analytical chemistry

pubs.acs.org/ac

Article

Spatiochemical Characterization of the Pancreas Using Mass Spectrometry Imaging and Topological Data Analysis

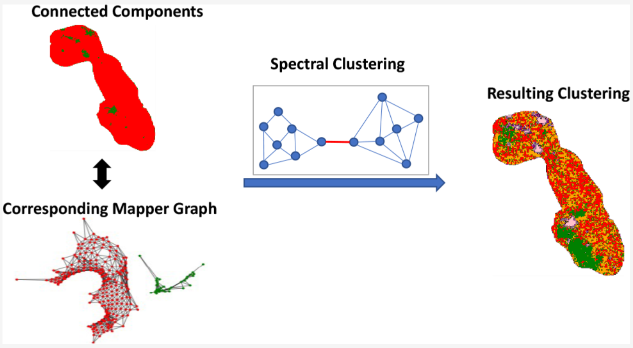
Helena Derwae,* Melanie Nijs, Axel Geysels, Etienne Waelkens, and Bart De Moor

Cite This: *Anal. Chem.* 2023, 95, 10550−10556

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Mass Spectrometry Imaging (MSI) is a technique used to identify the spatial distribution of molecules in tissues. An MSI experiment results in large amounts of high dimensional data, so efficient computational methods are needed to analyze the output. Topological Data Analysis (TDA) has proven to be effective in all kinds of applications. TDA focuses on the topology of the data in high dimensional space. Looking at the shape in a high dimensional data set can lead to new or different insights. In this work, we investigate the use of Mapper, a form of TDA, applied on MSI data. Mapper is used to find data clusters inside two healthy mouse pancreas data sets. The results are compared to previous work using UMAP for MSI data analysis on the same data sets. This work finds that the proposed technique discovers the same clusters in the data as UMAP and is also able to uncover new clusters, such as an additional ring structure inside the pancreatic islets and a better defined cluster containing blood vessels. The technique can be used for a large variety of data types and sizes and can be optimized for specific applications. It is also computationally similar to UMAP for clustering. Mapper is a very interesting method, especially its use in biomedical applications.



MASS SPECTROMETRY IMAGING

Mass spectrometry imaging (MSI), which is based on mass spectrometry (MS), is used to visualize the spatial distribution of ions and molecules inside tissue samples.[1,2] Currently, a popular technique to analyze MSI data is unsupervised dimensionality reduction such as Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP), t-distributed Stochastic Neighbor Embedding (t-SNE), and Non-negative Matrix Factorization (NMF).[3−7]

TOPOLOGICAL DATA ANALYSIS

Topological Data Analysis (TDA) has proven efficient in analyzing high dimensional biological and medical data.[8−10] For the latest work on TDA for biomedical applications, we refer the reader to Singh et al. (2023) and Skaf and Laubenbacher (2022).[8,11]

TDA is based on topology, which is the mathematical study of shape. In fields such as biology and medicine, much information can be obtained from looking at the shape of objects, which can lead to an increase in understanding. In this manner, shape can be seen as a form of data by itself.[12]

TDA builds on the ideas of simplices and homology. A point cloud is used to represent data, which is often high dimensional and thus not easily visualizable. The data can be interpreted as multiple building blocks called simplices, where a 0-dimensional simplex is a point, and 1-, 2-, and 3-dimensional simplices are edges, triangles, and tetrahedra, respectively. Higher dimensional simplices are rarely used because of their lack of interpretability. When multiple simplices are combined, they form a simplicial complex, under the condition that all the faces of every simplex in the complex are part of the complex. A simplicial complex can be interpreted as the skeleton of the data set, which represents the global structure, while mainly passing over the local structures. For example, when looking at a skeleton it shows the overall "shape" of a body.[13] An MSI data set is typically very noisy, meaning that the local structures can be very uncertain. The global structures, however, are less sensitive to this noise. Because of this, it makes sense to look at the global structures (or the skeleton) of the data set. The use of simplicial complexes also demonstrates a fundamental idea in TDA: connectivity is more important than distance.[14] The topology of a simplicial complex, and its corresponding data set, can be analyzed using its homology.[13] Zero-dimensional homology

