates of two different models. Finally, we utilise the map and the insulin-related peaks to assess whether we gain more insights.

6.1 Spatial Distribution using the Map

This section is written by Chris Butcher.

The insulin-related peaks must be co-located with insulin, hence they must be found in the islets of Langerhans. However, an insulin-related peak and a separate compound may have similar m/z ratios. Therefore, we cannot expect that m/z ratios corresponding to insulin-related peaks are not found in other parts of the tissue. However, we can assume that the insulin-related peaks are found significantly more in islets than in the rest of the tissue.

We construct our null hypothesis and alternative hypothesis as following:

H_0: There is no difference in the intensity of the compound at the specified m/z ratio between the islets of Langerhans and the rest of the pancreatic tissue.
H_1: The intensity of the compound at the specified m/z ratio is larger in the islets of Langerhans compared to the rest of the pancreatic tissue

In chapter 3, we have defined the map of the islets for the given pancreatic tissue (Figure 3.9a). Since we are only interested in the Islets of Langerhans, we merged Cluster 1 into Cluster 0; and Cluster 3 into Cluster 2, labelling it as Cluster 1 (Figure 6.1).

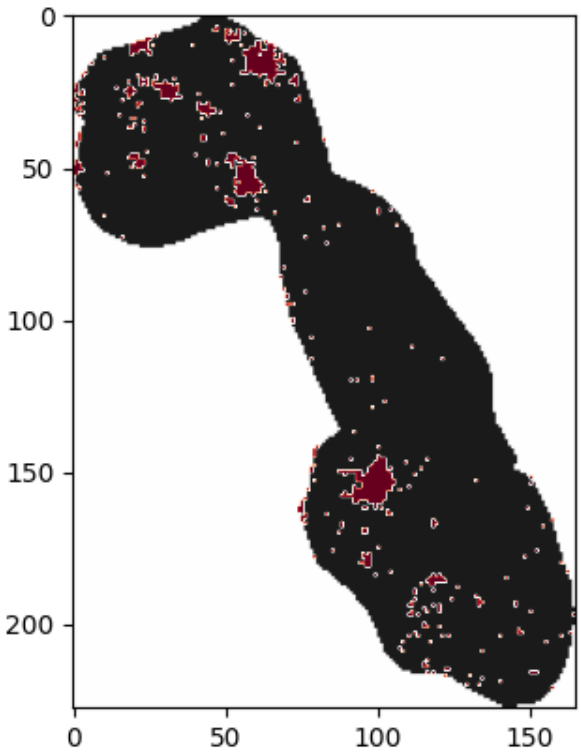


Figure 6.1: The Map of the Islets of Langerhans. Red areas indicate the islets.

To test our hypothesis, we applied the parametric one-tailed t-test and non-parametric one-tailed Mann-Whitney U test with a significance level of 0.05 and 0.01. We have listed the rejected m/z ratios for each test and the significance level for the three approaches GMM fitted model with means determined by peak detection, GMM fitted model with the number of components determined by peak detection, and sliding window with Spearman Correlation, in Table 6.1, Table 6.2, and Table 6.3, respectively. The example spatial distribution and the intensity distribution of rejected and accepted m/z ratios are shown in Figure 6.2a. Significant values for each method are plotted in Figure 6.3.

Table 6.1: Rejected m/z Ratios of GMM fitted Model with Given Means. Parametric and non-parametric tests with different confidence levels are done.

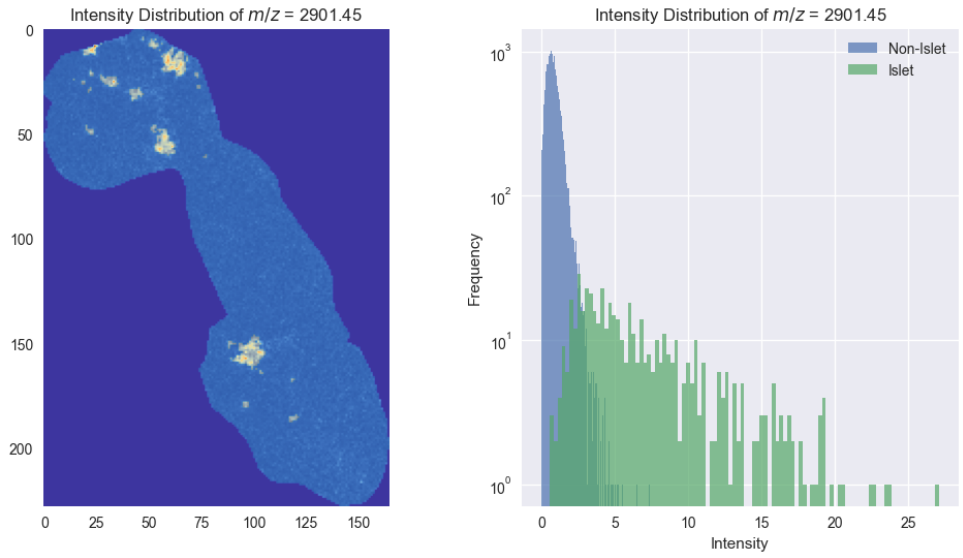
$\alpha = 0.05$		$\alpha = 0.01$	
T-test	U-test	T-test	U-test
2921.39		2921.39	2921.39
3329.33	3329.33	3329.33	3329.33
3526.22	3526.22	3526.22	3526.22
	4326.89		4326.89
	4338.25		4338.25
4348.51	4348.51	4348.51	4348.51
5698.01	5698.01	5698.01	5698.01
	5705.62	5705.62	5705.62
5734.18	5734.18	5734.18	5734.18
6079.00	6079.00	6079.00	6079.00
6086.83	6086.83	6086.83	6086.83
6107.05	6107.05	6107.05	6107.05

Table 6.2: Rejected m/z Ratios of GMM fitted Model with Given Number of Components. Parametric and non-parametric tests with different confidence levels are done.

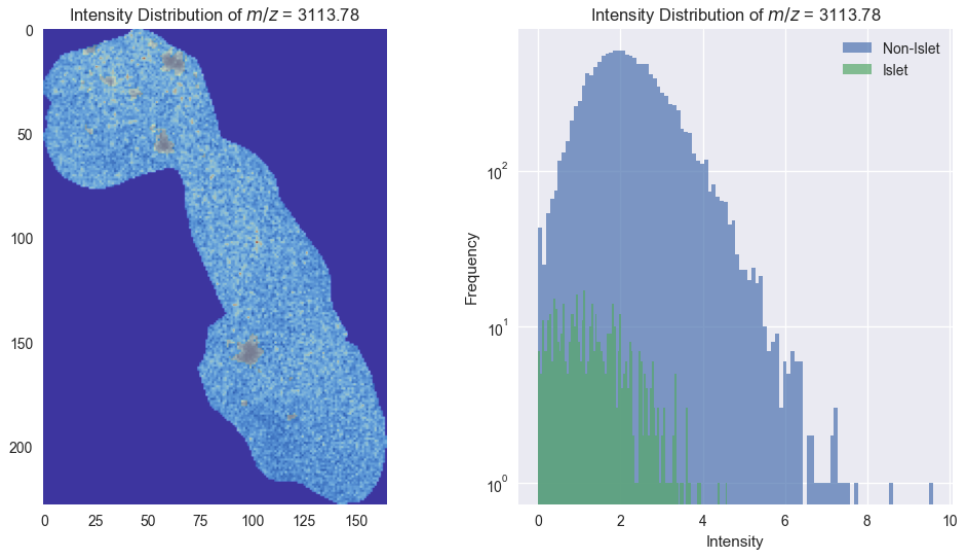
$\alpha = 0.05$		$\alpha = 0.01$	
T-test	U-test	T-test	U-test
2921.39		2921.39	2921.39
3329.33	3329.33	3329.33	3329.33
3526.22	3526.22	3526.22	3526.22
	4324.94		4324.94
	4336.74		4336.74
4352.40	4352.40	4352.40	4352.40
5697.71	5697.71	5697.71	5697.71
5733.21	5733.21	5733.21	5733.21
6075.01	6075.01	6075.01	6075.01
6085.07	6085.07	6085.07	6085.07
6103.98	6103.98	6103.98	6103.98
6110.78	6110.78	6110.78	6110.78

Table 6.3: Rejected m/z Ratios of Sliding Window Approach. Parametric and non-parametric tests with different confidence levels are done.

