



UNIVERSITY OF MILANO-BICOCCA

School of Science

Department of Informatics, Systems and Communication

Master's Degree course in Data Science

Improving distance-based clustering methods with spatial information for mass spectrometry imaging data

Supervisor: Prof. Paola Rebora

Co-supervisor: Dr. Giulia Capitoli

Master's thesis of:

Youssef Benchaib

ID 844515

Academic Year 2022-2023

Abstract

Mass spectrometry is emerging as a valuable tool that allows the detection of biological traits that would be overlooked with a simple visual morphological assessment of a sample by measuring the abundance of several specific molecules over multiple locations of a biological sample. This thesis will focus on the use of different metrics of distance used in clustering for the analysis of protein mass spectrometry data. Alternative methods will be compared with those currently considered standard (i.e., Euclidean distance between mass spectra) to improve their performance. The thesis will aim to use these methods to identify groups of spots that present similar patterns to extract interesting insights, such as anomalies that pathologists cannot capture from the original morphological image. In addition, it is intended to develop a method of spatial penalization within clustering algorithms in order to consider the geographical location of the spot, thus improving the consistency and interpretation of the results obtained. The results of the entire process will then be compared with a goal standard to verify the actual improvement in the results obtained.

Contents

1	Introduction	2
1.1	Importance of Clustering in MSI Data Analysis	2
1.2	Motivation and aim of the thesis	2
2	Background on Mass Spectrometry Imaging	3
2.1	Mass Spectrometry Data	3
2.2	Basics of Mass Spectrometry Imaging (MSI) Data	3
2.3	Workflow of MSI	4
2.4	Recent Advances in MSI	5
2.5	Common Distance Metrics in MSI: Advantages and Limitations .	6
2.6	Integration of Spatial Information in Clustering	6
3	Data and Methodology	7
3.1	Data Description and Preprocessing	7
3.2	Clustering Methods	9
3.3	Integration of Spatial Information into Clustering	14
3.4	Evaluation of the clustering performance	17
4	Results	20
4.1	Results for Traditional Clustering Methods	20
4.2	Results for Spatial-Enhanced Clustering Algorithms	26
5	Discussion	34
5.1	Insights from the Resulting Clusters	34
5.2	Consistency and Interpretability of the Results	34
6	Conclusions and Future Work	34
6.1	Key Findings and Implications	34
6.2	Limitations and Challenges	35
6.3	Recommendations for Future Research	35
A	Implementation of Algorithms and Code Fragments	36
1	Hierarchical Clustering (HCLUST)	36
2	K-means	37
3	Spatial Penalization	38

1 Introduction

Mass spectrometry imaging (MSI) is a powerful technique that combines the precision of mass spectrometry with spatial localization to investigate the molecular composition of biological samples. As an analytical tool, it allows researchers to obtain detailed spatially resolved molecular maps, which reveal the distribution of proteins, lipids, and other biomolecules across a sample's surface.[7]

The uniqueness of MSI lies in its ability to produce abundant information without the need for molecular labels¹, which can sometimes interfere with the original molecular state of the specimen. Its capability to detect a wide range of molecules across multiple locations of a biological sample provides a comprehensive molecular snapshot, which offers researchers a richer understanding of biological systems.

1.1 Importance of Clustering in MSI Data Analysis

MSI generates vast amounts of complex, high-dimensional data. With numerous spectra acquired from different spatial regions of a sample, making sense of this data necessitates advanced data analysis techniques. Clustering emerges as a fundamental tool in this context. By grouping together spectra that are similar, clustering facilitates the identification of patterns and trends within the data. It helps in condensing large datasets into coherent groups, wherein each cluster represents a particular molecular profile or signature. The ability to identify these molecular profiles has significant implications for various fields, especially in health sciences, where it can help in identifying anomalies or distinguishing between diseased and healthy tissues.

1.2 Motivation and aim of the thesis

The effectiveness of clustering depends on the choice of the distance metric used to measure the similarity (or dissimilarity) between data points. The commonly adopted metric, the Euclidean distance, "however, the choice of this distance measure can have a considerable impact on the results, as meaningful patterns might be lost due to the high-dimensionality of the data, and noise or outliers can become amplified." [23]

Additionally, traditional clustering methods in MSI, while effective in some cases, do not include the "geographical/spatial" distance, that can be derived by MSI and that can be very important for the clustering. For instance, two spots that are geographically close and show similar mass spectra might be more biologically relevant if considered together, as opposed to two spots that are far apart but have similar mass spectra. Thus, incorporating spatial information can provide a more accurate representation of the biological sample and enhance the precision of clustering.

This thesis aims to address these challenges by exploring alternative distance metrics for clustering MSI data and comparing their performance versus a pre-specified gold-standard in order to include their important characteristics in the clustering. Moreover, a secondary objective of the thesis is to integrate spatial information into the clustering process to test a potential improvement. The overarching goal is to enhance the inter-

¹Traditional labeling techniques, such as fluorescent, radioactive, or enzymatic tags, can potentially alter the molecule's natural behavior or require complex preparation steps. MSI circumvents these challenges by directly mapping the spatial distribution of molecules based on their mass, offering a more straightforward and potentially unbiased view of a sample's molecular landscape.

pretability and reliability of MSI data analysis, ensuring that the insights extracted are both scientifically relevant and biologically consistent.

In the chapters that follow, a thorough examination of existing methodologies, the introduction of proposed methods, experimental results, and their implications will be presented, all centered on the core objective of refining the clustering methods for MSI data.

2 Background on Mass Spectrometry Imaging

Mass spectrometry (MS) is a pivotal analytical technique with qualitative and quantitative capabilities, predominantly employed to identify and quantify an array of analytes with clinical relevance. The synergy of MS with gas or liquid chromatographs has significantly broadened the analytical scope for various clinical applications. Given its adeptness at identifying and quantifying proteins, MS has cemented its position as an indispensable tool in proteomics.

2.1 Mass Spectrometry Data

Mass spectrometry (MS) is a versatile analytical technique used to determine the mass, chemical composition, and molecular structure of individual molecules. In essence, MS identifies the amount and type of chemicals present in a sample by producing spectra of the masses of their molecules and molecular fragments.

The process begins by ionizing chemical compounds to generate charged molecules or molecule fragments and measuring their mass-to-charge ratios. The ionized molecules are then separated based on their mass-to-charge ratio and detected. This produces a spectrum that represents the distribution of ions in the sample.

There are several methods to ionize samples, including electron impact and matrix-assisted laser desorption/ionization (MALDI). The choice of ionization method can depend on the specific molecular species in the sample and the desired information.

One of the fundamental parameters in mass spectrometry is the mass-to-charge ratio, often represented by the symbol m/z . It denotes the ratio of the mass (m) of an ion to its charge (z). In the simplest terms, m/z describes how heavy an ion is relative to its charge.

In MS, most ions have a single positive or negative charge, making the m/z value equivalent to the mass of the ion. However, ions can sometimes carry multiple charges, altering the m/z value. This aspect is crucial because the MS instrument does not directly measure the mass of ions; instead, it measures how they behave under the influence of an electric or magnetic field, which is determined by their m/z value.

When interpreting a mass spectrum, the x-axis typically displays the m/z values, while the y-axis shows the intensity or relative abundance of ions. Peaks in a mass spectrum represent ions with specific m/z values, with the height of the peak indicating the ion's relative abundance. Understanding the m/z values is essential for deciphering the chemical composition and structure of molecules in the sample.[11]

2.2 Basics of Mass Spectrometry Imaging (MSI) Data

Mass spectrometry imaging (MSI) has emerged as a transformative analytical tool, allowing for detailed and comprehensive exploration into the spatial distribution of molecular

species across a plethora of sample types. Its potential lies in its capability to simultaneously image an extensive array of molecules, spanning from metabolites, lipids, peptides, to proteins and glycans, all within a single experimental setup and crucially, without the requirement of any labeling.[7]

One of the fundamental advantages of MSI is the amalgamation of data from mass spectrometry (MS) which provides comprehensive molecular information, with the added dimension of spatial distribution. This spatial-molecular duality ensures a richer understanding of biological samples. The general workflow, as illustrated in Figure 1, commences with meticulous sample preparation. Following this, an (x, y) grid is superimposed onto the sample’s surface, with the grid’s dimensions being user-specified. The mass spectrometer subsequently ionizes molecules on the sample’s surface, extracting a mass spectrum from each pixel on the designated section. The spatial resolution, in this case, is dictated by the size of the pixel. Once the spectra have been acquired, computational tools can be engaged to extract a specific mass-to-charge (m/z) value from each pixel’s spectrum. These values are then synthesized into a heat map, thereby offering a visual representation of the relative abundance of that specific m/z value across the sample.[7]

2.3 Workflow of MSI

Following the acquisition of Mass Spectrometry Imaging (MSI) data, it is imperative to refine this raw data to ensure subsequent analyses are accurate and meaningful. This refinement process is initiated with a series of preprocessing steps:

1. **Baseline correction:** By removing low-frequency noise and any inherent trends, the authentic spectral features are made more prominent.
2. **Smoothing:** Applied to curtail high-frequency noise, this step aids in preserving the true molecular patterns within the data.
3. **Normalization:** This ensures consistent intensity values across the dataset, counteracting potential variability that may arise from disparate samples or different acquisition sessions.
4. **Spectral recalibration:** Ensures aligned spectra across the dataset, thereby maintaining consistent peak positions.

Post-preprocessing, addressing the data’s high dimensionality and intricacy is paramount. This is tackled through *data compression*, streamlining the dataset to reduce computational intensity for ensuing analyses. Data compression can be bifurcated into:

- *Supervised compression*, where categorization is based on predefined groupings.
- *Unsupervised compression*, where the data’s inherent structure and relationships guide the compression without any prior labels. Techniques such as k-means, hierarchical clustering, and bisecting k-means are particularly efficacious for unsupervised compression, given their prowess in unearthing natural data groupings.

Once compressed, discriminant analyses are pivotal in evaluating the effectiveness of the selected classification or clustering techniques.

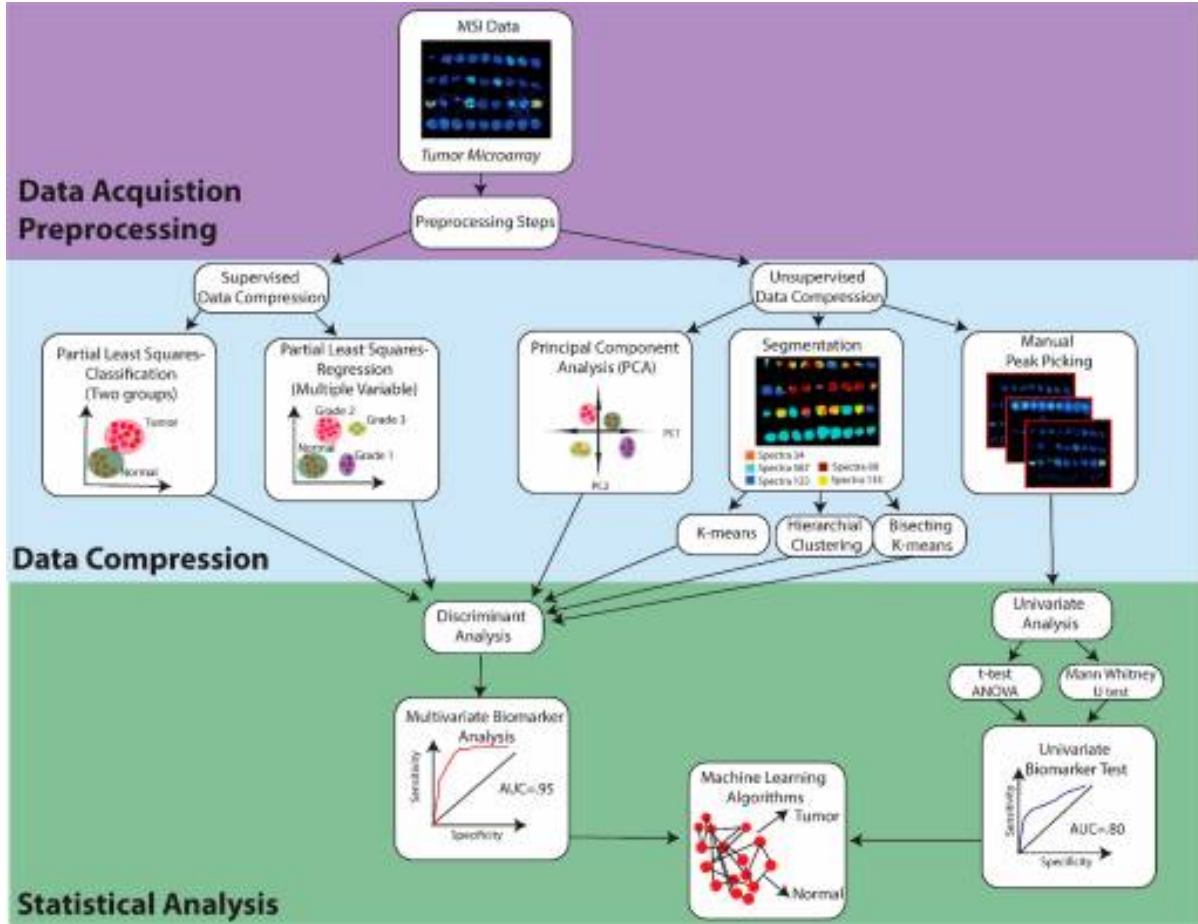


Figure 1: A descriptive workflow detailing the data processing and analysis for MSI data.[7]

In the realm of data processing and analysis for Mass Spectrometry Imaging (MSI) data, various strategies exist. The workflow presented in the accompanying Figure 1 provides an overview of the comprehensive steps involved in MSI data analysis. It's worth noting that this research specifically focuses on an unsupervised methodology. Such an approach is uniquely suited to the intricate nature of MSI data, aiming to discover inherent structures and patterns without relying on prior knowledge or labels. This strategy not only addresses the high-dimensional challenges but also endeavors to capture latent molecular patterns. By steering clear of external guidance, the unsupervised approach ensures a genuine extraction of insights, potentially revealing unanticipated patterns or relationships in the MSI dataset.

2.4 Recent Advances in MSI

The trajectory of MSI's evolution has been notably steep in the recent past. Technological innovations have propelled MSI to the forefront of clinical practices and pharmaceutical research arenas. Sample preparation techniques have been refined, ensuring consistent and reproducible data.[12] Instrumentation advancements have paved the way for swifter data acquisition coupled with enhanced spatial resolution. Further bolstering the credibility of MSI are methodologies enabling the absolute quantitation of detected molecules. To cope with the computational demands of vast imaging datasets generated by contemporary instruments, machine learning algorithms and statistical workflows have been

integrated into the analysis process. This integrative approach, blending hardware enhancements with software innovations, is carving the path for future developments in MSI, such as 3-dimensional renderings and single-cell resolution imaging.[7]

2.5 Common Distance Metrics in MSI: Advantages and Limitations

The success of clustering hinges largely on the chosen distance metric. In the context of MSI data analysis, the most commonly employed metric has been the Euclidean distance. Its mathematical simplicity and intuitive nature made it a preferred choice. When comparing mass spectra, it offers a straightforward method to gauge similarity based on the magnitude of individual peaks.

However, the Euclidean distance has its limitations. As first, it doesn't account for potential shifts in the mass-to-charge ratio (m/z) axis, which are frequent in MSI data. Additionally, it may not capture the true biological similarities between different spectra, especially when minor variations can result in significant Euclidean distances.

Over time, other metrics have been explored, such as the Manhattan distance, cosine similarity, and dynamic time warping (DTW). Each has its strengths and potential applications, with some offering robustness against shifts in the m/z axis (like DTW) or emphasizing angular differences (as with cosine similarity).