Received: December 15, 2022
Accepted: June 7, 2023
Published: July 4, 2023

ACS Publications

https://doi.org/10.1021/acs.analchem.2c05606

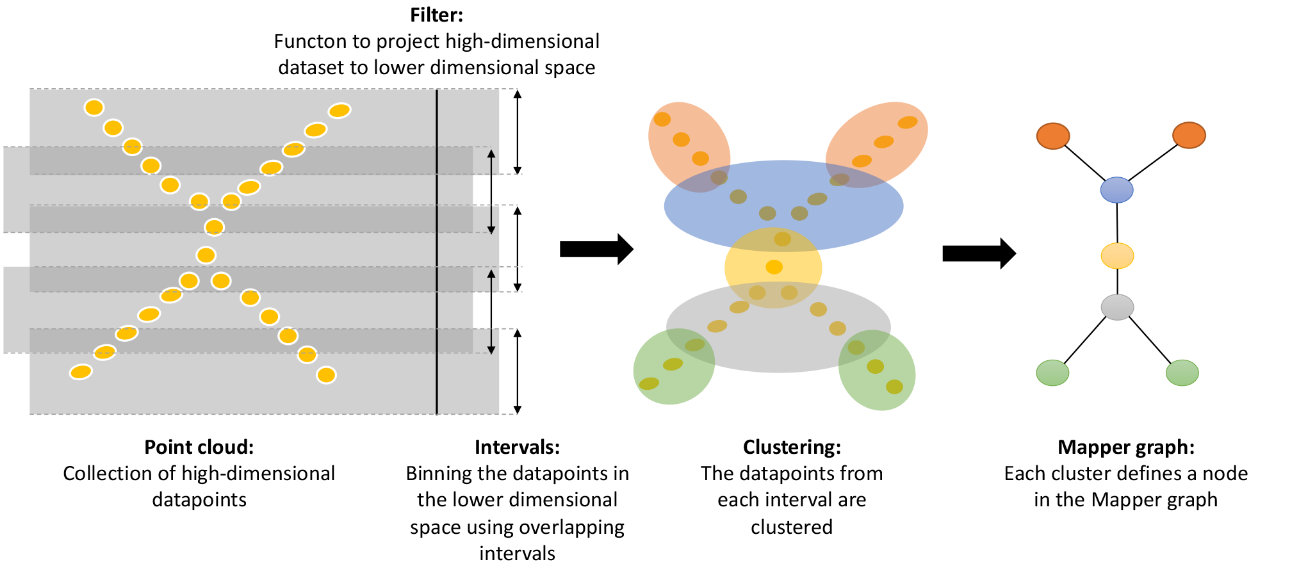

Filter:
Functon to project high-dimensional dataset to lower dimensional space
Point cloud:
Collection of high-dimensional datapoints
Intervals:
Binning the datapoints in the lower dimensional space using overlapping intervals
Clustering:
The datapoints from each interval are clustered
Mapper graph:
Each cluster defines a node in the Mapper graph

Figure 1. Example of the Mapper algorithm for a two-dimensional data set. Figure modified from Chazal and Michel.[18]

analyzes the connected components in a simplicial complex, while 1- and 2-dimensional homologies look at loops and voids, respectively. Again, higher dimensional homologies are not commonly used, since they are harder to inspect visually.

TDA has three main properties: coordinate invariance, deformation invariance, and compression.[14−16] These properties show that one could alter the data set and its simplicial complex without changing its topology. For example, one could deform the simplicial complex under the condition that no connected components, loops, or voids are broken or filled in. Because of these properties, TDA is more robust than many other data analysis techniques and it is highly scalable, giving it the qualities needed for analyzing big data sets.[15−17]

Two main techniques are used within the field of TDA. The first technique is called persistent homology[13,17,18] and was first introduced in 1990.[19−23] Recent developments such as persistent landscapes[24] have increased the popularity of the technique in the latest years. Two recent examples of its use in biomedical applications can be found in Takahashi et al. (2022) and Klaila et al. (2023).[25,26]

The second technique, called Mapper, was introduced by Singh et al. in 2007.[27] Mapper has shown to be promising for clustering and classification applications[28−30] and will be the focus of this work.