$\alpha = 0.05$		$\alpha = 0.01$	
T-test	U-test	T-test	U-test
3113.78	3113.78	3113.78	3113.78
	4336.31		4336.31



(a) Distribution of an Example Accepted Candidate m/z Ratio. (Left) Spatial distribution, (Right) Histogram of the $m/z = 2901.45$.



(b) Distribution of an Example Rejected Candidate m/z Ratio. (Left) Spatial distribution, (Right) Histogram of the $m/z = 3113.78$.

Figure 6.2: Example Results of the Hypothesis Test. (a) Accepted example, (b) Rejected example.

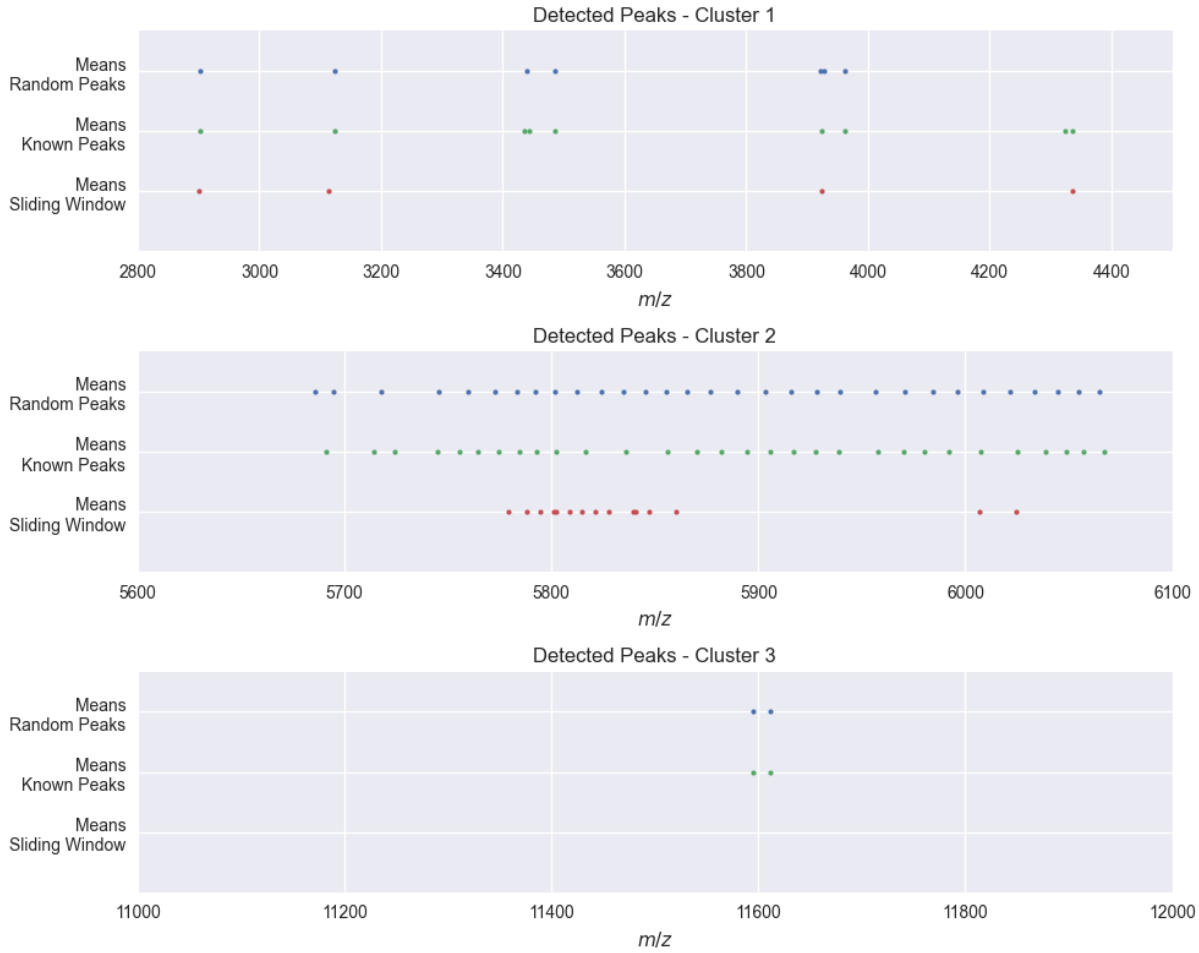


Figure 6.3: Significant Insulin-related Peaks Candidates. The candidates of different methods are plotted.

6.2 Comparing the Outputs of the Models

This section is written by Chris Butcher.

In Chapter 4, we started our analysis by applying Pearson Correlation. Then, we applied a peak detection algorithm to detect the number of components. Finally, we fitted GMM. We either used only the number of detected peaks as the number of components or we used the number of peaks as the number of components and the means of the peaks as the means of the components. We will call these two approaches Model 1 and Model 2 from now on. This resulted in 2 different candidate lists. In Chapter 5, the spatial distribution of the m/z values was used to group them. We then looked at the correlation between these groups to determine candidate insulin-related peak values. Two correlation methods were utilised, namely Pearson and Spearman. However, we use the Spearman correlation output since it also contains the values that were obtained from utilising the Pearson correlation. From now on we call them Model 3. We apply spatial distribution tests to find the m/z ratio candidates that are present within the islets significantly more than the non-islets. We use a strict significance level and do not include the values if

they are not significant at level 0.01 when either the parametric or non-parametric test is applied (not significant at one test is enough).

We have 3 different candidate lists, outputted by 3 different models, and we want to check whether these values correspond to each other (are they the same). We defined two m/z ratios to be the same if they differ by less than a given value, δ.

$$f(x_1, x_2) = \begin{cases} x_1 \sim x_2 & \text{if } |x_1 - x_2| < \delta \\ x_1 \not\sim x_2 & \text{if } |x_1 - x_2| \geq \delta \end{cases} \tag{6.1}$$

We pick $\delta = 10$ and first compared significant m/z ratios of Model 1 and Model 2. The output of Model 2 is found to be similar to the output of Model 1. From now on, we only use the outputs of Model 1. Then, we apply the same analysis to Model 1 and Model 3. The m/z ratios are either found in both models or only found in Model 1 (Figure 6.4 and Table 6.4).