2.6 Integration of Spatial Information in Clustering

One of the unique facets of MSI (Mass Spectrometry Imaging) data is its inherent spatial context. Each mass spectrum corresponds to a specific location on the biological sample, a fact that has been under-utilized in traditional clustering approaches. Recognizing this gap, recent research has focused on integrating spatial information to enhance clustering outcomes.

Spatially-constrained clustering methods have been introduced, which factor in the geographical proximity of mass spectra when forming clusters. Such methods have shown promise in preserving tissue-specific or region-specific molecular profiles.

Another approach has been the incorporation of spatial penalization techniques, which prioritize clustering of neighboring spectra, thereby producing clusters that are geographically cohesive. This not only yields biologically meaningful clusters but also aids in the visualization and interpretation of results.

The preference for the latter approach, incorporating spatial penalization techniques, arises from several considerations. First, the method ensures a higher degree of homogeneity within clusters due to the prioritization of geographically proximate spectra. This often leads to more representative and distinct cluster centroids. Secondly, it aligns with the biological rationale that neighboring regions in a sample would possess similar molecular compositions, thereby emphasizing the importance of local spatial context. Lastly, the resulting clusters, being geographically cohesive, are easier to interpret and visualize, enhancing the practical utility of the clustering outcomes for researchers and practitioners.

3 Data and Methodology

3.1 Data Description and Preprocessing

The data used in this thesis derived by the application of the MALDI-MSI technology on a section of mouse brain tissue, we consider the sequential imaging of lipids, glycans, and tryptic peptides on a single section. MALDI-MSI uses a laser energy-absorbing matrix to create ions from large molecules with minimal fragmentation. In other words, this technology visualizes the distribution of molecules such as peptides, lipids, and glycans in a biological sample.[20] For this work, a slice from the mouse’s brain is partitioned into a grid of pixels. Then, MALDI-MSI acquires a mass spectrum for each pixel. In particular, MALDI-MSI creates a grid of pixel-by-pixel spectra. Each spectra measures the mass-to-charge (m/z) values representing analytes of interest along the x-axis and the corresponding abundance along the y-axis.[6] Finally, a three-dimensional MSI dataset is obtained by organizing the mass spectra alongside the grid coordinates for each respective pixel intended for deeper analysis. The entire process is illustrated in Figure 1.[3] Because of experimental constraints, noise in the MALDI-MSI spectra can influence the statistical evaluation. To address this concern, a well-established biological pipeline is executed.

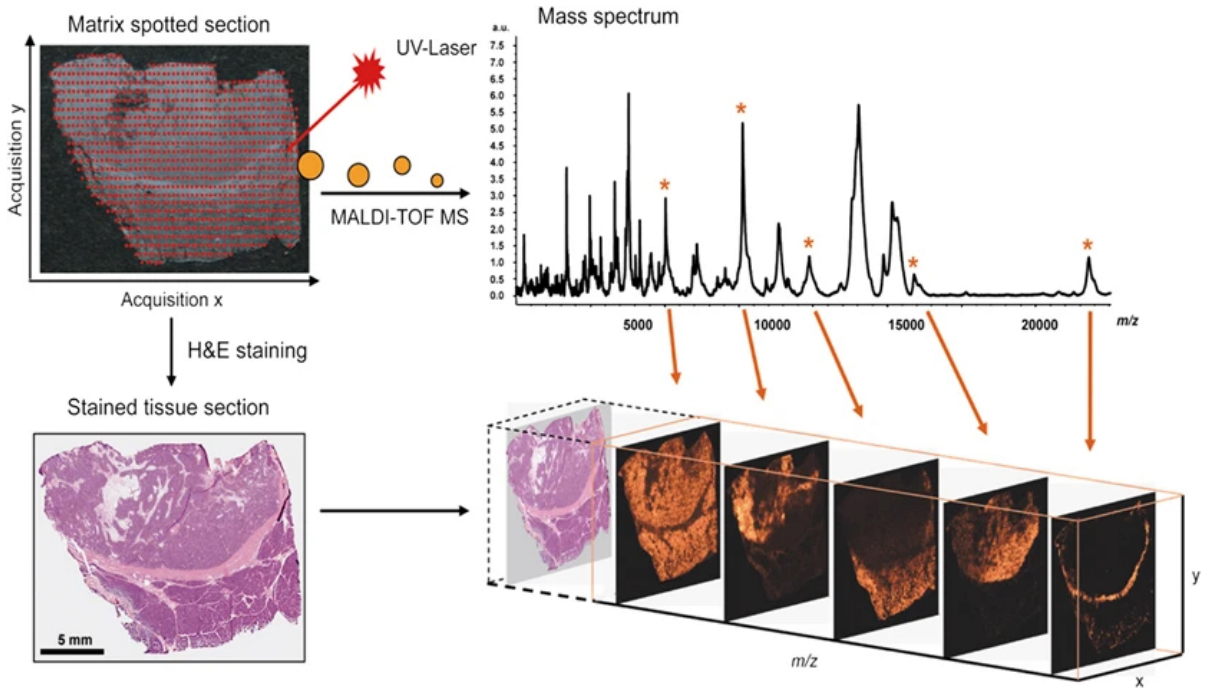


Figure 2: MALDI-MSI imaging workflow and spectrum, from Aichler [2015]

The data from the MALDI-MSI technology is first refined by filtering out any disturbances. This involves baseline adjustment, smoothening, standardizing, aligning spectra, and pinpointing and extracting peaks. It’s essential to minimize differences within each sample and rectify any discrepancies due to analysis and equipment after preparing the sample, ensuring precise m/z positioning. For this statistical study, an image of biological tissue examined with a 50- μm grid, representing about 18,000 laser shots (defining the pixels) is used. This process returns thousands of unique spectra, each labeled by pixel position (x, y).[8]

During the dataset analysis, an abundance of missing values was detected. This is

because whenever the incoming beam of molecules falls below the instrument threshold, a missing value is recorded. In this case, the missing values represent 53.58%. Guided by biological information, we interpret the lack as indicating the absence or presence below the detection threshold of a particular molecule. Therefore, all missing values were replaced with zeros. The work was carried out on a dataset that corresponds to the union of three separate datasets, all related to the same tissue section, each focusing on a specific range of m/z values and relative spatial coordinates of a specific macro-group of molecules: lipids, glycans and peptides.

The final dataset used is thus composed of 18'102 points of mouse brain tissue and the mass spectra analysis of 316 proteins contained within it with a vast range of the spectra (from 0.0 to 51.34 for lipids, from 0.0 to 79.37 for glycans and from 0.0 to 210.52 for peptides). Below in Figures 3, 4 and 5 mean and variance (on the left) and the coefficient of variation (on the right) of, in order, lipids, peptides, and glycans.

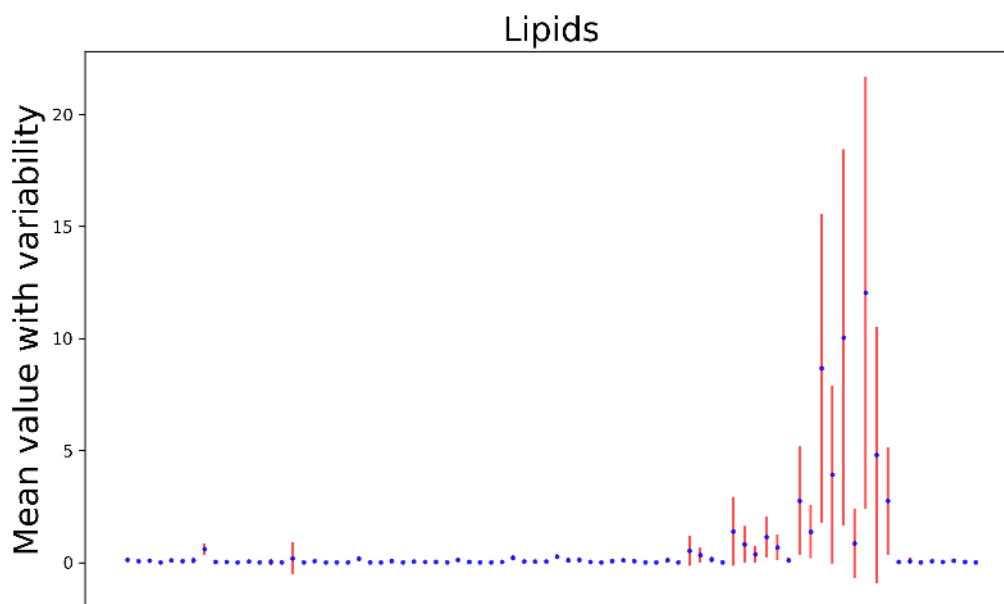


Figure 3: Lipids, Mean value with variability

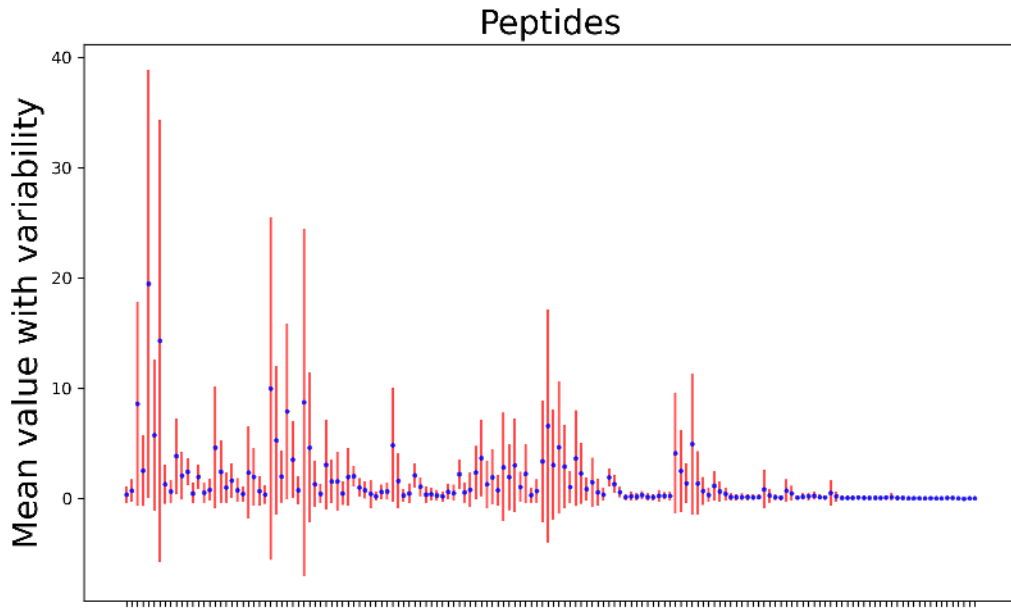


Figure 4: Peptides, Mean value with variability

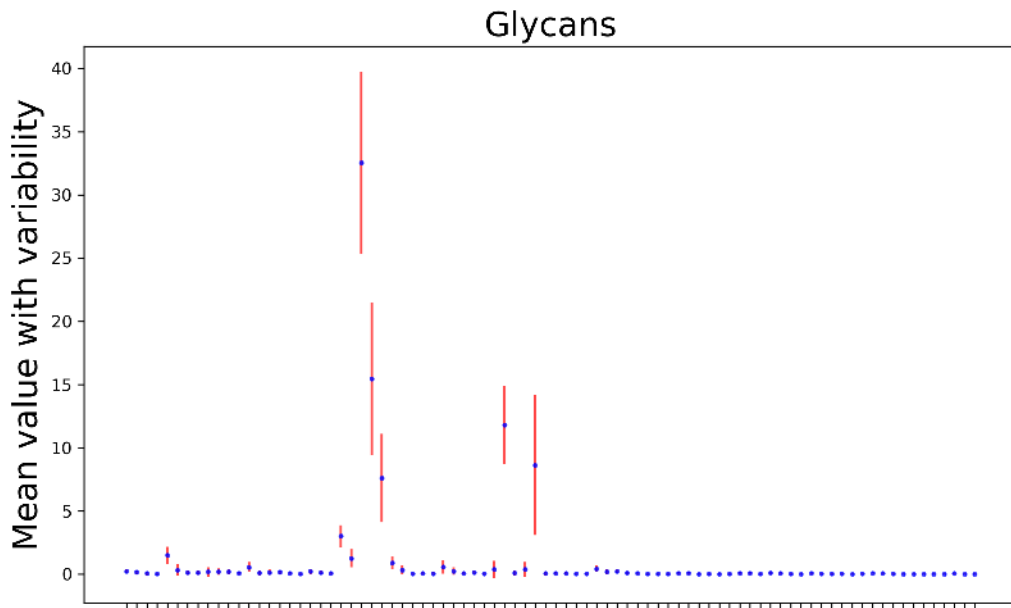


Figure 5: Glycans, Mean value with variability

3.2 Clustering Methods

Clustering, a primary form of unsupervised learning, has been a pivotal technique in MSI data analysis. Its ability to classify data into meaningful groups or clusters based

on inherent similarities is very useful. This section delves into the specifics of clustering methodologies, discussing both traditional and the proposed alternative methods.

3.2.1 Traditional Clustering Methods

In the domain of data clustering, two major techniques stand out: K-means clustering and hierarchical clustering. These methods are foundational for segmenting datasets into subgroups or clusters, enabling the identification of patterns and categorizing data points based on similarities. The default metric for measuring similarity in both algorithms is the Euclidean distance.

Notation Definitions Let's define the notation used in the subsequent sections:

- A and B : These represent two distinct data points in our dataset, for which we will be computing the distance.
- a_h and b_h : These symbols represent the h^{th} feature (or dimension) of data points A and B , respectively.
- p : This symbolizes the total number of features or dimensions in our dataset. Thus, A and B are located in a p -dimensional space.
- $d(A, B)$: Represents the computed distance between data points A and B .
- μ_k : The centroid or mean of k -th cluster.
- C : A cluster, which is a set of vector of p dimensions.
- C_j : Represents the j^{th} cluster in the dataset.
- $|C_j|$: The number of data points (or the cardinality) in cluster C_j .

K-means Clustering K-means clustering is an iterative, centroid-based technique that partitions data into a predefined number of clusters, denoted by k . [10]

Algorithm and Objective The main objective of K-means is to minimize the within-cluster sum-of-squares (WCSS), represented mathematically as:

$$WCSS = \sum_{k=1}^K \sum_{\mathbf{A} \in C_k} \|\mathbf{A} - \mu_k\|^2$$

The K-means procedure can be broken down into the following steps:

1. **Initialization:** k data points are selected randomly as the initial centroids. Advanced initialization techniques, such as k-means++, can provide better starting points for smaller datasets. [10]
2. **Cluster Assignment:** Each data point is assigned to the nearest centroid's cluster.
3. **Centroid Recalculation:** The centroid of each cluster is recalculated as the mean of its data points.

4. **Convergence Evaluation:** The assignment and centroid recalculation processes are iterated until the centroids stabilize or a maximum number of iterations is achieved.

Practical Considerations K-means clustering is not without its challenges:

- **Choice of k :** The optimal number of clusters isn't always apparent. Approaches like the Elbow Method or Silhouette Analysis can be instrumental in estimating k . These methods evaluate the model's performance across different k values, aiding in determining a value that best captures the data's structure.
- **Initialization Sensitivity:** The algorithm's outcome can vary depending on the initial centroids. Thus, employing methods like k-means++ or running the algorithm multiple times with different initializations and choosing the best outcome can mitigate this issue.
- **Outliers:** Outliers can distort the position of centroids. One remedy is using the k-medoids algorithm, which uses actual data points as centroids, making it less susceptible to outliers. Alternatively, preprocessing the data to remove or mitigate outliers can help.

Hierarchical Clustering Hierarchical clustering creates a tree of clusters without requiring the specification of the number of clusters upfront. This tree, visualized through a dendrogram, can offer a detailed view of how data points relate to each other.[13]

Algorithmic Overview Two primary strategies exist for hierarchical clustering:

1. **Agglomerative (bottom-up):** Starts with each observation as its own cluster and merges them based on a distance metric until one single cluster remains.
2. **Divisive (top-down):** Begins with all observations in one cluster and partitions them recursively.