The Mapper algorithm results in a graph representing the high dimensional data set in a more comprehensible manner. Similar data points will either be inside the same node in the graph or in closely connected nodes. Different areas of the graph will represent different parts of the data.[29] In this work, Mapper will be used to dive into a specific data set with the aim of discovering new tissue types.

Since UMAP is one of the current state-of-the-art techniques for MSI analysis, we will compare our results with a clustering using UMAP, based on the spatial clustering part of Smets et al. (2020).[31] UMAP too is based on TDA principles, as it builds fuzzy simplicial complexes.[32] A more extensive explanation about TDA can be found in Skaf and Laubenbacher (2022).[8]

EXPERIMENTAL SECTION

Data Acquisition. The used data comes from MSI experiments performed on healthy mouse pancreatic tissue, using a Bruker rapifleX MALDI-TOF mass spectrometer. The data sets correspond to Sample 2 and Sample 1 in Smets et al. (2020) and (2019),[3,31] respectively, and contain 14 791 and 10 606 pixels, both with 14 000 m/z bins. Sample 2 will be used as main data set in this work. For more information about the data acquisition, we refer the reader to these papers.

Mapper for Clustering. The Mapper algorithm is composed of three steps:

1. A filter function has to be chosen to project the data points to a lower dimensional space.
2. This space is divided in a fixed number of overlapping intervals. Each data point will belong to one or more intervals.
3. For each interval, a clustering is done of the data points belonging to that interval. This clustering is executed in the original high dimensional space.

Each cluster forms a node in the resulting Mapper graph. Clusters containing common data points will be connected in the graph. This method can be seen in Figure 1.

In order to make the Mapper graph, the Python toolbox giotto-tda[33] was used. To cluster the resulting Mapper graph, the scikit-learn implementation[34,35] of the spectral clustering method was used.[36] The Calinski-Harabasz (CH) index[37] was used as a quantitative metric for the clusters, again, using the scikit-learn implementation.[35,38] The CH score is an unsupervised clustering evaluation metric based on the Between-Cluster Dispersion (BCD) and the Within-Cluster Dispersion (WCD), which can be calculated as follows:

$$\mathrm{BCD} = \frac{\sum_{k=1}^{K} n_k \lvert c_k - c \rvert^2}{K - 1} \tag{1}$$

$$\mathrm{WCD} = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} \lvert d_i - c_k \rvert^2}{N - K} \tag{2}$$