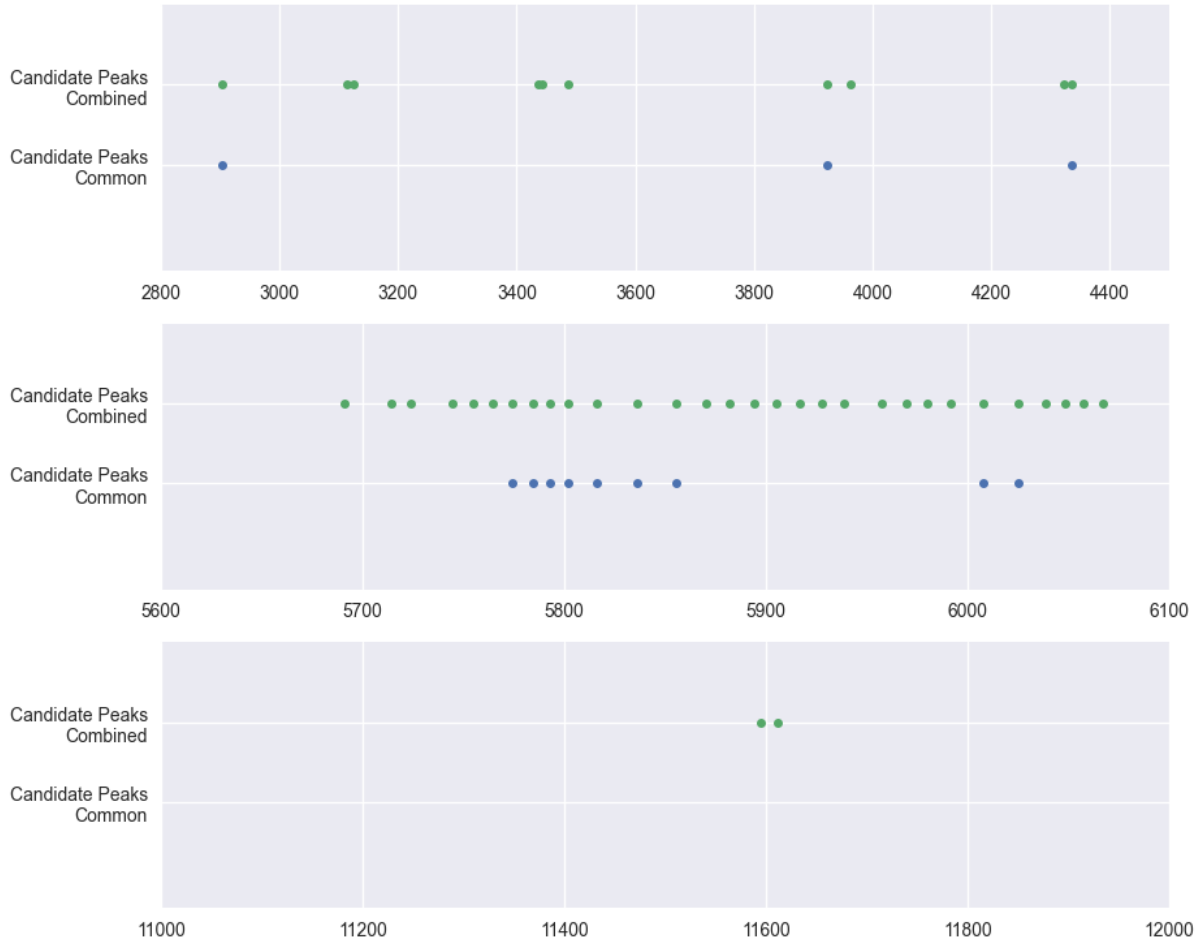


Figure 6.4: Common and Unique Candidates Outputted by Different Methods. Common candidates are outputted by all of the methods and candidates are considered as common if they differ less than 10 m/z.

Table 6.4: The Final Insulin-Related Peak Candidates. The candidates found by one approach are listed as 'Unique Candidates' and the candidates found by two approaches are listed as 'Common Candidates'.

Unique Candidates	Unique Candidates (cont.)	Common Candidates
2901.45	5859.97	2903.38
3124.08	5880.55	3924.51
3436.28	5890.84	5776.97
3443.73	5900.53	5789.85
3486.33	5912.09	5802.07
3923.23	5925.67	5812.34
3961.83	5939.01	5821.39
5685.69	5954.71	5831.64
5694.48	5964.15	5842.60
5718.41	5973.86	5854.19
5748.66	5983.50	5867.70
5763.45	5991.87	6000.26
5778.90	6006.67	6010.52
5787.91	6024.69	6021.46
5794.34	6044.64	6032.42
5808.50	6058.16	
5814.93	6071.79	
5827.80	11594.58	
5839.38	11611.35	
5847.10		

6.3 UMAP Analysis of Data

So far we have analysed the MSI data, identified the mapping of the islets of Langerhans, tried to identify insulin-related peaks, and generated 2 lists of insulin-related peak candidate values. We now wish to see whether we can gather more insights by using our findings. First, we will check whether our mapping can help us to identify different structures within the pancreatic tissue: an organisation within the islets and/or structures other than the islets of Langerhans. Then, we try to remove the insulin-related peaks from the islets and apply a similar analysis to gather more insights.

6.3.1 UMAP Analysis of All Data with Cluster Labelling

Uniform Manifold Approximation and Projection (UMAP) is a dimensionality reduction technique, which performs well at visualising high dimensional data by projecting it into a lower dimensional space [36]. UMAP captures essential patterns within the data by constructing a graph of the data's nearest neighbours, and then optimising a low dimensional representation such that local and global structures are preserved. The method allows for insightful visualisations, whilst retaining important relationships. This enables users to uncover trends and clusters in the data that were previously unclear.

We apply UMAP to the data and colour them with the clusters assigned to them in Chapter 3 (Figure 6.5). Blue is the outside of the tissue and is omitted, cyan is the islets of Langerhans, pink is the tissue surrounding the islets, and red is the rest of the tissue. It is seen that red and pink are co-allocated, which is expected since they do not particularly correspond to different structures; pink is just outside of the islets and red is the rest of the tissue. However, it is seen that some of the pink and red coloured data are co-located with the islets, highlighted by the blue rectangle, the area of interest (AoI) (Figure 6.5). This can be either a labelling mistake when the map is constructed, or those pixels might correspond to different structures.

To further investigate the AoI, and the underlying reason why those data points are co-located with the islets, we check the mean intensity of m/z ratios of the AoI, the non-islets, and the islets (Figure 6.6). It is observed that the AoI share similar mean intensity of m/z ratios, except around m/z ratio of 5800 and m/z ratio of 6000. $Ins2$ has the m/z ratio of 5800 (Figure 3.1). So this indicates these pixels might be another structure surrounding islets that do not produce or store insulin which can be observed when the spatial distribution of AoI is plotted (Figure 6.7). These results indicate that the pixels highlighted within the AoI might be corresponding to a different structure and further anatomical investigation is needed.