Linkage Criteria and Formulations The core of hierarchical clustering is linkage - the method to compute the distance between clusters.

Single Linkage:

$$d_{\text{single}}(C_k, C_j) = \min_{A \in C_k, B \in C_j} d(A, B)$$

Complete Linkage:

$$d_{\text{complete}}(C_k, C_j) = \max_{a \in C_k, b \in C_j} d(a, b)$$

Average Linkage:

$$d_{\text{average}}(C_k, C_j) = \frac{1}{|C_k||C_j|} \sum_{A \in C_k} \sum_{B \in C_j} d(A, B)$$

Centroid Linkage:

$$d_{\text{centroid}}(C_k, C_j) = d(\mu_{C_k}, \mu_{C_j})$$

Challenges and Considerations While hierarchical clustering offers depth and a detailed view of relationships, it can be computationally demanding for large datasets. Choosing where to "cut" the dendrogram, i.e., deciding the number of clusters, is a nuanced task. It might require domain expertise or additional analytical methods.

Conclusion Both K-means and hierarchical clustering serve as robust methodologies to unveil latent structures in mass spectrometry imaging data. Applying them effectively requires not just an understanding of their mathematical workings but also a comprehensive appreciation of the data's characteristics and the overarching research objectives.

3.2.2 Distance Metrics in Clustering

The efficacy of clustering algorithms, notably k-means and hierarchical clustering, hinges on the nuances of distance computation. By default, these algorithms leverage the Euclidean distance unless specified otherwise. The choice of distance metric plays a crucial role, particularly in the intricate landscapes of datasets like mass spectrometry imaging. This section delves deep into the various distance metrics to underscore their significance and applications.

Euclidean Distance Euclidean distance traces its roots back to Pythagoras' theorem, providing an intuitive geometric method for measuring distance. It determines the "straight line" distance between two points in space, making it highly suitable for spatial datasets.

$$d(A, B) = \sqrt{\sum_{h=1}^p (b_h - a_h)^2}$$

In many applications, especially in machine learning and clustering, Euclidean distance is the default choice due to its natural geometric interpretation. However, it's sensitive to magnitudes and might not always be ideal when the data have different scales or units.[15]

Manhattan Distance Also known as the "Taxicab" or "City Block" distance, the Manhattan distance is reminiscent of a city's grid layout. Instead of the direct path as in Euclidean distance, it calculates the distance a taxi would take moving along the grid.

$$d(A, B) = \sum_{i=1}^p |b_i - a_i|$$

Manhattan distance is often used in grid-based pathfinding algorithms. It's less sensitive to outlier coordinates than the Euclidean distance, which can be advantageous in noisy datasets.[17]

Chebyshev Distance Originating from chessboard distances, the Chebyshev distance determines the maximum absolute difference across all dimensions.

$$d(A, B) = \max_i |b_i - a_i|$$

Because it focuses on the maximum difference, Chebyshev distance can be used in scenarios where the worst-case deviation is of primary importance. It effectively bounds other distance measures.[5]

Bray-Curtis Distance The Bray-Curtis distance assesses the compositional dissimilarity between two entities.

$$d(A, B) = \frac{\sum_{i=1}^p |b_i - a_i|}{\sum_{i=1}^p (b_i + a_i)}$$

This metric is often used in ecology to quantify the compositional difference between different sites based on species occurrences. It ranges between 0 (complete similarity) and 1 (complete dissimilarity).[19]

Cosine Distance The cosine distance focuses on the angle between two vectors, making it especially useful for high-dimensional, sparse datasets.

$$d(A, B) = 1 - \frac{\sum_{i=1}^p a_i \times b_i}{\sqrt{\sum_{i=1}^p a_i^2 \times \sum_{i=1}^p b_i^2}}$$

Commonly used in text analysis, the cosine distance effectively deals with situations where the magnitude of vectors is less relevant than their orientation.[1]

Correlation - Pearson Distance The Pearson distance, derived from Pearson correlation, gauges the linear relationship between two datasets.

$$d(A, B) = 1 - \frac{\sum_{i=1}^p (a_i - \bar{A}) \times (b_i - \bar{B})}{\sqrt{\sum_{i=1}^p (a_i - \bar{A})^2 \times \sum_{i=1}^p (b_i - \bar{B})^2}}$$

Where $\bar{A}$ and $\bar{B}$ are the average values of data points A and B , respectively. The Pearson distance is particularly effective when considering the linear relationship between variables, like in time series analysis.[9]

Canberra Distance The Canberra distance weighs the absolute differences between points by their sum.

$$d(A, B) = \sum_{i=1}^p \frac{|a_i - b_i|}{|a_i| + |b_i|}$$

It is sensitive to points near the origin, making it especially relevant in contexts where small absolute differences at low magnitudes are significant.[14]

Conclusion As we traverse the vast and complex domain of mass spectrometry imaging data, the choice of distance metric becomes a lynchpin. It's not merely a mathematical consideration but a strategic decision that can illuminate hidden patterns, correlations, and clusters. A judicious blend of metric understanding, domain knowledge, and empirical validation is pivotal for clustering success.

3.3 Integration of Spatial Information into Clustering

The integration of spatial information in clustering techniques has been highlighted as a significant development in the realm of spatial transcriptomics. A notable work that underscores this development is presented by Avesani et al. [4]. This study introduced Stardust, a clustering method that effectively leverages the amalgamation of spatial and transcriptomic information in the clustering process. A key challenge addressed by Stardust is discerning the extent to which spatial information should influence clustering across diverse tissues.

While the approach proposed offers promising insights, this investigation seeks to delve deeper into the realm of spatial information integration in clustering. Specifically, the emphasis lies on examining how spatial penalization on clusters can be juxtaposed with known distance methodologies.

In this section, the specific approach to the integration of spatial information will be elucidated, delineating how it differentiates and positions itself relative to existing methodologies like Stardust.[4]

3.3.1 Need for Spatial Penalization

In the domain of mass spectrometry imaging (MSI), data is not merely a collection of spectral intensities; it inherently embodies spatial information that is located in a specific pixel of the image. This spatial dimension adds a layer of complexity but also offers an opportunity to enhance the accuracy and relevance of clustering results.

The traditional clustering methods, although powerful, often operate under the assumption that data points are independent entities. This assumption may lead to inaccuracies when applied to MSI data, where the spatial proximity of data points often corresponds to their molecular similarity. For instance, neighboring pixels on a tissue slide might originate from the same type of cell and hence, should ideally be clustered together.

Without considering the spacial proximity of two pixels, clustering algorithms might overlook these subtleties, leading to clusters that might be mathematically optimal but not biologically or contextually relevant. The challenge is analogous to segmenting an image based solely on pixel values without considering the pixels' positions. Just as in image processing, where spatial filters and morphological operations can enhance segmentation results, in MSI data analysis, spatial penalization can guide clustering algorithms to produce more biologically relevant clusters.

Spatial penalization does not imply an overshadowing of the original data's spectral properties. Instead, it acts as a balancing factor, ensuring that the clustering does not stray too far from the spatial realities of the sample. The degree of penalization can be tuned, allowing a balance between spatial and spectral clustering influences.

Furthermore, spatial penalization can enhance the interpretability of clustering results. Clusters shaped by both spectral and spatial properties can be more easily related to physical structures or regions within the sample, aiding researchers in drawing more concrete conclusions from their data.

While traditional distance metrics provide a foundational approach to clustering MSI data, the integration of spatial penalization is imperative to capture the holistic essence of the data and derive meaningful, actionable insights.

3.3.2 Spatial Penalization Method and Spatial-Enhanced Clustering Algorithms

Mass spectrometry imaging (MSI) data inherently possesses spatial attributes. Conventional clustering algorithms primarily focus on the feature-space, making it essential to consider the spatial relationships between data points. Ignoring this spatial context might lead to loss of valuable insights. This section introduces a method designed to enhance traditional clustering techniques with spatial considerations.

Spatial penalization integrates spatial proximity as an additional factor to influence clustering outcomes. This method adjusts clustering algorithms to be more attuned to the inherent spatial structures in MSI data.

In traditional clustering, distances are primarily calculated based on feature values. However, with spatial penalization, these distances are modified using the spatial relationships between data points. The principle is straightforward: regions in close spatial proximity probably have shared characteristics, and this closeness should influence clustering decisions.

Formula for Spatial Penalization For spatial penalization, we first define the spatial distance using the Euclidean norm, which we'll denote as $d_E(A, B)$, representing the spatial Euclidean distance between two coordinates A and B . The spatial penalization factor is then derived from the following formula:

$$\text{PenalizationFactor}(A, B) = 1 - e^{-d_E(A, B)}$$

In the figure 6 a representation of the penalty function:

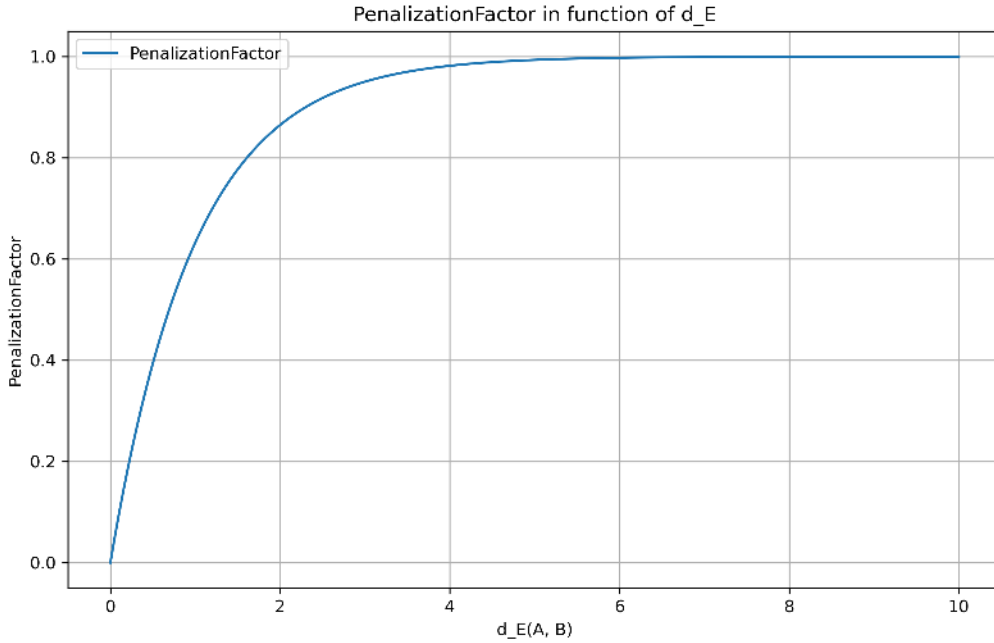


Figure 6: Representation of the Penalization Factor function curve

The mentioned function converts raw spatial distances into a scale between 0 and 1. It ensures that the penalization is neither too lenient nor too stringent, respecting the inherent data distribution.

Integrating Spatial Penalization into Clustering Algorithms Spatial penalization can enhance both k-means and hierarchical clustering:

1. **Distance Computation:** Calculate the primary distance (Euclidean, Manhattan, etc.) between data points based on their features.
2. **Spatial Penalization:** Compute and transform the spatial distance between data points using the spatial penalization formula.
3. **Adjusted Distance:** The primary distance is modified using the penalization factor, represented as:

$$\text{AdjustedDistance}(A, B) = d(A, B) \times \text{PenalizationFactor}(A, B)$$

With these adjustments, both the intrinsic feature similarities and spatial proximities significantly shape the clustering outcome.

3.3.3 Spatial Penalization Effects

Incorporating spatial penalization imparts unique characteristics to the way points are clustered, especially concerning their mutual proximities:

- **Mutually Close Points:** For two points that are mutually close in the spatial domain, their computed spatial distance will be small. When passed through the spatial penalization formula, the resulting SpatialDistance will be closer to 0. Therefore, mutually close points are more likely to be clustered together, maintaining their spatial relationships.
- **Distant Points:** Conversely, for two points situated farther apart, their spatial distance will be relatively large. The penalization formula will thus yield a value that's closer to 1, significantly influencing the original feature-based distance. This adjustment, in essence, acts as a penalty, making it less likely for distant points to be clustered together unless their feature similarities are overwhelmingly strong.

Benefits and Considerations The integration of spatial penalization in clustering algorithms offers significant advantages, particularly when dealing with Mass Spectrometry Imaging (MSI) data. One of the standout benefits of this approach is the assurance of enhanced spatial coherence. By giving due importance to spatial attributes, the resulting clusters adhere more closely to the inherent spatial structures present within the data. Such adherence often translates into clusters that are both more interpretable and carry profound meaning in the context of MSI data.

Moreover, the spatial penalization method is marked by its adaptability. Its design allows for a seamless integration into various clustering algorithms, underlining its versatility in data analysis. However, this approach is not without its intricacies. There's a noted sensitivity to parameters, indicating that the strength of the spatial penalization

might need adjustments based on the specific dataset and the corresponding research questions.

Spatial attributes are critical in MSI data, and neglecting them can lead to a loss of meaningful information. By integrating spatial penalization into clustering algorithms, we ensure that the inherent spatial structures of the data are respected, paving the way for more insightful analyses.

3.4 Evaluation of the clustering performance

3.4.1 Gold Standard for Comparison

To evaluate the performance of the spatial penalization technique for clustering in mass spectrometry imaging data we need a rigorous gold standard. In this context, Bayesian clustering methods, namely the Gaussian Mixture Model (GMM) and the Hierarchical Dirichlet Process Mixture (HPM), have been shown to be effective for the analysis of mass spectrometry imaging data, and are often used as benchmarks for evaluating the performance of new methods.[16] These methods offer a probabilistic framework for data clustering and are recognized for their adaptability and robustness.[2] Both GMM and HPM have been validated by biologists in numerous applications involving complex biological datasets. Their credibility stems from the consistent and accurate clustering results they produce, which align well with known biological phenomena and experimental findings.

Gaussian Mixture Model (GMM) The Gaussian Mixture Model, often abbreviated as GMM, is a probabilistic model that assumes all data points are generated from a mixture of several Gaussian distributions with unknown parameters. In essence, each cluster adheres to a Gaussian distribution.

1. **Principle:** Instead of assigning a data point to a definitive cluster, GMM determines the probability that the point belongs to a given cluster. This soft-clustering method proves valuable in scenarios with ambiguous cluster boundaries.
2. **Expectation-Maximization:** The GMM employs the Expectation-Maximization (EM) algorithm to determine the optimal parameters for the Gaussian distributions. The EM algorithm iteratively adjusts cluster assignments and Gaussian parameters until convergence is achieved.

The key advantage of GMM lies in its capacity for soft clustering, offering nuanced categorization that is especially beneficial when cluster boundaries are not distinct. However, a limitation of GMM is its assumption that clusters are Gaussian, which may not always hold true for every dataset. Thus, while offering flexibility, GMM might struggle with highly non-linear or non-Gaussian data. For comparative purposes, the standard GMM serves as the non-spatial Bayesian clustering benchmark.[18]

Hierarchical Dirichlet Process Mixture (HPM) The Hierarchical Dirichlet Process Mixture, or HPM, enhances the Bayesian non-parametric clustering methodology to include hierarchical and spatial structures.

1. **Principle:** Unlike the GMM which assumes a set number of clusters, HPM can deduce the number of clusters directly from the data, making it apt for modeling complex data distributions and hierarchies.
2. **Spatial Analysis:** A spatially-enhanced version of HPM is considered for gold standard comparison. By incorporating spatial data directly into the clustering process, this model facilitates a comprehensive analysis of mass spectrometry imaging data. The combination of spatial proximity and feature similarity ensures a robust clustering outcome.