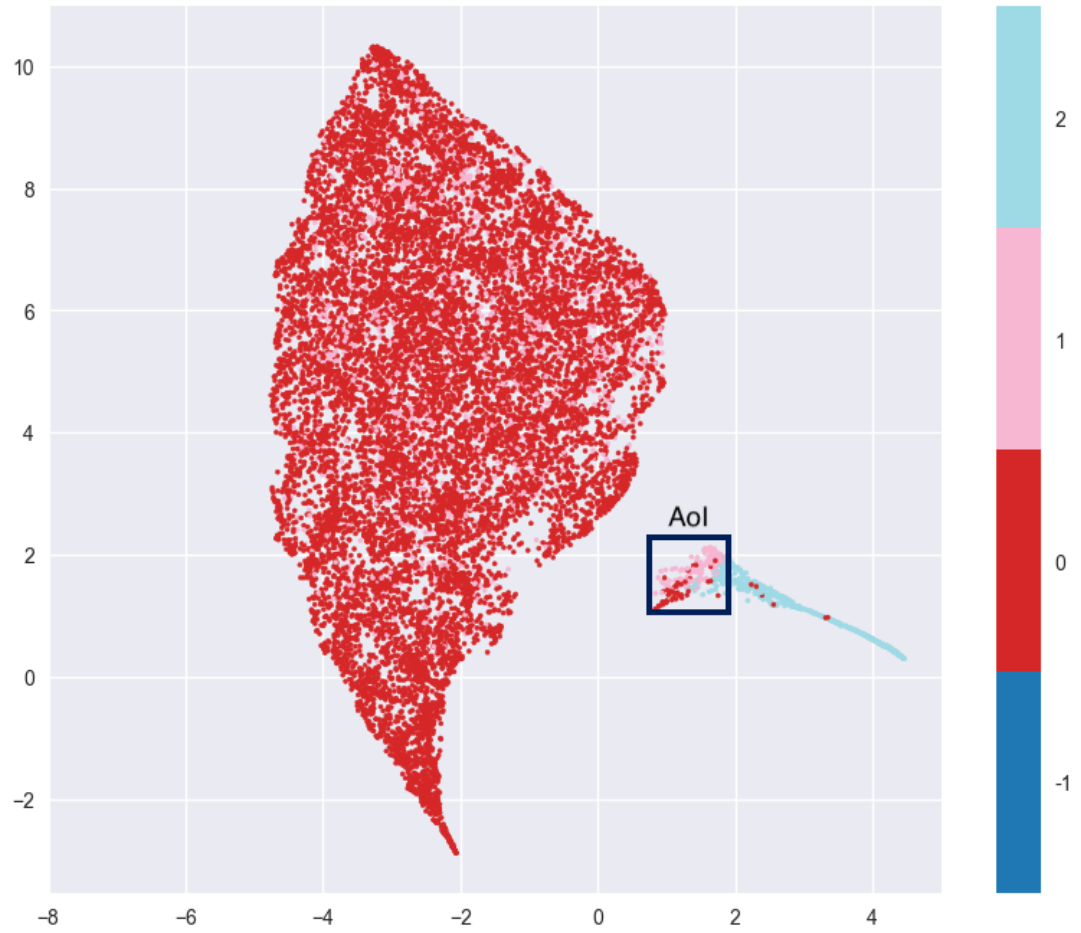


Figure 6.5: UMAP Analysis of Original Data. Each point is a pixel and colours represent the clusters assigned in Chapter 3: Cyan (2) represents the islets. The area of interest (AoI) is labelled different than the neighbouring pixels.

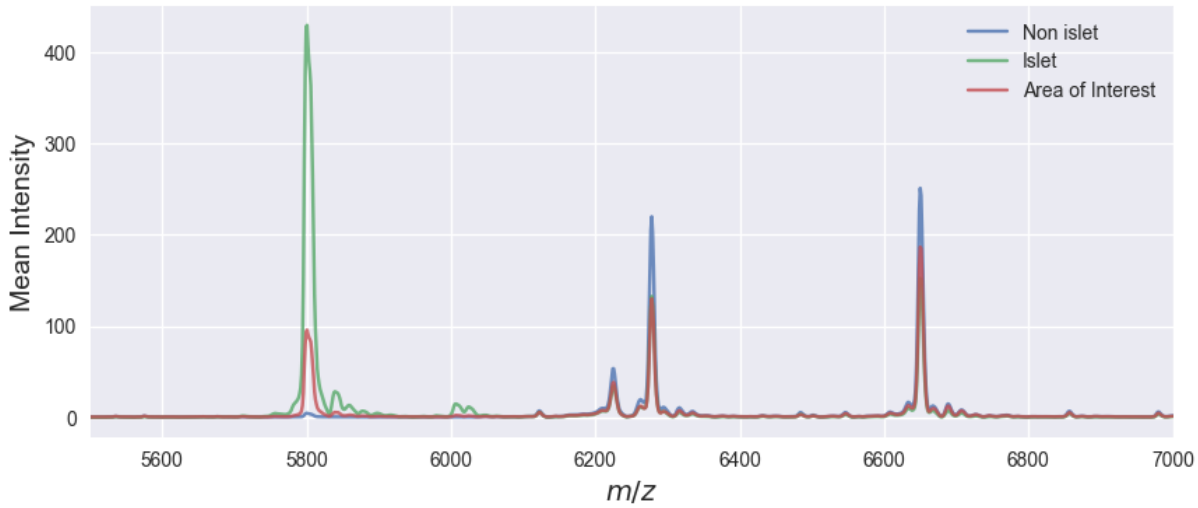


Figure 6.6: The Mean Intensity of AoI. AoI has different distribution than the islet and non-islet.

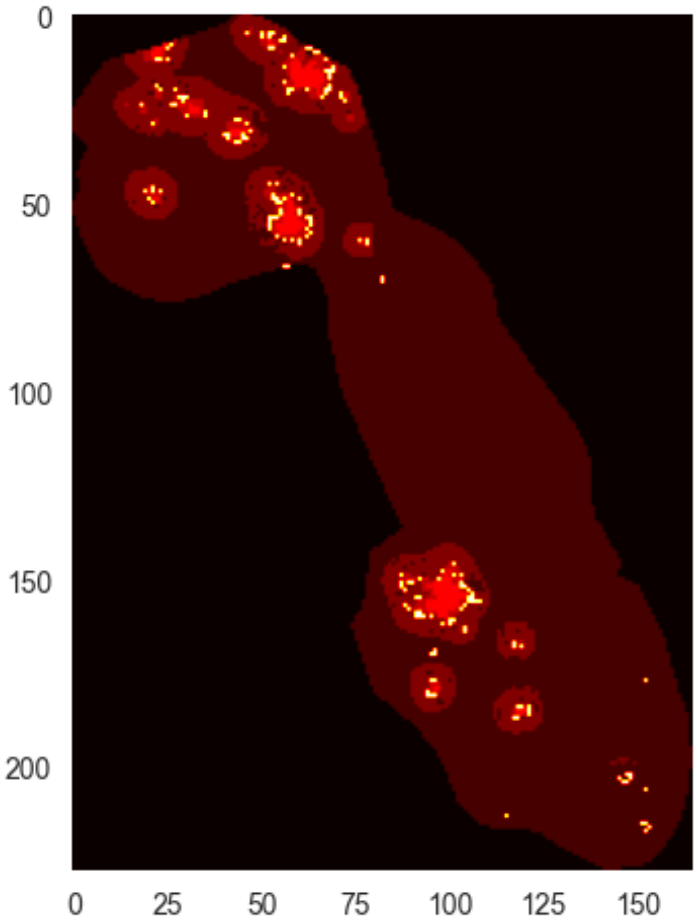
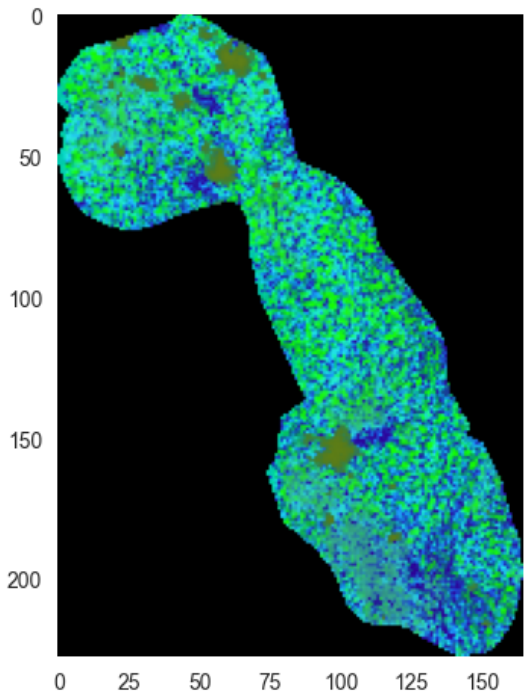


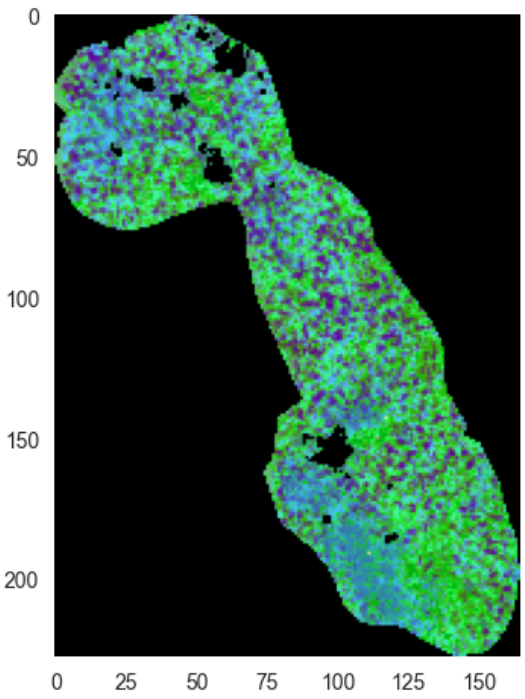
Figure 6.7: The Distribution of the Pixels within the AoI. The islets are highlighted with red and the AoI is distributed surrounding the islets.

6.3.2 UMAP Analysis of Non-Islet Pixels

We apply UMAP analysis to the non-islet pixels to observe whether we can gather further information that is masked by the insulin. We apply UMAP to the full data and data without the islet pixels and visualise the result of UMAP analysis (Figure 6.8). Since the islets are omitted, they are coloured as black (Figure 6.8b). Even though the assigned colours differ, no major changes are visualised when the analysis of full data (Figure 6.8a) and the the analysis of non-islet data (Figure 6.8b) are compared. Omitting the islets of Langerhans does not provide additional structural organisation of the non-islet pixels in our sample.



(a) UMAP Analysis of All Data. Islets are coloured with brown.



(b) UMAP Analysis of Non-Islet Pixels. The Islets are coloured with black.

Figure 6.8: UMAP Analysis of MSI Data. (1) Analysis is applied to all data, (2) analysis is applied to the non-islet data.

6.3.3 Removal of the Insulin-related Peaks

In previous subsections, we investigated whether there is additional information we can gather by identifying the map of the islets. In this and the following subsections, we want to investigate whether we can gather further insights about the islets of Langerhans, which might have been masked due to the abundance of insulin, by removing the insulin and insulin-related peaks from the data.

We only use the m/z ratios that are found in both models, Model 1 and Model 3, to be cautious (Table 6.4). Since Model 1 investigates the m/z ratios that are correlated with $Ins2$, the ratios that are found in Model 1 but not in Model 3 might be other compounds that are co-located with insulin, such as glucagon [11], but not insulin-related peaks.

We simply summed the GMM components of Model 1 - the correlation value of the peak m/z^i ratio over the m/z spectrum whose peaks values correspond to the correlation with $Ins2$, and capped at 1 (if the sum $1 -$ sum $= 1$) (Figure 6.9a).

$$\rho^{\mathrm{m/z}^i} = \sum_{j=0}^{14000} K\left(\rho^{\mathrm{m/z}_j} \mid \rho^{\mathrm{m/z}^i}\right), \tag{6.2}$$

$$\rho^{\mathrm{m/z}_j^i} = \begin{cases} \rho^{\mathrm{m/z}_j^i} = \rho^{\mathrm{m/z}_j^i} & \text{if } \rho^{\mathrm{m/z}_j^i} \leq 1, \\ \rho^{\mathrm{m/z}_j^i} = 1 & \text{if } \rho^{\mathrm{m/z}_j^i} > 1. \end{cases}, \tag{6.3}$$

$$\rho^{\mathrm{m/z}} = [\rho^{\mathrm{m/z}^0}, \rho^{\mathrm{m/z}^1}, ..., \rho^{\mathrm{m/z}^{14000}}], \tag{6.4}$$

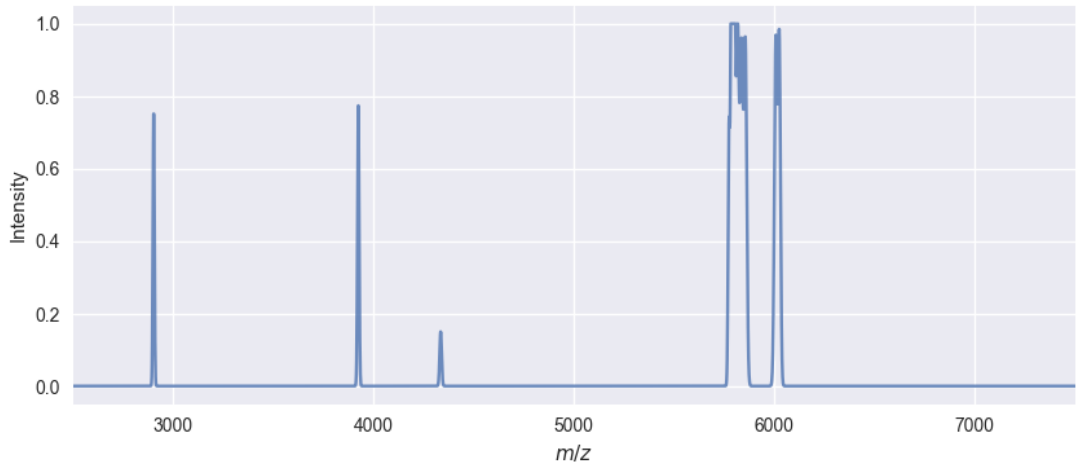
where $K\left(\rho^{\mathrm{m/z}_j} \mid \rho^{\mathrm{m/z}^i}\right)$ is the correlation value of the j-th m/z ratio for the given m/z^i.

Now we have correlation of the insulin-related peaks within the islets, $\rho^{\mathrm{m/z}}$. To remove the related peaks, for the pixels within the islets of Langerhans, $D_{insulin}^{islet}$, we multiply the insulin-related peak correlations with the intensity of $Ins2$ to get the intensity of the insulin-related peaks $I_{artefacts}$ (Figure 6.9b). Then we remove it from the pixel data. Now, we have the reduced islet data, $D_{insulin}^{reduced\_islet}$, which does not contain insulin and insulin-related peaks.