The HPM method stands out for its adaptability to the inherent structure of the data, having the capability to infer the optimal number of clusters. This is especially valuable when the number of clusters is not known a priori. Additionally, its ability to integrate spatial information offers a more holistic view of mass spectrometry imaging data. However, HPM is more computationally intensive than GMM, and its performance can be sensitive to the specification of hyperparameters.[24]

Conclusion The selection of Bayesian methods, GMM and HPM, as gold standards provides a reliable foundation for comparison. While the GMM offers a foundational, non-spatial comparison point, the spatially-aware HPM is essential to understand the benefits of including spatial information in the clustering process. Such a comparison facilitates discerning insights regarding the advantages and potential refinements required for the spatial penalization technique.

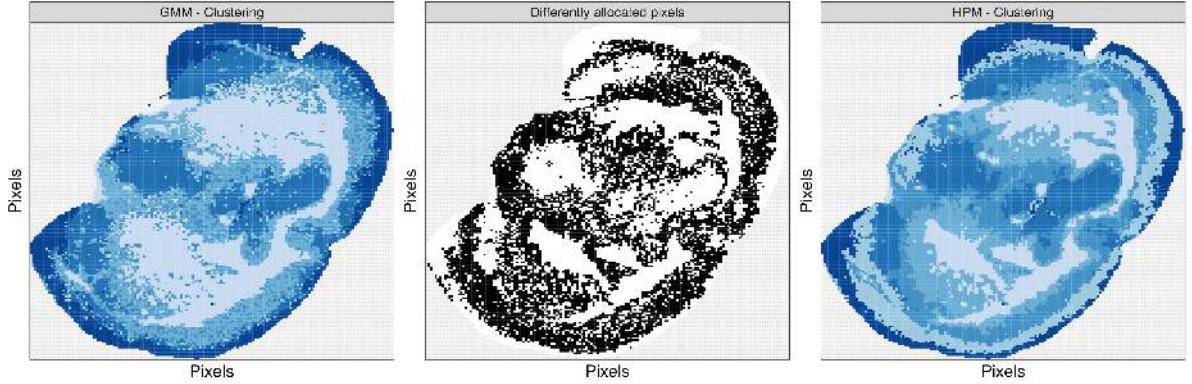


Figure 7: GMM and HPM clustering results, with a comparison of differently allocated pixels

3.4.2 Metrics for evaluation as compared with gold standard

Three primary metrics were chosen to assess the quality of clustering:

- a) Adjusted Rand Index (ARI) [21]: The Rand Index (RI) measures the similarity between two data clusterings. It is given by:

$$RI = \frac{a + b}{a + b + c + d}$$

Where:

- a is the number of pairs of objects that are in the same group in both the true and predicted clusterings.
- b is the number of pairs of objects that are in different groups in both clusterings.
- c is the number of pairs of objects that are in the same group in the true clustering but in different groups in the predicted clustering.
- d is the number of pairs of objects that are in different groups in the true clustering but in the same group in the predicted clustering.

However, RI does not guarantee that random cluster assignments will yield a value close to zero.

The Adjusted Rand Index introduces a correction factor:

$$ARI = \frac{RI - \text{Expected_RI}}{\max(RI) - \text{Expected_RI}}$$

Where **Expected_RI** is the expected RI value for random cluster assignments.

b) Normalized Mutual Information (NMI) [22]:

Mutual Information (MI) between two variables Y and C is given by:

$$MI(Y, C) = \sum_{y \in Y} \sum_{c \in C} P(y, c) \log \frac{P(y, c)}{P(y)P(c)}$$

Where $P(y, c)$ is the joint probability distribution of Y and C , and $P(y)$ and $P(c)$ are the marginal probability distributions of Y and C respectively.

However, to account for the individual entropies of the variables, we use Normalized Mutual Information:

$$NMI = \frac{2 \times MI(Y, C)}{H(Y) + H(C)}$$

Where **H** denotes entropy.

c) Cluster Overlap:

Beyond the quantitative evaluations given by ARI and NMI, a third approach was employed based on the overlap between clusters. Clusters were visualized individually and overlapped with the reference gold standard (GMM for the unweighted methods and HPM for the weighted methods). A manual remapping was subsequently conducted in collaboration with domain experts to discern the correct and incorrect cluster correspondences beyond mere result overlaps, as is the case with ARI and NMI. This third approach produced more meaningful and consistent results in line with the visualizations obtained from cluster plots. Consequently, it will be considered to be more reliable.

4 Results

It should be noted that the clustering data used as the standard goal, specifically the GMM and HPM models, were set to generate 6 clusters. The choice of these six clusters was not arbitrary but was firmly rooted in both the spatial patterns observed in the data and the biological underpinnings associated with the region of interest. Explicitly, within the given data, discernible spatial patterns are evident. These patterns prominently highlight various regions such as the hippocampal formation (represented by light blue pixels), the dense white matter (characterized by high-intensity blue pixels), and a conglomerate region of gray matter paired with the cerebral cortex (visible as white pixels). Beyond these observable spatial patterns, from a biological vantage point, there's an anticipation to distinctly identify both the Thalamus and Hypothalamus as integral components within the overarching hippocampal formation. Drawing from this detailed biological information, we prudently determined the number of mixture components, denoted as K , to be set at 6. Hence, to ensure a consistent and meaningful comparison, all subsequent clustering endeavors were aligned to this configuration, working towards partitioning the data into six well-defined clusters.

4.1 Results for Traditional Clustering Methods

As part of this thesis, both k-means and hierarchical clustering techniques are explored using different distance metrics. Table 1 reports the performance of these methods in comparison to the gold-standard GMM, which does not take spatial information into account.

	GMM			
	ARI		NMI	
	k-means	hclust	k-means	hclust
Euclidean	0.267	0.054	0.395	0.245
Manhattan	0.306	0.127	0.455	0.340
Chebyshev	0.221	-0.009	0.32	0.068
Bray Curtis	0.525	0.336	0.529	0.369
Cosine	0.312	0.254	0.392	0.297
Cor - Pearson	0.308	0.226	0.39	0.324
Canberra	0.357	0.460	0.409	0.438

Table 1: Performance of the cluster methods with different distances as compared with the gold-standard without considering the spatial distance (GMM)

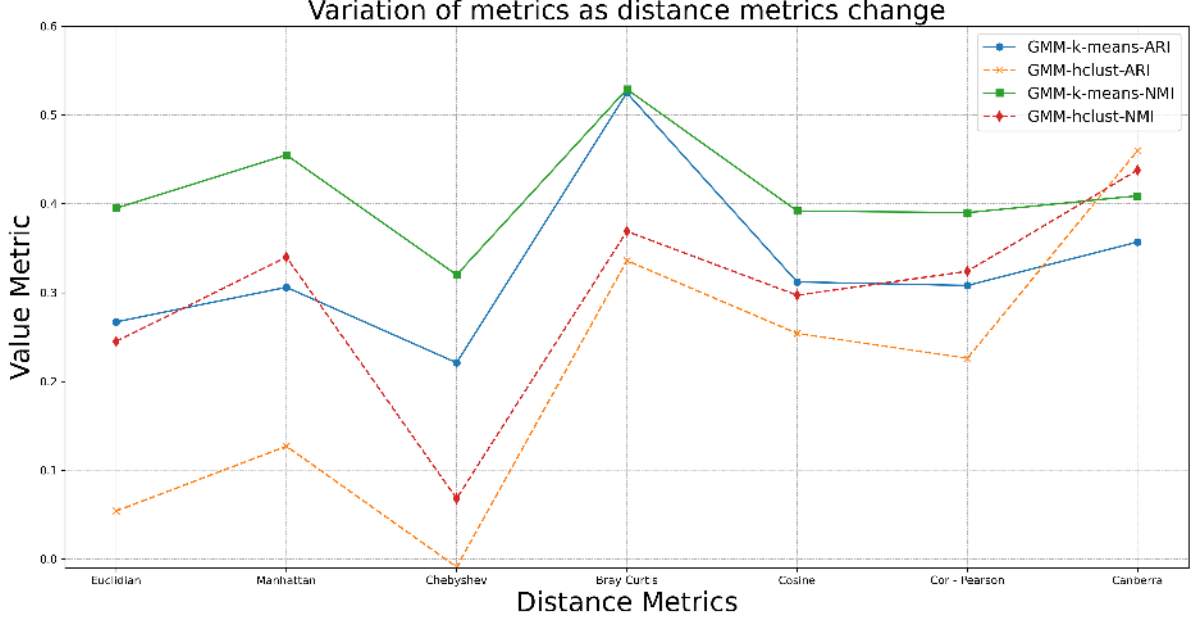


Figure 8: Results for Traditional Clustering Methods on Line Chart

Considering Table 1, with the non-spatial penalized GMM as a comparative benchmark, the results provide a clear insight into the performance of different distance metrics with both k-means and hierarchical clustering (hclust) methods.

From a cursory glance, the Bray Curtis metric, when compared with the GMM model, yields the highest ARI score with k-means at 0.525. Notably, the Euclidean distance metric, which is a standard and widely adopted metric, demonstrates a moderate performance in this setting (ARI of 0.267 with k-means). This suggests that while the Euclidean distance provides a baseline for comparison, it is not necessarily the most effective metric for every clustering task.

When focusing on the NMI score, Bray Curtis again leads the pack in the k-means method with a score of 0.529. This further reinforces the effectiveness of this metric in the context of mass spectrometry data.

The hierarchical clustering method (hclust) consistently scores lower in the ARI metric across all distance measurements in comparison to k-means, indicating that, for this specific dataset and the GMM model, k-means might be a closer clustering algorithm.

Following the third evaluation metric used, Cluster Overlap, Euclidean k-means (to evaluate the behavior of the standard method) and Bray Curtis k-means (being the best performing based on the previous metrics) are considered.

Figures 9, 10, 11 show the comparison between GMM clustering and Euclidean k-means clustering, in particular Figure 10 shows the pixels allocated differently in the two clustering types, the result is an overlap of 39.19%.

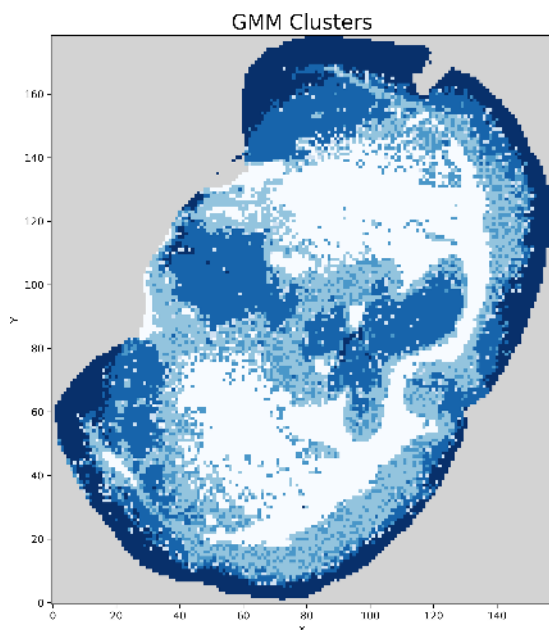


Figure 9: GMM clusters

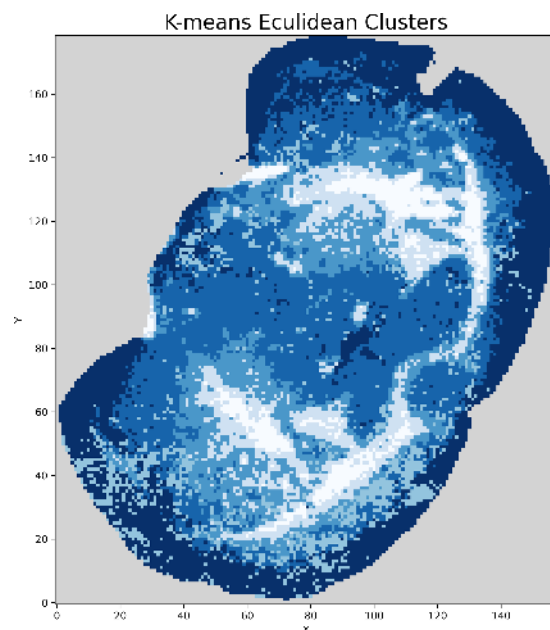
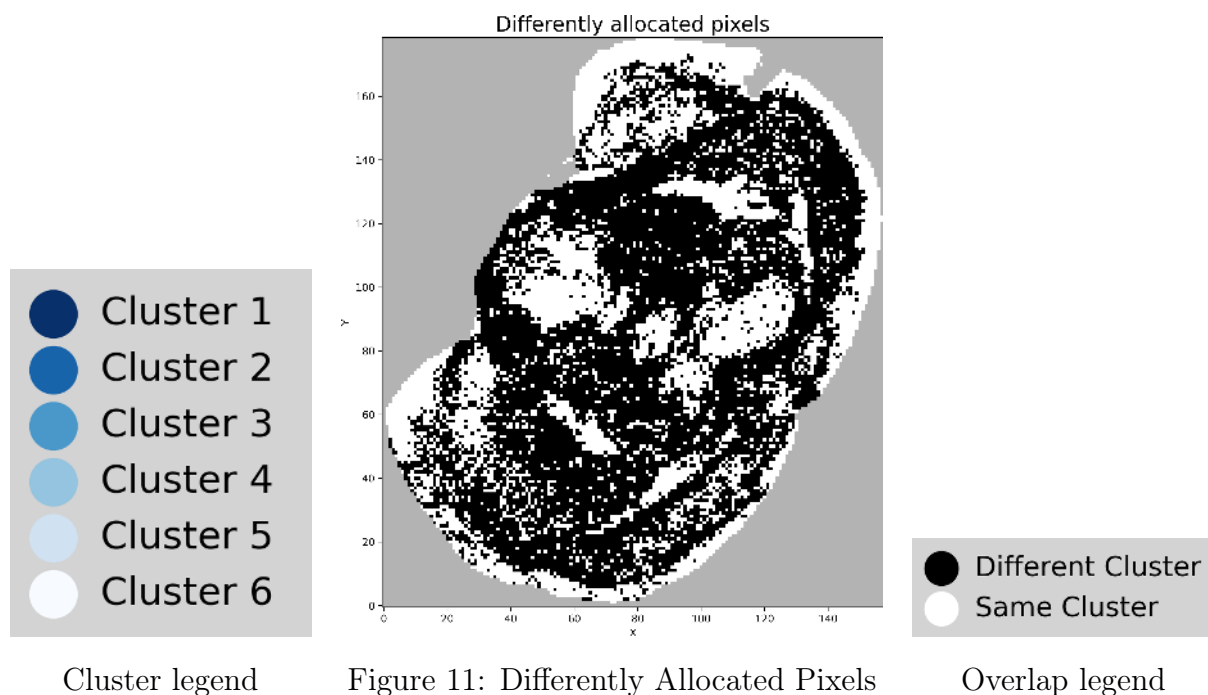


Figure 10: K-means Euclidean clusters



Cluster legend

Figure 11: Differently Allocated Pixels

Overlap legend

In the Figure 12 and 13, again for the K-means Euclidean clusters, their internal composition and coefficient of variation is shown, divided by peptides, lipids, and glycans so that we can visually confirm that the content of the clusters is consistent with the assignment made manually by the biologist. At the top of the figure you can also find the number of elements present for each cluster.