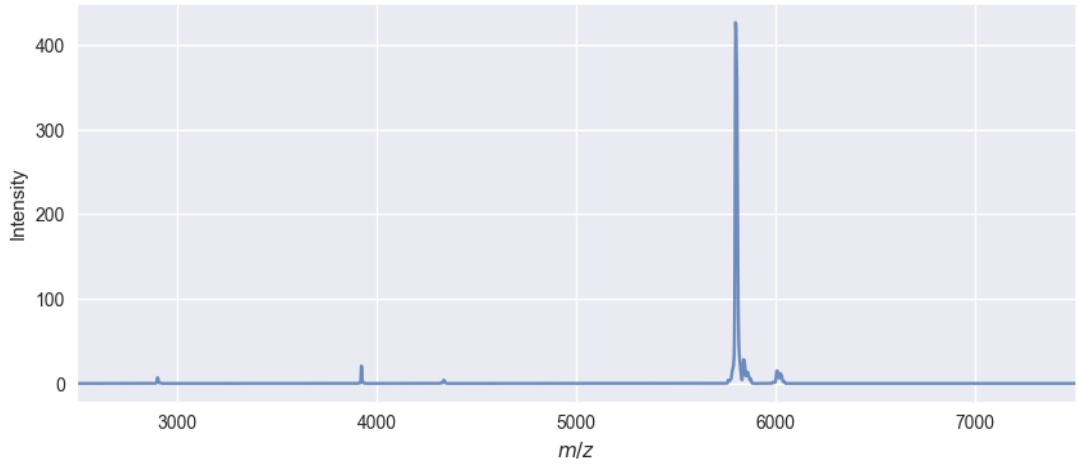
$$D_{insulin}^{islet} = D_{insulin}[label = islet], \tag{6.5}$$

$$I_{aretefacts} = \rho^{\mathrm{m/z}} \circ D_{insulin}^{islet}, \tag{6.6}$$

$$D_{insulin}^{reduced\_islet} = D_{insulin}^{islet} - I_{aretefacts}. \tag{6.7}$$



(a) Combined Insulin-related Peaks GMM Components. Insulin-Related Peak components are summed together.



(b) The Mean Intensity of Insulin-Related Peaks. The mean intensity of the insulin-related peaks within the islets are calculated by multiplying combined insulin-related peak component correlation with $Ins2$ intensity.

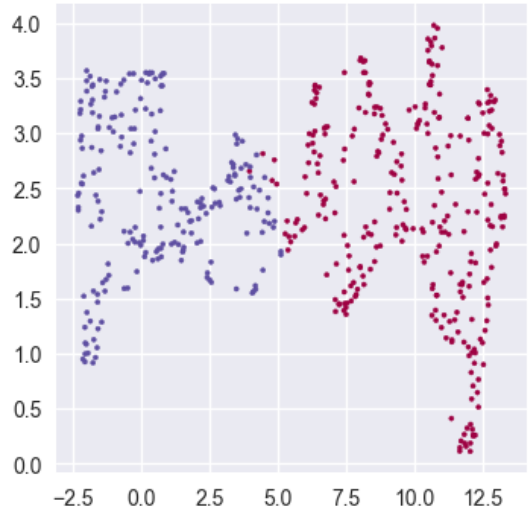
Figure 6.9: Detected Insulin-related Peaks GMM Components. (a) The correlation of the m/z ratios, (b) The mean intensity of the m/z ratios.

6.3.4 Effect of Insulin-Related Peak Removal from the Islets on UMAP Analysis

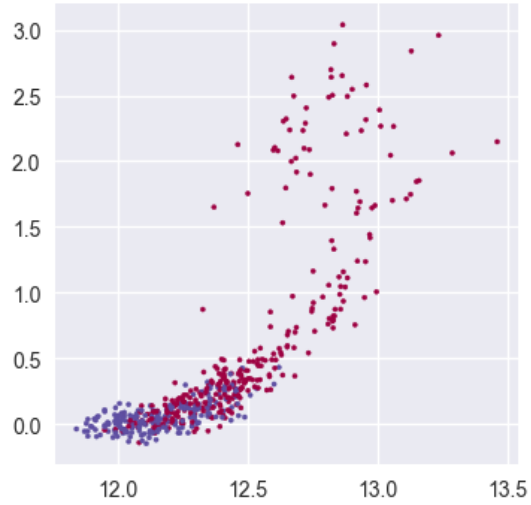
Insulin and insulin-related peaks may mask some of the organisation within the islets of Langerhans, and we want to see whether we can gather more insights from the reduced islet data $D_{insulin}^{reduced\_islet}$ which does not contain insulin and insulin-related peaks.

We first apply UMAP to the islet data and colour the pixels with their corresponding structures (Figure 6.10). Magenta corresponds to the inner parts of the islets and violet corresponds to the outer parts of the islets (still inside of the islets but close to the border). When insulin and insulin-related peaks are present, there is a clear distinction between the clusters (Figure 6.10a), however, when the insulin-related peaks are removed,

we see a continuum (Figure 6.10b). This indicates that there is not a strict difference in the intensities of non-insulin-related peaks m/z ratio between the inner and outer parts of the islets. Instead, some structures are common to both the inner and outer parts, along with structures that belong to only one of them. In addition, it can be seen that the inner parts are much more spread apart, whilst the outer parts are more tightly clustered. This suggests that there is more variation in the structures that are found in the inner parts, compared to those found in the outer parts.



(a) UMAP Analysis of the Original Islet. The colour represent 2 clusters (see Chapter 3) within the islets.



(b) UMAP Analysis of the Reduced Islet. The colour represent 2 clusters (see Chapter 3) within the islets.

Figure 6.10: The Effect of Insulin-Related Peak Removal on UMAP Analysis. Datapoints represent pixels. (a) UMAP analysis on original islet data, (b) UMAP analysis on reduced islet data.

6.4 Conclusion

We began this chapter by assessing the candidate insulin-related peaks by comparing their intensities at the islets of Langerhans and non-islets. Later, we compared the candidates identified by different approaches. It was observed that the candidates identified by the Sliding Window Approach were also identified by the Correlation Analysis and GMM Approach (Table 6.4 and Figure 6.4). We hypothesize that the Correlation Analysis and GMM Approach, in addition to identifying the insulin-related peaks, also identify the m/z ratios that are co-localized with $Ins2$, resulting in a higher number of candidates.

UMAP analysis was conducted to further investigate the data and gather additional information. In Figure 6.8a, some of the pixels labeled as non-islets neighbor the pixels labeled as islets. When the spatial distribution of the pixels is examined (Figure 6.6) and the m/z composition of these pixels is observed (Figure 6.7), it is suggested that these pixels might correspond to a different type of cellular compartment that shares a similar molecular composition with both islets and non-islets.

UMAP analysis was also performed on the non-islets (Figure 6.8) and on the islets of Langerhans after removing the insulin-related peaks (Figure 6.10). Removing the insulin-related peaks indicated that the molecular composition of the islets, except for the insulin-related peaks, is quite similar and does not differ significantly when the islets are divided into inner islet and outer islet regions. However, omitting the islets did not provide any additional information.

Chapter 7

Conclusions

This chapter is written by both authors.

The aim of this thesis is twofold. Initially, we attempt to construct the map of the islets of Langerhans in our tissue sample through the use of edge detection and clustering algorithms. Following this, the goal is to identify insulin-related peaks that are present in the data such that they can be removed to allow for more detailed analyses.

When constructing the map of the islets of Langerhans, edge detection is used to obtain an initial map. Binary dilation is then applied to make the map less restrictive. Finally, three clustering algorithms are used to attain a well defined map, namely HDBSCAN, Fuzzy K and K-means. Utilising the K-means algorithm produces a map that aligns most readily with domain knowledge, and so is the approach that is chosen to create the final map of the islets of Langerhans.

Two separate methods are used to attempt to identify insulin-related peaks that are present. The first method utilises Pearson correlation and Gaussian Mixture Models, while the second approach compares spatial distributions of m/z values, before applying correlation analysis. The number of candidate insulin-related peaks returned by the first method is 52, whereas only 19 are detected by the second method. This could be due to the first model detecting m/z values that are co-located with insulin whilst not being peaks that are related to insulin, such as glucagon. When testing whether the candidate m/z values are more abundant in the islet cells compared to non-islet cells, 12 values are rejected from the first method's output, and 2 values are rejected from the second method's output. All candidate insulin-related peaks that were identified by the second method were also identified by the first method, indicating that these are true insulin-related peaks.

Finally, we explored to see whether it is possible to gain more insight into the structure of pancreatic tissue. UMAP is applied to the original data and it is seen that the cells located near the islets according to our labelling have 2 distinct groups. The molecular decomposition of one of the groups lies between the islets and the rest of the tissue, indicating the presence of another structure. We then applied UMAP to the non-islet data, however, no additional organisation was observed. We finally applied UMAP to both the original data and the data after the insulin-related peaks have been removed.

The results suggest that there is not a definitive difference between the inner and outer parts of the islets as indicated by the mapping on the original data, but instead there is a gradual change in the structure of the islets.

In this dissertation, we applied two different methods that labeled similar m/z ratios as insulin-related peaks. The spatial distribution test also confirmed that these insulin-related peaks are found within the islets. The results indicate that a combination of Pearson correlation analysis with GMMs, along with spatial comparison and correlation analysis methods, is promising for identifying additional peaks created during the MSI process. Identifying these peaks can lead to greater insights into the data.

In future steps, we would consult with relevant literature to compare our results with molecules whose m/z values are already known to science. Unfortunately, the exact values for many molecules are still unknown, which would make the process challenging. However, it could provide further insight into our results, and what they represent. We did not apply this approach here as it is beyond the scope of what this thesis was trying to achieve.

Appendix

The GitHub repository for Chapter 3, Chapter 4 and Chapter 6, and the GitHub repository for Chapter 5 are available.

Bibliography

[1] Walid M. Abdelmoula et al. "Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of Mass spectrometry imaging data". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.43 (2016). ISSN: 10916490. DOI: 10.1073/pnas.1510227113.

[2] Walid M. Abdelmoula et al. "Peak learning of mass spectrometry imaging data using artificial neural networks". In: *Nature Communications* 12.1 (2021). ISSN: 20411723. DOI: 10.1038/s41467-021-25744-8.

[3] Bruce Alberts et al. *Molecular Biology of the Cell*. 2007. DOI: 10.1201/9780203833445.

[4] Theodore Alexandrov. *MALDI imaging mass spectrometry: statistical data analysis and current computational challenges*. 2012. DOI: 10.1186/1471-2105-13-s16-s11.

[5] Susan Ashton et al. "Aurora kinase inhibitor nanoparticles target tumors with favorable therapeutic index in vivo". In: *Science Translational Medicine* 8.325 (2016). ISSN: 19466242. DOI: 10.1126/scitranslmed.aad2355.

[6] Mitra Basu. "Gaussian-based edge-detection methods - A survey". In: *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* 32.3 (Aug. 2002), pp. 252–260. ISSN: 10946977. DOI: 10.1109/TSMCC.2002.804448.

[7] Alex Bateman et al. "UniProt: the Universal Protein Knowledgebase in 2023". In: *Nucleic Acids Research* 51.D1 (2023). ISSN: 13624962. DOI: 10.1093/nar/gkac1052.

[8] Birte Beine et al. "Tissue maldi mass spectrometry imaging (MALDI MSI) of peptides". In: *Methods in Molecular Biology*. Vol. 1394. 2016. DOI: 10.1007/978-1-4939-3341-9{_}10.

[9] Amanda Rae Buchberger et al. *Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights*. 2018. DOI: 10.1021/acs.analchem.7b04733.

[10] Kenneth P. Burnham and David R. Anderson. *Multimodel inference: Understanding AIC and BIC in model selection*. 2004. DOI: 10.1177/0049124104268644.

[11] Over Cabrera et al. "The unique cytoarchitecture of human pancreatic islets has implications for islet cell function". In: *Proceedings of the National Academy of Sciences of the United States of America* 103.7 (2006). ISSN: 00278424. DOI: 10.1073/pnas.0510790103.

[12] Ricardo J.G.B. Campello, Davoud Moulavi, and Joerg Sander. "Density-based clustering based on hierarchical density estimates". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 7819 LNAI. PART 2. 2013. DOI: 10.1007/978-3-642-37456-2{_}14.

[13] Robert D. Cardiff, Claramae H. Miller, and Robert J. Munn. "Manual hematoxylin and eosin staining of mouse tissue sections". In: *Cold Spring Harbor Protocols* 2014.6 (2014). ISSN: 15596095. DOI: 10.1101/pdb.prot073411.

[14] Matteo D'Aloia, Annalisa Longo, and Maria Rizzi. "Noisy ECG signal analysis for automatic peak detection". In: *Information (Switzerland)* 10.2 (2019). ISSN: 20782489. DOI: 10.3390/info10020035.

[15] Gabriela Da Silva Xavier. *The cells of the islets of langerhans*. 2018. DOI: 10.3390/jcm7030054.

[16] Chhabil Dass. *Fundamentals of Contemporary Mass Spectrometry*. 2006. DOI: 10.1002/9780470118498.

[17] N. E. DAY. "Estimating the components of a mixture of normal distributions". In: *Biometrika* 56.3 (1969). ISSN: 0006-3444. DOI: 10.1093/biomet/56.3.463.

[18] Jurij Dolenšek, Marjan Slak Rupnik, and Andraž Stožer. *Structural similarities and differences between the human and the mouse pancreas*. 2015. DOI: 10.1080/19382014.2015.1024405.

[19] Felix Draude et al. "Characterization of freeze-fractured epithelial plasma membranes on nanometer scale with ToF-SIMS". In: *Analytical and bioanalytical chemistry* 407.8 (2015). ISSN: 16182650. DOI: 10.1007/s00216-014-8334-2.

[20] A A Elayat, M M el-Naggar, and M Tahir. "An immunocytochemical and morphometric study of the rat pancreatic islets." In: *Journal of anatomy* 186 (Pt 3) (1995). ISSN: 0021-8782.

[21] Andrew H Fischer et al. "Hematoxylin and Eosin (H & E) staining". In: *CSH protocols* 2008.4 (2005). ISSN: 00165085.

[22] Timo Gaber, Cindy Strehl, and Frank Buttgereit. *Metabolic regulation of inflammation*. 2017. DOI: 10.1038/nrrheum.2017.37.

[23] Juliana P.L. Gonçalves, Christine Bollwein, and Kristina Schwamborn. *Mass Spectrometry Imaging Spatial Tissue Analysis toward Personalized Medicine*. 2022. DOI: 10.3390/life12071037.

[24] L. C. Groop, E. Widén, and E. Ferrannini. "Insulin resistance and insulin deficiency in the pathogenesis of Type 2 (non-insulin-dependent) diabetes mellitus: errors of metabolism or of methods?" In: *Diabetologia* 36.12 (1993). ISSN: 0012186X. DOI: 10.1007/BF00400814.

[25] Group. Lancet. "Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). UK Prospective Diabetes Study (UKPDS)". In: *Lancet* 352.837 (1998).

[26] Sven Heiles. *Advanced tandem mass spectrometry in metabolomics and lipidomics—methods and applications*. 2021. DOI: 10.1007/s00216-021-03425-1.

[27] Abiodun M. Ikotun et al. "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data". In: *Information Sciences* 622 (2023). ISSN: 00200255. DOI: 10.1016/j.ins.2022.11.139.