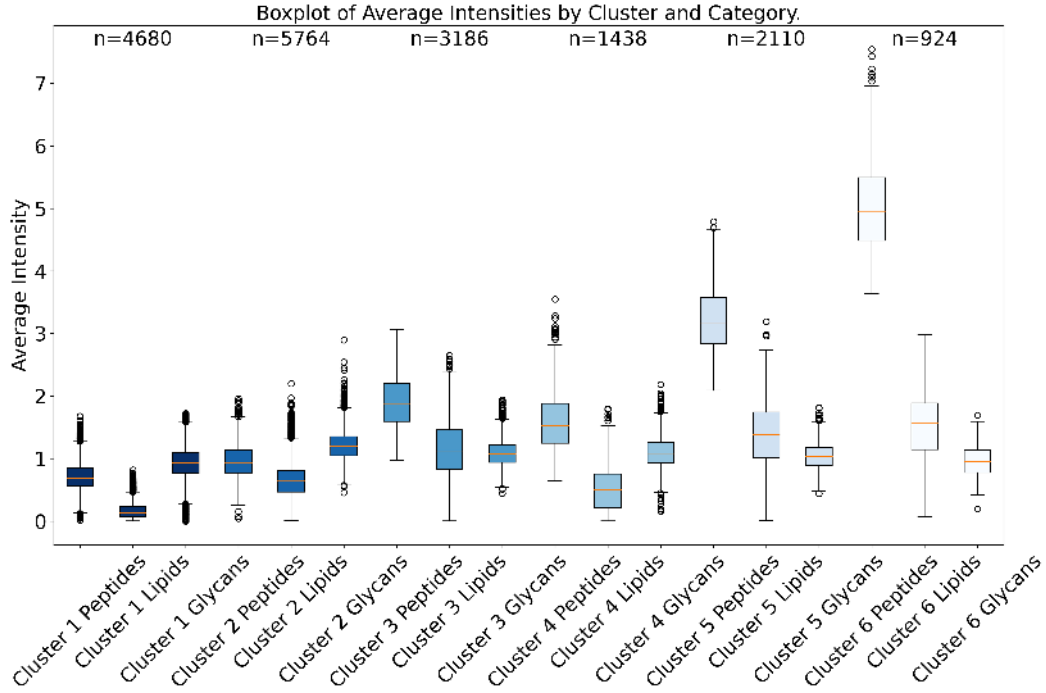


Figure 12: Boxplot of Intensities by Cluster and Category with K-means - Euclidean

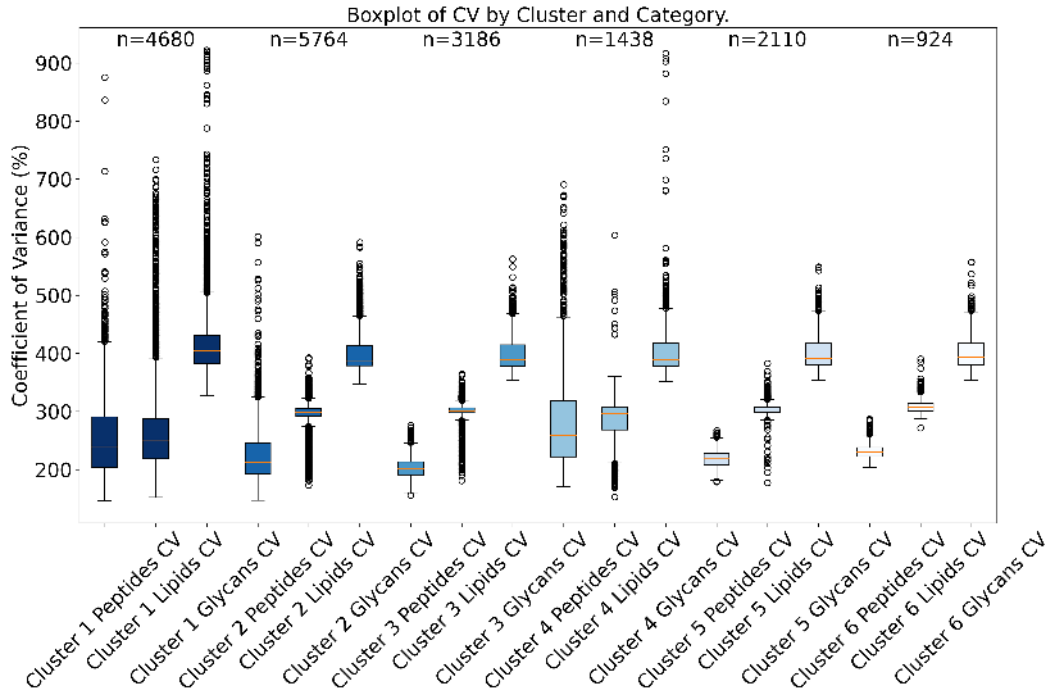


Figure 13: Boxplot of Coefficient Variation by Cluster and Category with K-means - Euclidean

Figures 14, 15, 16 show the comparison between GMM clustering and Bray Curtis k-means, in particular Figure 15 shows the pixels allocated differently in the two clustering types, the result is an overlap of 3%.

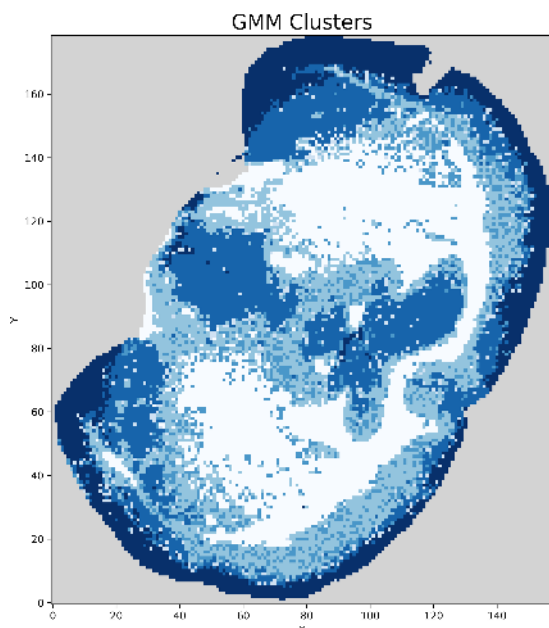


Figure 14: GMM clusters

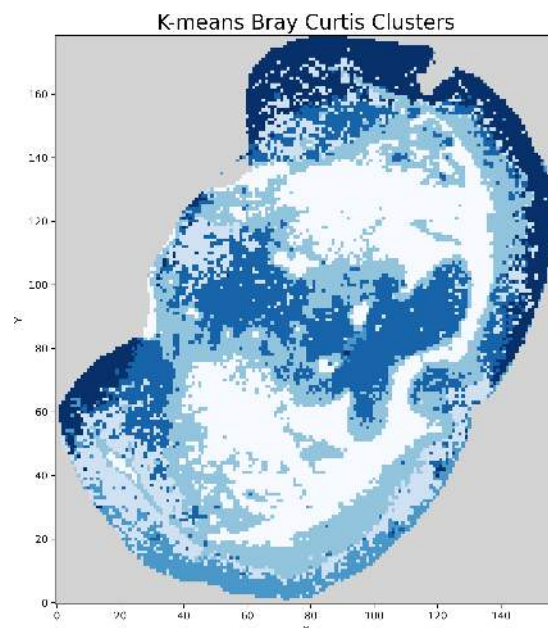
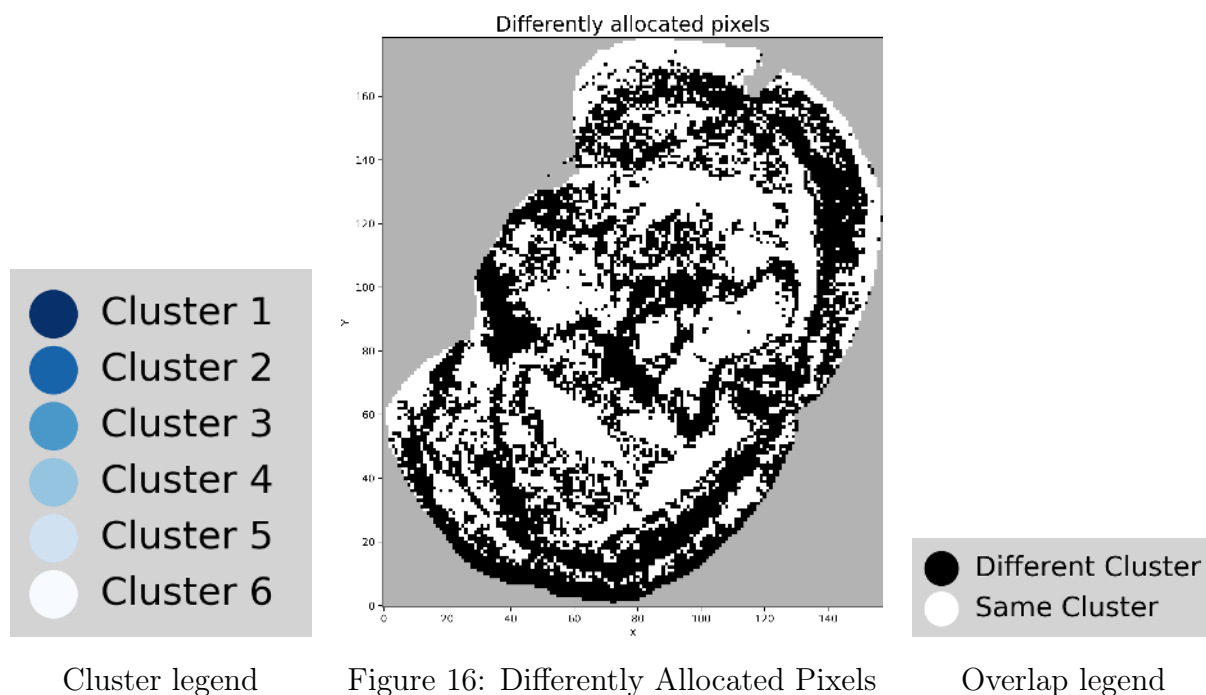


Figure 15: K-means Bray Curtis clusters



Cluster legend

Figure 16: Differently Allocated Pixels

Overlap legend

Below in Figure 17 and 18, as done previously, is shown the internal composition and coefficient of variation of the Bray Curtis k-means clusters, divided by peptides, lipids, and glycans so that we can visually confirm that the content of the clusters is consistent with the assignment made manually by the biologist. At the top of the figure you can also find the number of elements present for each cluster.

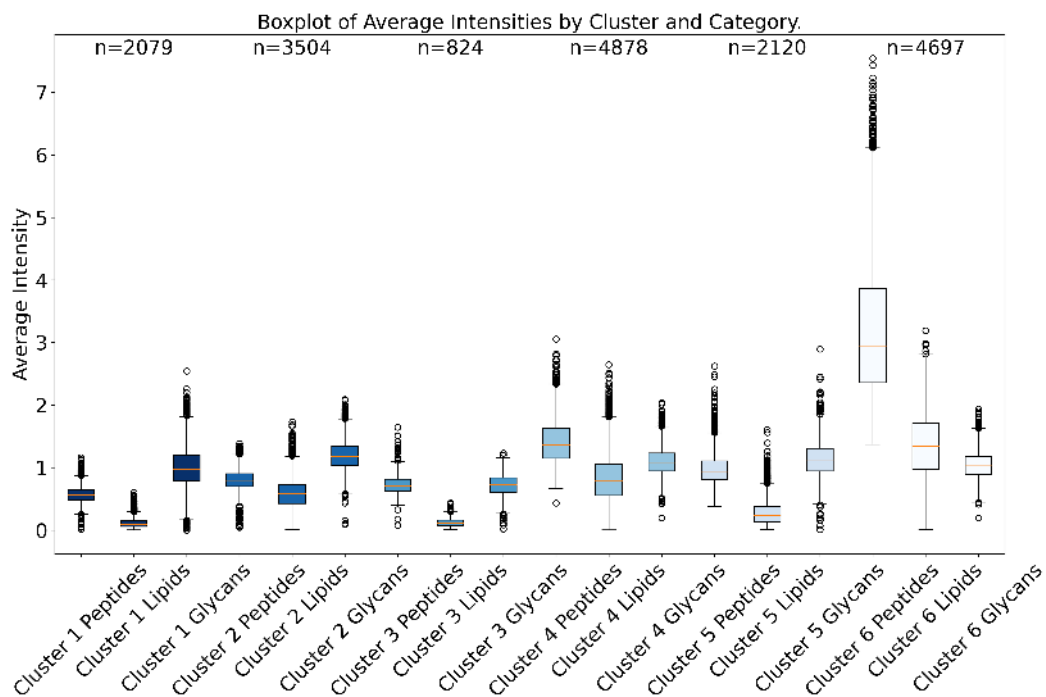


Figure 17: Boxplot of Intensities by Cluster and Category with K-means - Bray Curtis

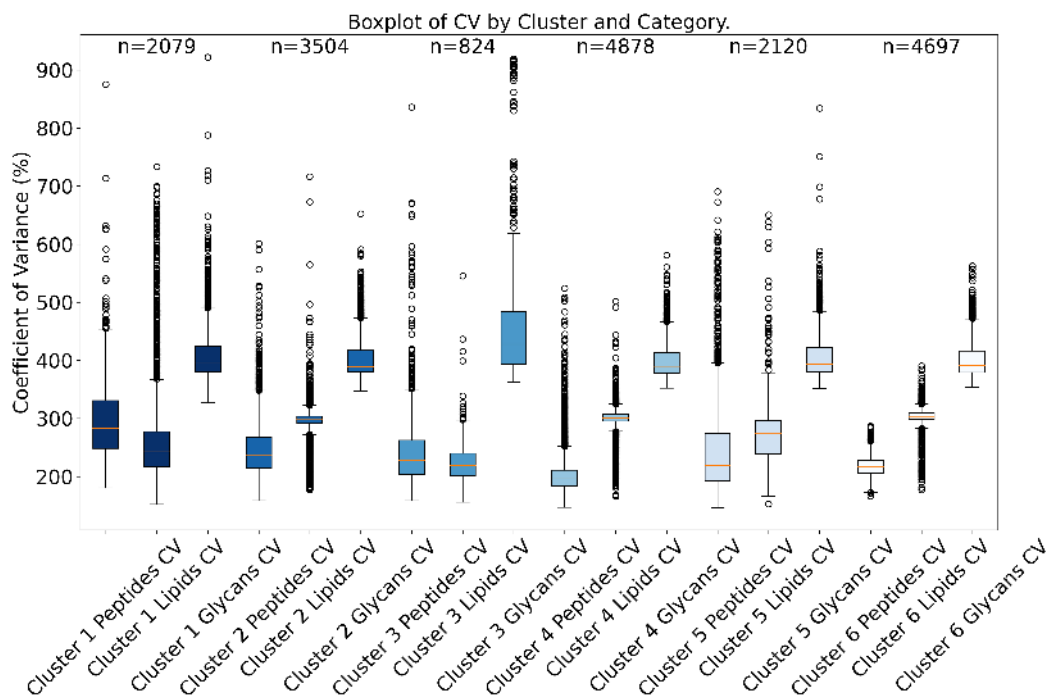


Figure 18: Boxplot of Coefficient Variation by Cluster and Category with K-means - Bray Curtis

4.2 Results for Spatial-Enhanced Clustering Algorithms

	HPM			
	ARI		NMI	
	k-means	hclust	k-means	hclust
Euclidean	0.434	0.156	0.551	0.369
Manhattan	0.396	0.227	0.572	0.436
Chebyshev	0.223	0.031	0.364	0.156
Bray Curtis	0.440	0.270	0.553	0.406
Cosine	0.335	0.238	0.477	0.348
Cor - Pearson	0.330	0.226	0.474	0.359
Canberra	0.375	0.391	0.489	0.511

Table 2: Performance of the cluster methods with different distances as compared with the gold-standard with penalization for spatial distance (HPM penalized)

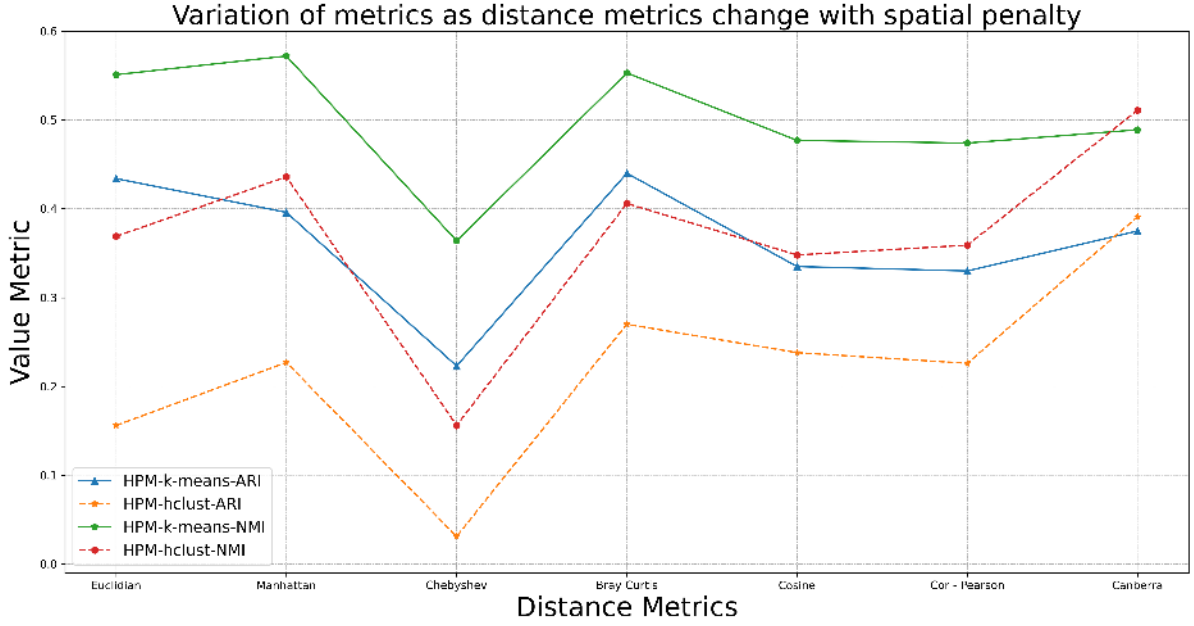


Figure 19: Results for Spatial-Enhance Clustering Methods on Line Chart

Transitioning to the results from the Table 2, which integrates spatial considerations akin to the HPM model, several observations become evident.