[28] Rima Kaddurah-Daouk, Bruce S. Kristal, and Richard M. Weinshilboum. *Metabolomics: A global biochemical approach to drug response and disease*. 2008. DOI: 10.1146/annurev.pharmtox.48.113006.094715.

[29] Julia Laskin et al. "Tissue imaging using nanospray desorption electrospray ionization mass spectrometry". In: *Analytical Chemistry* 84.1 (2012). ISSN: 00032700. DOI: 10.1021/ac2021322.

[30] Allison N. Lau and Matthew G. Vander Heiden. *Metabolism in the Tumor Microenvironment*. 2020. DOI: 10.1146/annurev-cancerbio-030419-033333.

[31] Olalla López-fernández et al. *Determination of polyphenols using liquid chromatography–tandem mass spectrometry technique (LC–MS/MS): A review*. 2020. DOI: 10.3390/antiox9060479.

[32] Chao Lu and Craig B. Thompson. *Metabolic regulation of epigenetics*. 2012. DOI: 10.1016/j.cmet.2012.06.001.

[33] Siyuan Ma et al. *High spatial resolution mass spectrometry imaging for spatial metabolomics: Advances, challenges, and future perspectives*. 2023. DOI: 10.1016/j.trac.2022.116902.

[34] M. Mann, R. C. Hendrickson, and A. Pandey. *Analysis of proteins and proteomes by mass spectrometry*. 2001. DOI: 10.1146/annurev.biochem.70.1.437.

[35] Raymond E. March. *An introduction to quadrupole ion trap mass spectrometry*. 1997. DOI: 10.1002/(SICI)1096-9888(199704)32:4<351::AID-JMS512>3.0.CO;2-Y.

[36] Leland McInnes et al. "UMAP: Uniform Manifold Approximation and Projection". In: *Journal of Open Source Software* 3.29 (2018). DOI: 10.21105/joss.00861.

[37] Eugene N. Nikolaev, Yury I. Kostyukevich, and Gleb N. Vladimirov. *Fourier transform ion cyclotron resonance (FT ICR) mass spectrometry: Theory and simulations*. 2016. DOI: 10.1002/mas.21422.

[38] Peter Juma Ochieng et al. "Adaptive Savitzky–Golay Filters for Analysis of Copy Number Variation Peaks from Whole-Exome Sequencing Data". In: *Information (Switzerland)* 14.2 (2023). ISSN: 20782489. DOI: 10.3390/info14020128.

[39] Andrew Palmer, Dennis Trede, and Theodore Alexandrov. "Where imaging mass spectrometry stands: here are the numbers". In: *Metabolomics* 12.6 (2016). ISSN: 15733890. DOI: 10.1007/s11306-016-1047-0.

[40] Gary J. Patti, Oscar Yanes, and Gary Siuzdak. *Innovation: Metabolomics: the apogee of the omics trilogy*. 2012. DOI: 10.1038/nrm3314.

[41] Ralph G. Pearson. "Hard and Soft Acids and Bases". In: *Journal of the American Chemical Society* 85.22 (1963). ISSN: 15205126. DOI: 10.1021/ja00905a001.

[42] Daniel Petras, Alan K. Jarmusch, and Pieter C. Dorrestein. *From single cells to our planet—recent advances in using mass spectrometry for spatially resolved metabolomics*. 2017. DOI: 10.1016/j.cbpa.2016.12.018.

[43] Alexandra van Remoortere et al. "MALDI imaging and profiling MS of higher mass proteins from tissue". In: *Journal of the American Society for Mass Spectrometry* 21.11 (2010). ISSN: 10440305. DOI: 10.1016/j.jasms.2010.07.011.

[44] Svitlana Rozanova et al. "Quantitative Mass Spectrometry-Based Proteomics: An Overview". In: *Methods in Molecular Biology*. Vol. 2228. 2021. DOI: 10.1007/978-1-0716-1024-4{_}8.

[45] Stanislav S. Rubakhin et al. *Imaging mass spectrometry: Fundamentals and applications to drug discovery*. 2005. DOI: 10.1016/S1359-6446(05)03458-6.

[46] Enrique H. Ruspini. "A new approach to clustering". In: *Information and Control* 15.1 (1969). ISSN: 00199958. DOI: 10.1016/S0019-9958(69)90591-9.

[47] Seong-Dae Kim, Jeong-Hwan Lee, and Jae-Kyoon Kim. "A new chain-coding algorithm for binary images using run-length codes". In: *Computer Vision, Graphics, & Image Processing* 41.1 (1988). ISSN: 0734189X. DOI: 10.1016/0734-189x(88)90121-1.

[48] Gil Sharon et al. *Specialized metabolites from the microbiome in health and disease*. 2014. DOI: 10.1016/j.cmet.2014.10.016.

[49] Meng Shin Shiao et al. "Adaptive evolution of the insulin two-gene system in mouse". In: *Genetics* 178.3 (2008). ISSN: 00166731. DOI: 10.1534/genetics.108.087023.

[50] Shuichi Shimma and Yuki Sugiura. "Effective Sample Preparations in Imaging Mass Spectrometry". In: *Mass Spectrometry* 3.Special_Issue (2014). ISSN: 2187-137X. DOI: 10.5702/massspectrometry.s0029.

[51] Lekha Sleno and Dietrich A. Volmer. *Ion activation methods for tandem mass spectrometry*. 2004. DOI: 10.1002/jms.703.

[52] Marcus Svensson et al. "Heat stabilization of the tissue proteome: A new technology for improved proteomics". In: *Journal of Proteome Research* 8.2 (2009). ISSN: 15353893. DOI: 10.1021/pr8006446.

[53] John G. Swales et al. "Mass spectrometry imaging of cassette-dosed drugs for higher throughput pharmacokinetic and biodistribution analysis". In: *Analytical Chemistry* 86.16 (2014). ISSN: 15206882. DOI: 10.1021/ac502217r.

[54] Magdalena K. Sznurkowska et al. "Tracing the cellular basis of islet specification in mouse pancreas". In: *Nature Communications* 11.1 (2020). ISSN: 20411723. DOI: 10.1038/s41467-020-18837-3.

[55] Adam J. Taylor, Alex Dexter, and Josephine Bunch. "Exploring Ion Suppression in Mass Spectrometry Imaging of a Heterogeneous Tissue". In: *Analytical Chemistry* 90.9 (2018). ISSN: 15206882. DOI: 10.1021/acs.analchem.7b05005.

[56] Victoria L. Tokarz, Patrick E. MacDonald, and Amira Klip. *The cell biology of systemic insulin function*. 2018. DOI: 10.1083/jcb.201802095.

[57] Jan Urban, Nils Kristian Afseth, and Dalibor Štys. *Fundamental definitions and confusions in mass spectrometry about mass assignment, centroiding and resolution*. 2014. DOI: 10.1016/j.trac.2013.07.010.

[58] Maxence Wisztorski et al. "MALDI direct analysis and imaging of frozen versus FFPE tissues: What strategy for which sample?" In: *Methods in Molecular Biology* 656 (2010). ISSN: 10643745. DOI: 10.1007/978-1-60761-746-4{_}18.

[59] Ying Xi et al. "SMART: A data reporting standard for mass spectrometry imaging". In: *Journal of Mass Spectrometry* 58.2 (2023). ISSN: 10969888. DOI: 10.1002/jms.4904.

Faculty of Science
Kasteelpark Arenberg 11 - box 2100
3001 Heverlee, BELGIË
tel. + 32 16 32 14 01
fax + 32 16 32 83 10
www.kuleuven.be