First, by taking a comparative look between the two tables, it's apparent that the inclusion of a spatial penalty has augmented the performance of the hclust method across most distance metrics. This enhancement is especially pronounced with the Euclidean metric, where the ARI score has risen from 0.054 to 0.164 and the NMI score from 0.245 to 0.324. Such an improvement underscores the value of incorporating spatial information, as it tends to provide a more holistic understanding and a refined clustering outcome.

The Bray Curtis metric, despite its high performance in the non-spatial penalized model, does not exhibit the same dominance in the spatially enhanced clustering scenario. While it still maintains commendable scores, other metrics like the Canberra distance, especially in the hclust method, demonstrate comparable or superior performance (ARI of 0.422 and NMI of 0.434).

Following the third evaluation metric used, Cluster Overlap, Euclidean k-means (to evaluate the behavior of the standard method), Manhattan k-means and Bray Curtis k-means (being both two best performers based on the previous metrics) are considered.

Figures 20, 21, 22 show the comparison between HPM clustering and Euclidean k-means clustering with spatial penalization, in particular Figure 21 shows the pixels allocated differently in the two clustering types, the result is an overlap of 34.03% (33.55% without the spatial penalization).

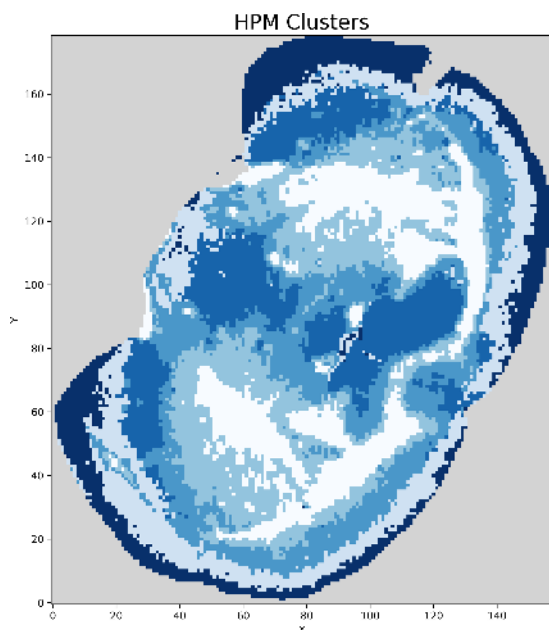


Figure 20: HPM clusters

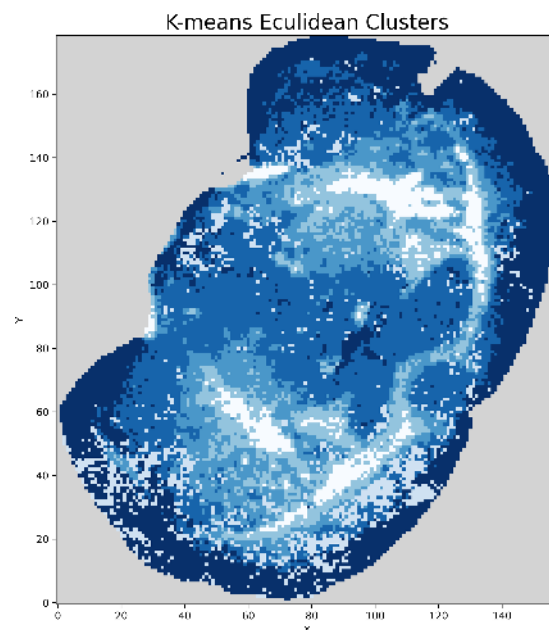
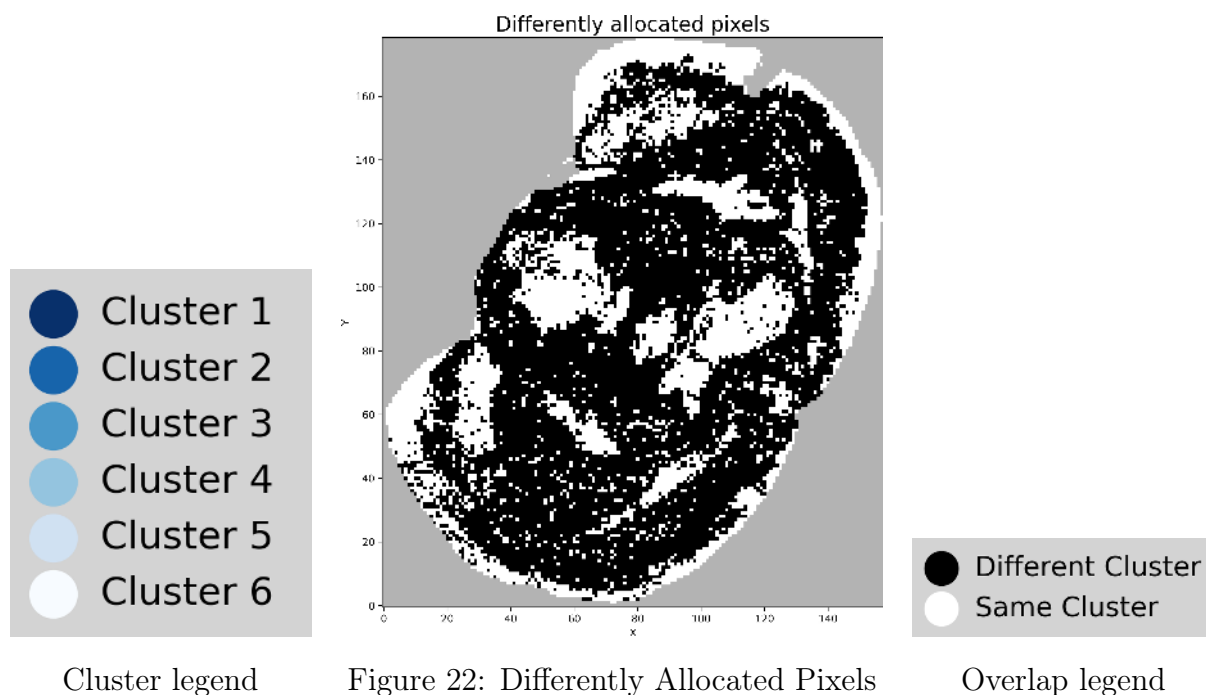


Figure 21: K-means Euclidean clusters



Cluster legend

Figure 22: Differently Allocated Pixels

Overlap legend

In the Figure 23 and 13, again for the K-means Euclidean clusters, their internal composition and coefficient of variation is shown, divided by peptides, lipids, and glycans so that we can visually confirm that the content of the clusters is consistent with the assignment made manually by the biologist. At the top of the figure you can also find the number of elements present for each cluster.

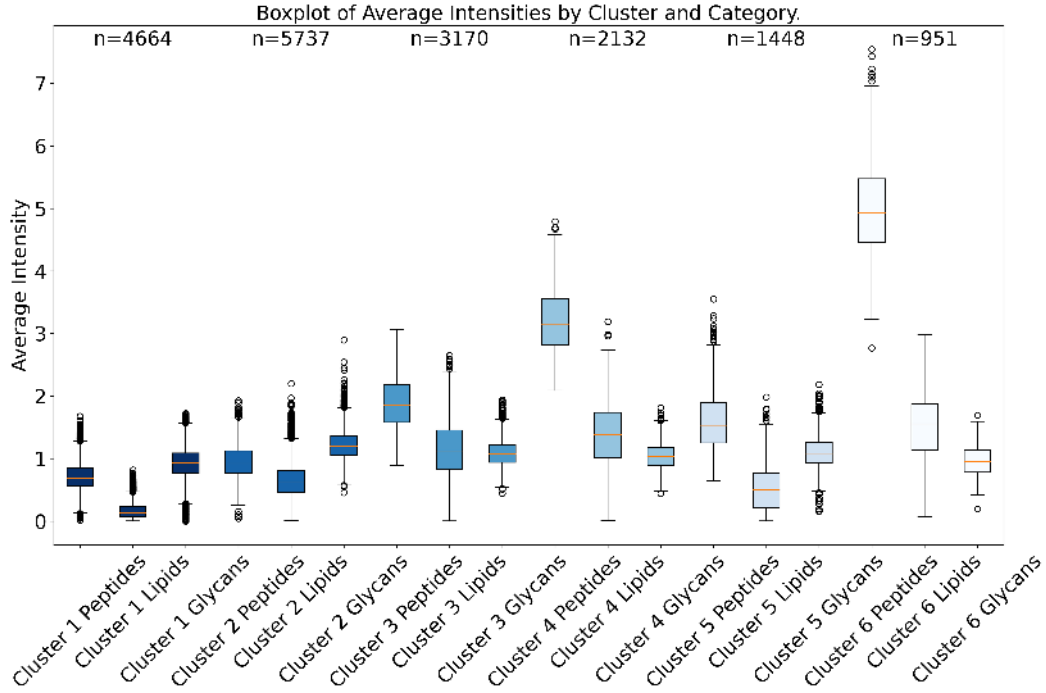


Figure 23: Boxplot of Intensities by Cluster and Category with K-means - Euclidean penalized

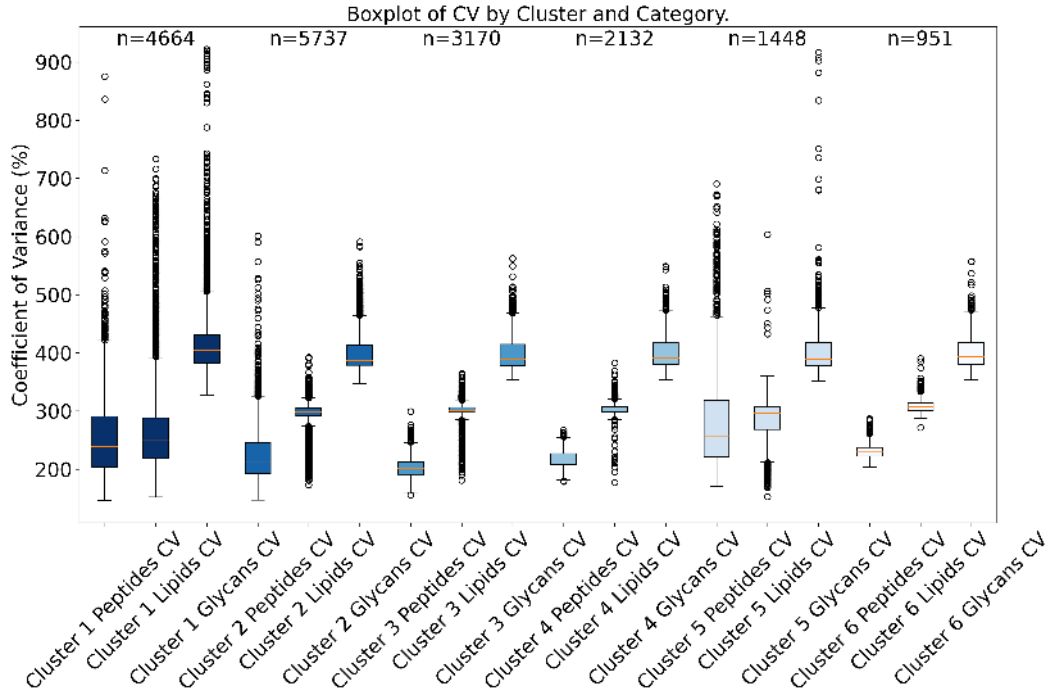


Figure 24: Boxplot of Coefficient Variation by Cluster and Category with K-means - Euclidean with Penalization

Figures 25, 26, 27 show the comparison between HPM clustering and Manhattan k-means clustering with spatial penalization, in particular Figure 26 shows the pixels allocated differently in the two clustering types, the result is an overlap of 44.03% (43.10% without the spatial penalization).

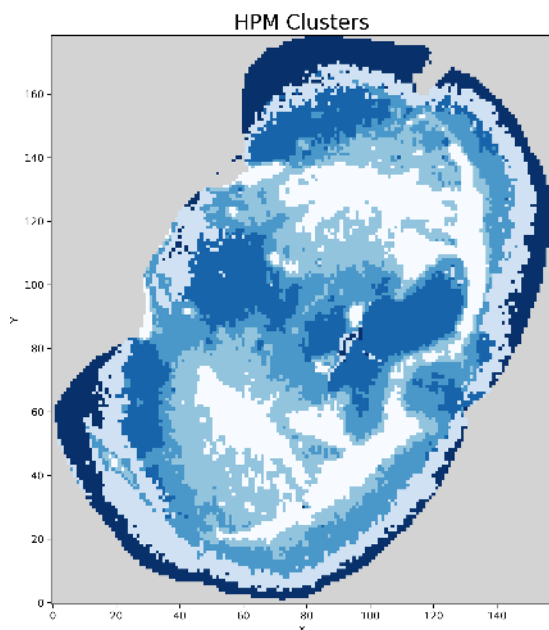


Figure 25: HPM clusters

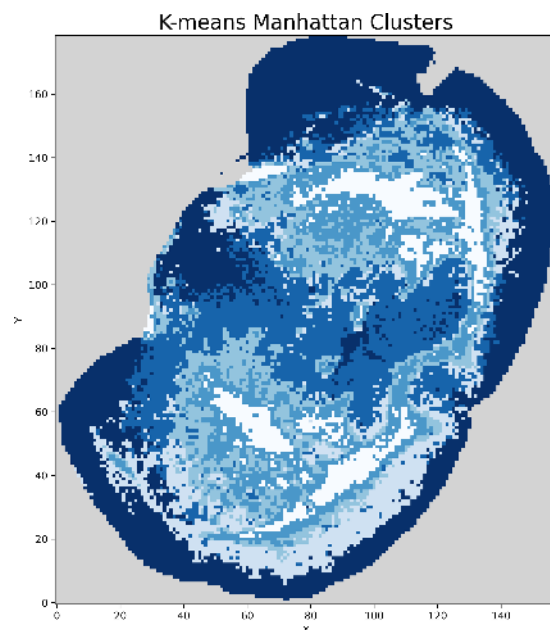
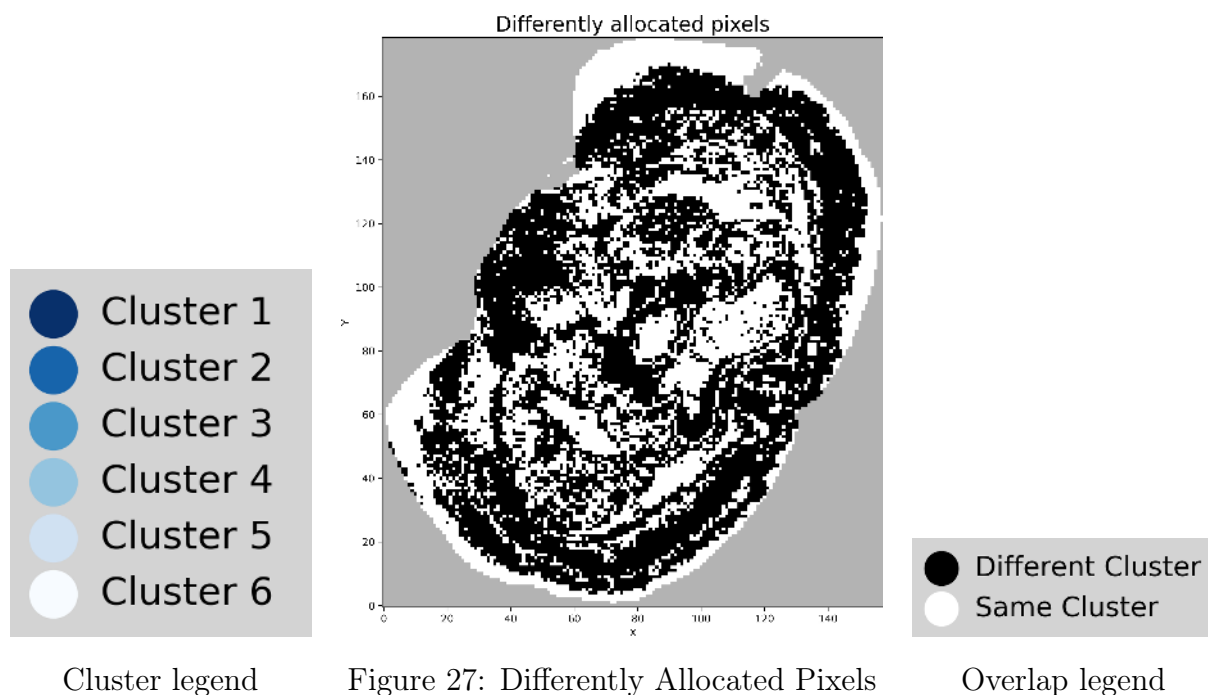


Figure 26: K-means Manhattan clusters



Cluster legend

Figure 27: Differently Allocated Pixels

Overlap legend

Below in Figure 28 and 29, as done previously, is shown the internal composition and coefficient of variation of the Manhattan k-means clusters, divided by peptides, lipids, and glycans so that we can visually confirm that the content of the clusters is consistent with the assignment made manually by the biologist. At the top of the figure you can also find the number of elements present for each cluster.

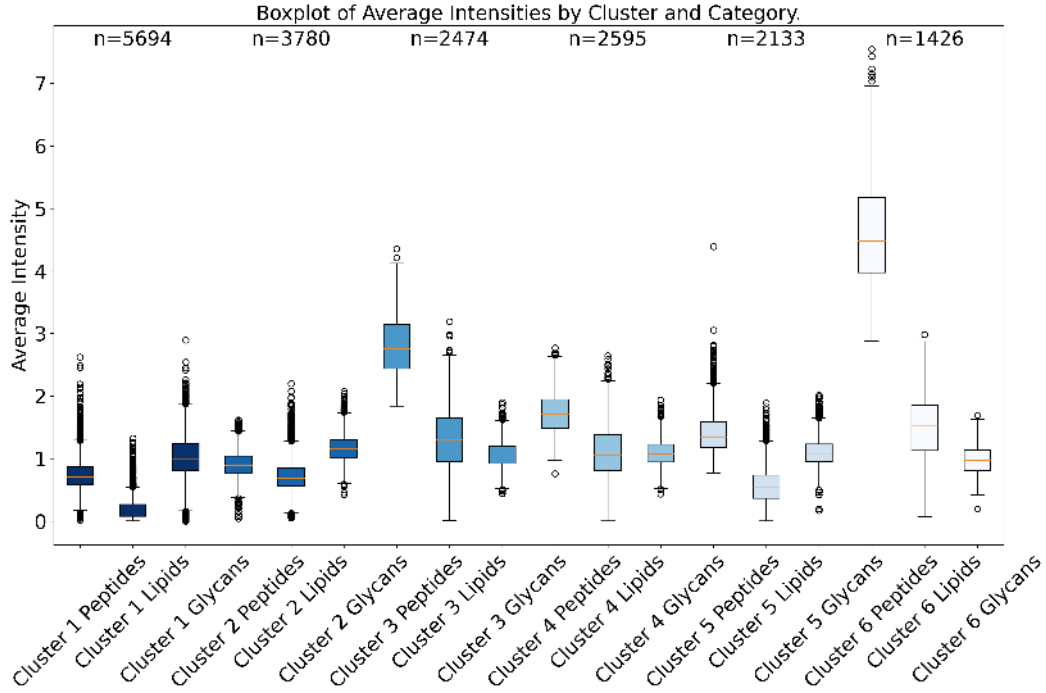


Figure 28: Boxplot of Intensities by Cluster and Category with K-means - Manhattan with Penalization

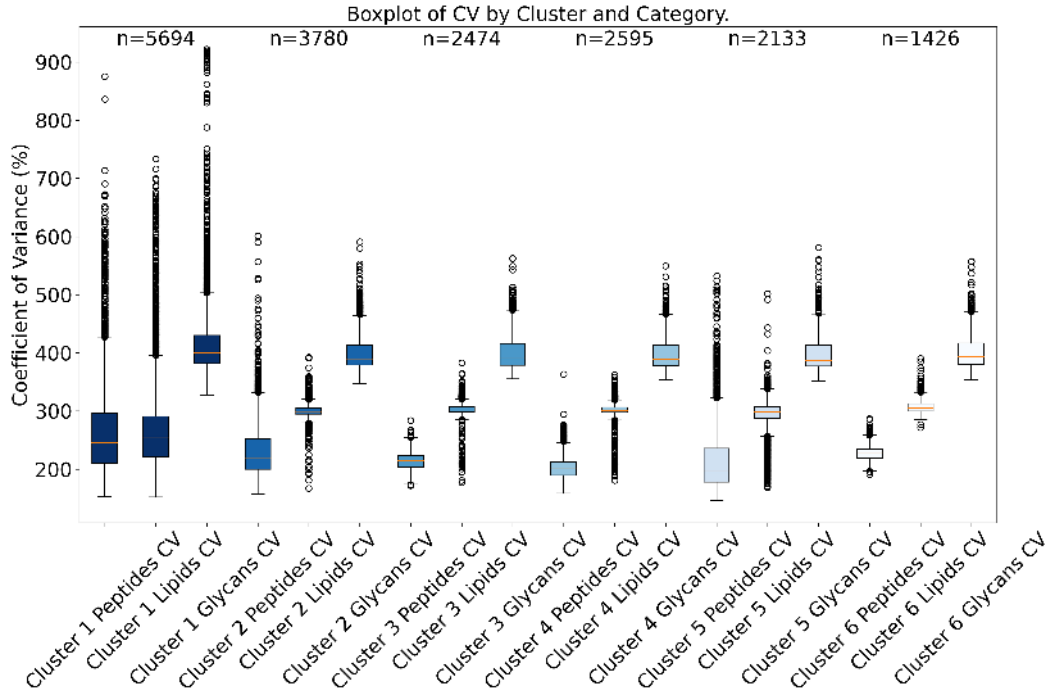


Figure 29: Boxplot of Coefficient Variation by Cluster and Category with K-means - Manhattan with Penalization

Figures 30, 31, 32 show the comparison between HPM clustering and Bray Curtis k-means clustering with spatial penalization, in particular Figure 31 shows the pixels allocated differently in the two clustering types, the result is an overlap of 58.22% (57.26% without the spatial penalization).

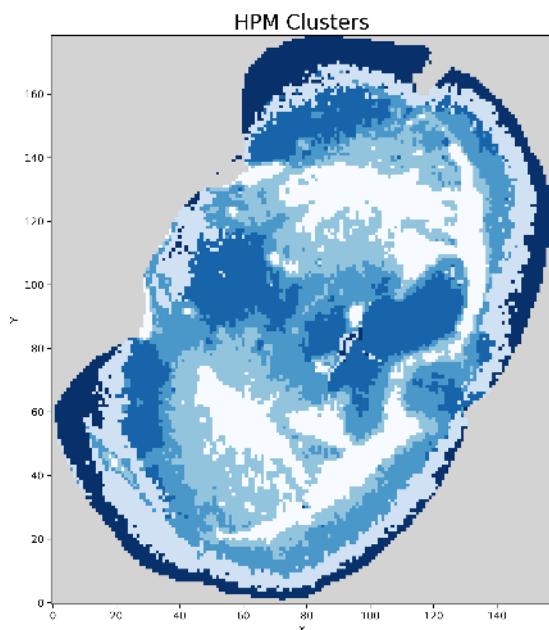


Figure 30: HPM clusters

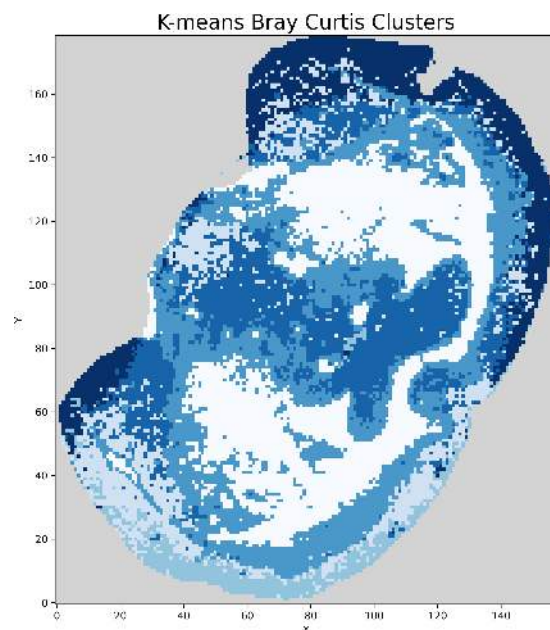
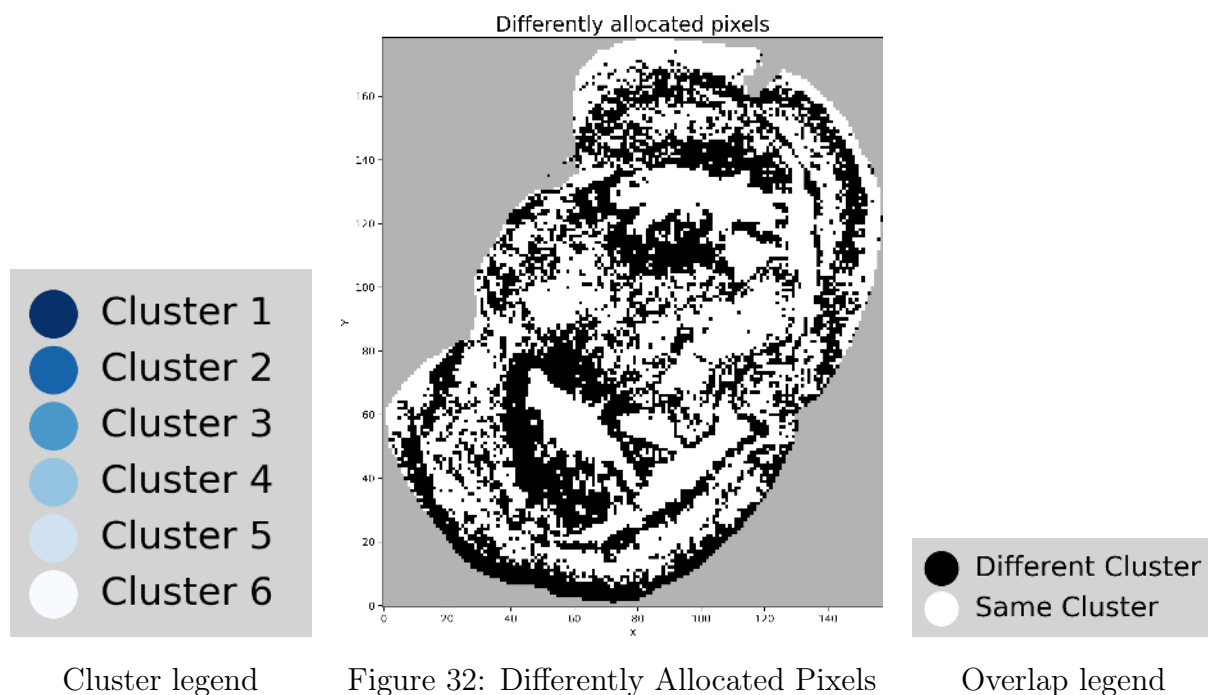


Figure 31: K-means Bray Curtis clusters



Cluster legend

Figure 32: Differently Allocated Pixels

Overlap legend

In Figure 33 and 34, is shown the internal composition and coefficient of variation of the Bray Curtis k-means clusters, divided by peptides, lipids, and glycans so that we can visually confirm that the content of the clusters is consistent with the assignment made manually by the biologist. At the top of the figure you can also find the number of elements present for each cluster.

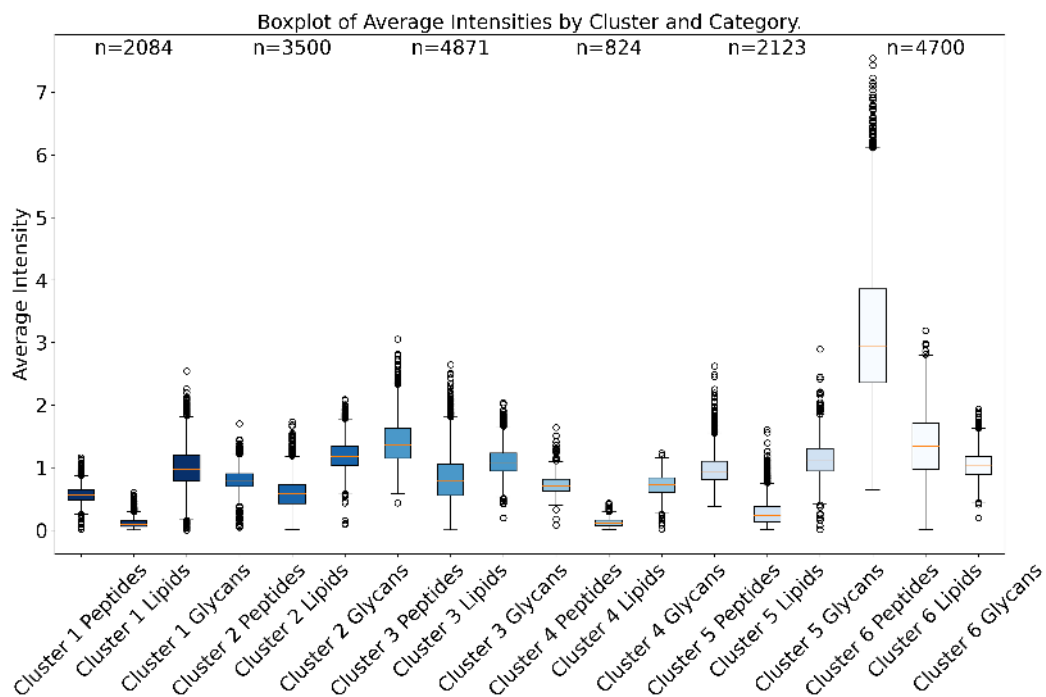


Figure 33: Boxplot of Intensities by Cluster and Category with K-means - Bray Curtis with Penalization

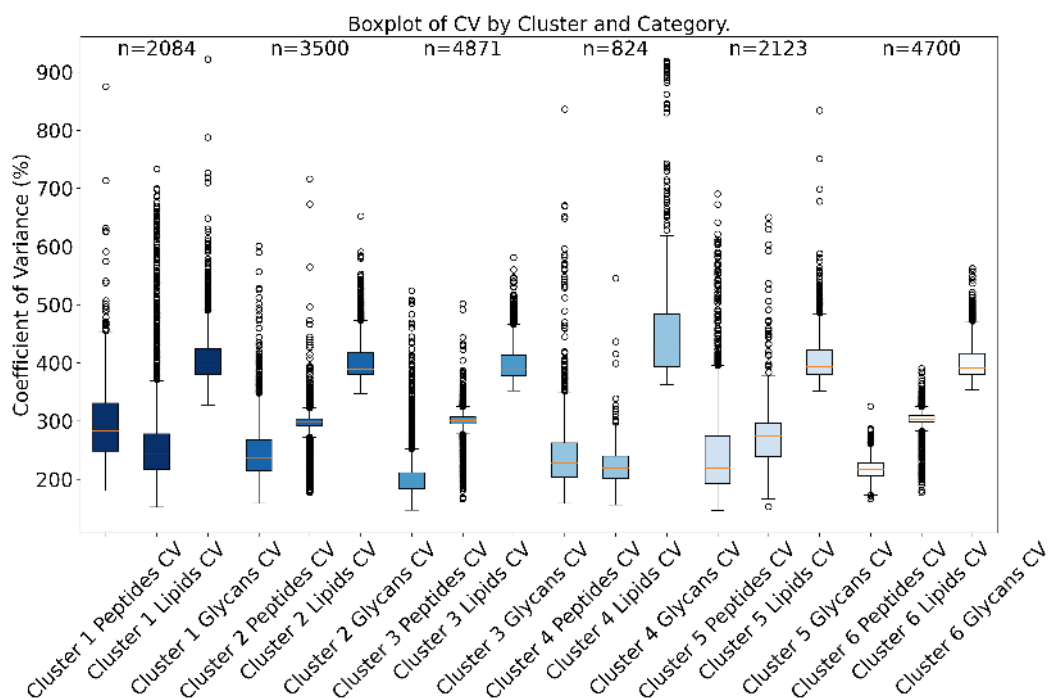


Figure 34: Boxplot of Coefficient Variation by Cluster and Category with K-means - Bray Curtis with Penalization

5 Discussion

5.1 Insights from the Resulting Clusters

The analysis of clusters, as revealed by the ARI, NMI, and Cluster Overlap metrics, provides a comprehensive understanding of the underlying structures within the dataset. The Cluster Overlap metric, in particular, provides direct evidence of the consistency between different clustering methods by evaluating the percentage of shared data points or pixels.

A noteworthy observation is the enhanced performance demonstrated by the Bray Curtis metric in conjunction with the k-means clustering. This is indicative of the possibility that certain patterns within the data might align more favorably with specific distance metrics. The results also underscore the slight yet consistent improvements upon the introduction of spatial penalties. Although the results were not groundbreaking, the good performance with spatial penalization is undeniable, reinforcing the importance of factoring in spatial attributes when analyzing data. Particularly, the Euclidean distance, which is often assumed to be the go-to metric for many applications, displayed suboptimal performance in this context. This observation lends credence to the supposition that Euclidean distance might not be the most suitable for mass spectrum data, especially considering the vast range of the spectra (from 0.0 to 210.52). This expansive range could render certain distance metrics less appropriate, as they might not effectively capture the intricacies and nuances present in such data.

The results also underscore the slight yet consistent improvements upon the introduction of spatial penalties. The improvements brought about by spatial penalization in terms of Cluster Overlap percentages further accentuate the importance of incorporating spatial features. Overlooking such attributes could risk missing out on vital relationships and links within the dataset, thereby possibly diluting the depth and precision of the garnered insights.

5.2 Consistency and Interpretability of the Results

The consistently high scores across various metrics, both in traditional and spatially-enhanced clustering models, are testament to the robustness and reliability of the methodologies employed. ARI, which emphasizes the accuracy of the clustering process, combined with NMI, which sheds light on the shared information across different clusters, and Cluster Overlap, which provides a quantitative measure of consistency, together paint a comprehensive picture. This tri-metric approach not only assures the robustness of the results but also enhances their interpretability, enabling a more seamless translation from raw data patterns into meaningful, actionable insights.

6 Conclusions and Future Work

6.1 Key Findings and Implications

This research journey has brought to the forefront the critical importance of judiciously selecting distance metrics for clustering tasks. The standout performance of the Bray Curtis metric, particularly within the non-spatial GMM framework, underscores this

point. Furthermore, the integration of spatial attributes into the clustering process, drawing inspiration from the HPM model, has been illuminated as a crucial aspect of data analysis, with its absence potentially leading to significant omissions in data insights. These discoveries bear wide-ranging implications, extending beyond the immediate realm of this research, to areas where spatial data plays an integral role, encompassing fields from biology to geospatial studies and more.

6.2 Limitations and Challenges

While the research results provide a solid foundation, it is important to acknowledge inherent limitations. The efficacy of a distance metric, although demonstrated in this study, might vary contingent upon the unique characteristics and nuances of different datasets. Furthermore, the function chosen to penalize distance, may not necessarily be the optimal one for every context or dataset. There is a wide array of potential penalization functions, and the one selected here represents only a singular approach from a broader toolkit. It is conceivable that other penalization functions might yield differing results, further influencing the clustering outcomes.

Some of the algorithms, particularly when grappling with voluminous datasets, are computationally intensive, raising concerns regarding scalability and efficiency. Moreover, the introduction of spatial penalties, although beneficial, operates on the assumption that spatial significance is uniformly applicable across all data points—an assumption that might not universally hold true. It is also worth noting that spatial penalization is one of several potential methods to incorporate spatial information. There might be other techniques, yet unexplored in this research, that could provide alternative or complementary benefits in integrating spatial information into clustering tasks.

6.3 Recommendations for Future Research

Venturing ahead, it would be worthwhile for subsequent research initiatives to delve into the realm of adaptive spatial penalization. This would involve dynamically ascertaining the spatial significance based on inherent data attributes, offering a more nuanced approach. There also exists a promising avenue in exploring hybrid distance metrics, potentially amalgamating the strengths of multiple metrics to achieve superior clustering outcomes. As the world stands on the cusp of a technological revolution with artificial intelligence at its helm, integrating sophisticated deep learning techniques into the clustering paradigm could usher in an era of unprecedented data analysis capabilities.

Appendix A

Implementation of Algorithms and Code Fragments

1 Hierarchical Clustering (HCLUST)

The hierarchical clustering algorithm was executed with the below Python code.

```
1 # Computation of the distance matrix using Euclidean distance
2 distances = pdist(df_IM.values, metric='euclidean')
3
4 if weight == True:
5     distances = distances * distanzeSpaziali
6
7 # Execution of the hierarchical clustering algorithm
8 linkage_matrix = hierarchy.linkage(distances, method='complete')
9
10 # Criterion for dendrogram cut
11 threshold = 6 # Modify based on desired clusters
12
13 # Obtain the clusters
14 clusters = hierarchy.fcluster(linkage_matrix, threshold, criterion
15                               ='maxclust')
```

Listing A.1: Hierarchical Clustering Implementation

Different metrics were used during the various runs by changing the content of the variable `metods`.

2 K-means

Prior to introducing the `np.random.seed(42)`, a seed for reproducibility of results, The k-means algorithm was initially run multiple times to ensure that the centroid initialization did not significantly affect the outcomes. Below is the general code for the k-means clustering:

```
1 from scipy.spatial.distance import cdist
2 import numpy as np
3 def kmeans_standard(X, df_XY, n_clusters, weight, max_iterations
4                     =100):
5     np.random.seed(42)
6     # Centroids initialization at random
7     index = np.random.choice(range(len(X)), n_clusters, replace=
8                             False)
9     centroids = X.values[index]
10    centroidsXY = df_XY.values[index]
11    for _ in range(max_iterations):
12        # Computation of distances between points and centroids
13        # using Euclidean distance
14        distances = cdist(X.values, centroids, metric='euclidean')
15        if weight:
16            cdistXY = cdist(df_XY.values, centroidsXY, metric='
17                            euclidean')
18            PenalizationFactor = 1 - np.exp(-cdistXY)
19            distances = distances * PenalizationFactor
20        # Points assignment to the cluster with the minimum
21        # distance
22        clusters = np.argmin(distances, axis=1)
23        # Centroids update as the mean of points in the cluster
24        new_centroids = np.array([X.values[clusters == k].mean(
25                                axis=0) for k in range(n_clusters)])
26        new_centroidsXY = np.array([df_XY.values[clusters == k].
27                                mean(axis=0) for k in range(n_clusters)])
28        # Convergence check
29        if np.all(centroids == new_centroids):
30            break
31        centroids = new_centroids
32        centroidsXY = new_centroidsXY
33    return clusters, centroids, distances
34
35# Multiple K-means execution
36num_iterations = 100
37n_clusters = 6
38all_clusters = [kmeans_standard(df_IM, df_XY, n_clusters, weight)
39                [0] for _ in range(num_iterations)]
40all_clusters[0] += 1
```

Listing A.2: K-means Implementation

Also here, as with the `hclust` different metrics were used during the various runs by changing the content of the variable `metods`.

3 Spatial Penalization

Spatial penalization was integrated into the clustering algorithms. For hierarchical clustering, the penalization was defined as:

```
1 pdistXY = pdist(df_XY.values, metric='euclidean')
2 PenalizationFactor = 1 - np.exp(pdistXY)
```

Listing A.3: Spatial Penalization for HCLUST

For k-means, it was defined as:

```
1 cdistXY = cdist(df_XY.values, centroidsXY, metric='euclidean')
2 PenalizationFactor = 1 - np.exp(-alpha*cdistXY)
3 distances = distances * PenalizationFactor
```

Listing A.4: Spatial Penalization for K-means

The penalty in both clustering algorithms is executed when weight equals True.

Bibliography

- [1] Cosine similarity. *Information and Software Technology*, 2019. Related reference: Jiawei Han, ... Jian Pei, in *Data Mining (Third Edition)*, 2012.
- [2] Walid M. Abdelmoula, Begona Gimenez-Cassina Lopez, Elizabeth C. Randall, Tina Kapur, Jann N. Sarkaria, Forest M. White, Jeffrey N. Agar, William M. Wells, and Nathalie Y. R. Agar. Peak learning of mass spectrometry imaging data using artificial neural networks. *Nature Communications*, 12:5544, 2021.
- [3] M. Aichler and A. Walch. Maldi imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. *Laboratory Investigation*, 2015.
- [4] Simone Avesani, Eva Viesi, Luca Alessandrì, Giovanni Motterle, Vincenzo Bonnici, Marco Beccuti, Raffaele Calogero, and Rosalba Giugno. Stardust: Improving Spatial Transcriptomics Data Analysis through Space-aware Modularity Optimization-based Clustering. *GigaScience*, 11:giac075, 08 2022.
- [5] Rafael Batres. Chebyshev distance. <https://www.sciencedirect.com/topics/computer-science/chebyshev-distance>, 2012. In: *Computer Aided Chemical Engineering, 11th International Symposium on Process Systems Engineering*.
- [6] Kristin J Boggio, Emmanuel Obasuyi, Ken Sugino, Sacha B Nelson, Nathalie YR Agar, and Jeffrey N Agar. Recent advances in single-cell maldi mass spectrometry imaging and potential clinical impact. *Expert Review of Proteomics*, 8(5):591–604, 2011.
- [7] Amanda Rae Buchberger, Kellen DeLaney, Jillian Johnson, and Lingjun Li. Mass spectrometry imaging: A review of emerging advancements and future insights. *Anal. Chem.*, 90(1):240–265, 2018.
- [8] G. Capitoli, S. Colombara, A. Cotroneo, F. De Caro, R. Morandi, C. Schembri, A.G. Zapiola, and Francesco Denti. Detecting latent spatial patterns in mass spectrometry brain imaging data via bayesian mixtures. In *Book of the Short Papers SEAS IN 2023*, pages 1127–1132. Pearson, 2023.
- [9] Dominic Edelmann, Tamás F. Móri, and Gábor J. Székely. On relationships between the pearson and the distance correlation coefficients. *Statistics Probability Letters*, 169:108960, February 2021.
- [10] Pasi Fränti and Sami Sieranoja. How much can k-means be improved by using better initialization and repeats? *Pattern Recognition*, 93:95–112, 2019.
- [11] Eshita Garg and Muhammad Zubair. Mass Spectrometer, January 2023. In: *Stat-Pearls [Internet]*. Treasure Island (FL): StatPearls Publishing; January 2023.

- [12] Richard J.A. Goodwin. Sample preparation for mass spectrometry imaging: Small mistakes can lead to big consequences. *Journal of Proteomics*, 75(16):4893–4911, August 2012.
- [13] Hans-Friedrich Köhn and Lawrence J. Hubert. Hierarchical Cluster Analysis. *Wiley StatsRef: Statistics Reference Online*, 2015. Update based on original article by Lawrence Hubert, Wiley StatsRef: Statistics Reference Online, 2014, John Wiley Sons, Ltd.
- [14] Sheroze Liaquat, Muhammad Fahad Zia, Omer Saleem, Zeeshan Asif, and Mohamed Benbouzid. Performance analysis of distance metrics on the exploitation properties and convergence behaviour of the conventional firefly algorithm. *Applied Soft Computing*, 126:109255, 2022.
- [15] Fatima A. Merchant and Kenneth R. Castleman. Euclidean distance. <https://www.sciencedirect.com/topics/engineering/euclidean-distance>, 2023. In: Microscope Image Processing (Second Edition).
- [16] Raf Van de Plas Nico Verbeeck, Richard M. Caprioli. Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. *Mass Spectrometry Reviews*, 39(3):245–291, October 2019.
- [17] Fred E. Szabo PhD. Manhattan distance. <https://www.sciencedirect.com/topics/mathematics/manhattan-distance>, 2015. In: The Linear Algebra Survival Guide.
- [18] Douglas Reynolds. *Gaussian Mixture Models*, pages 659–663. Springer, 2007.
- [19] C. Ricotta and J. Podani. On some properties of the bray-curtis dissimilarity and their ecological meaning. *Ecological Complexity*, 31:201–205, September 2017.
- [20] Tatiana C. Rohner, Dieter Staab, and Markus Stoeckli. Maldi mass spectrometric imaging of biological tissue sections. *Mechanisms of Ageing and Development*, 126(1):177–185, 2005. Functional Genomics of Ageing II.
- [21] scikit-learn developers. Adjusted rand score. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html, 2023. Scikit-learn: Machine Learning in Python.
- [22] scikit-learn developers. Normalized mutual information score. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.normalized_mutual_info_score.html, 2023. Scikit-learn: Machine Learning in Python.
- [23] Tina Smets, Nico Verbeeck, Marc Claesen, Arndt Asperger, Gerard Griffioen, Thomas Tousseyn, Wim Waelput, Etienne Waelkens, and Bart De Moor. Evaluation of distance metrics and spatial autocorrelation in uniform manifold approximation and projection applied to mass spectrometry imaging data. *Anal. Chem.*, 91(9):5706–5714, 2019.
- [24] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101:1566–1581, 2006.