$$\mathrm{CH} = \frac{\mathrm{BCD}}{\mathrm{WCD}} \tag{3}$$

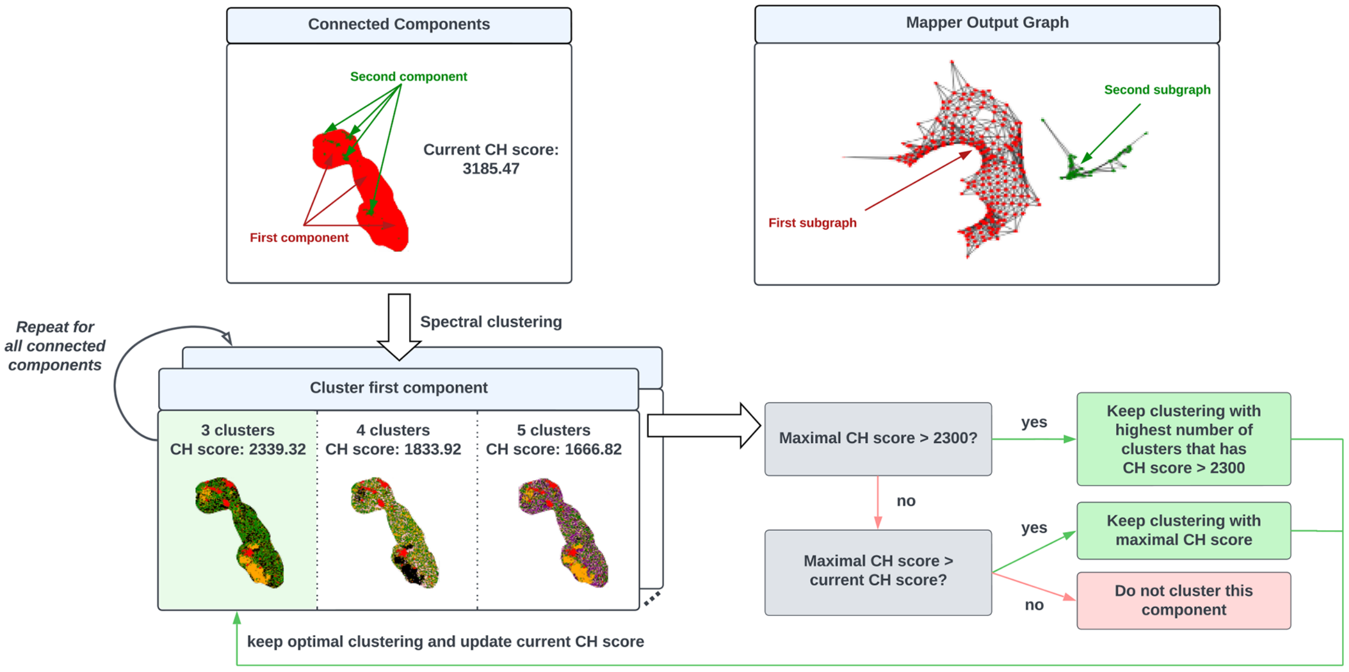

Connected Components
Second component
Current CH score: 3185.47
First component
Mapper Output Graph
Second subgraph
First subgraph
Spectral clustering
Repeat for all connected components
Cluster first component
3 clusters CH score: 2339.32
4 clusters CH score: 1833.92
5 clusters CH score: 1666.82
Maximal CH score > 2300?
yes
Keep clustering with highest number of clusters that has CH score > 2300
no
Maximal CH score > current CH score?
yes
Keep clustering with maximal CH score
no
Do not cluster this component
keep optimal clustering and update current CH score

Figure 2. Diagram of the clustering workflow together with the original constructed Mapper graph. The parameters to construct the Mapper graph were UMAP, 25 intervals, 25% overlap. The connected components correspond to the subgraphs of the Mapper graph. The original "current score" is the CH score obtained from the connected components. When a clustering is chosen for a component, the current score gets updated.

Here, K represents the number of clusters, n_k is the number of data points in cluster k, N is the total number of data points, c_k is the centroid of cluster k, c is the centroid of all data points, and d_i is a data point belonging to cluster k. A clustering with a higher CH score will have denser and more distinctive clusters.

Algorithm. In order to have a benchmark clustering score, we reproduced the spatial clustering part of Smets et al. (2020)[31] and calculated the CH score of the resulting clustering. Since UMAP is a stochastic method, the code was executed 50 times, and an average score was computed. The resulting benchmark score was 2297, which will be rounded up to 2300.

Making a Mapper graph from the data set resulted in multiple subgraphs, which will be called "connected components". These connected components represent different tissue types and form an initial clustering. For this clustering, an initial CH score was computed. Afterward, for each connected component separately, a clustering was made with three, four, and five clusters. For each component, the clustering with the highest number of clusters that has a CH score higher than 2300 or higher than the CH score using less clusters is kept. This algorithm can be seen in Figure 2. Only components containing more than 10 nodes were considered for clustering.

Parameters. The parameters that have to be determined for the Mapper method are the filter function, the dimensionality to which the filter function should project the data points, the number of overlapping intervals, the overlap percentage of these intervals, and the clustering method; see Table S1 in the Supporting Information. The filter functions that were tested are Singular Value Decomposition (svd)[39] and UMAP.[32] UMAP was tested because it is currently one of the state-of-the-art techniques for dimensionality reduction for MSI data, and svd was tested because it is a relatively fast and simple, commonly used method.[39]

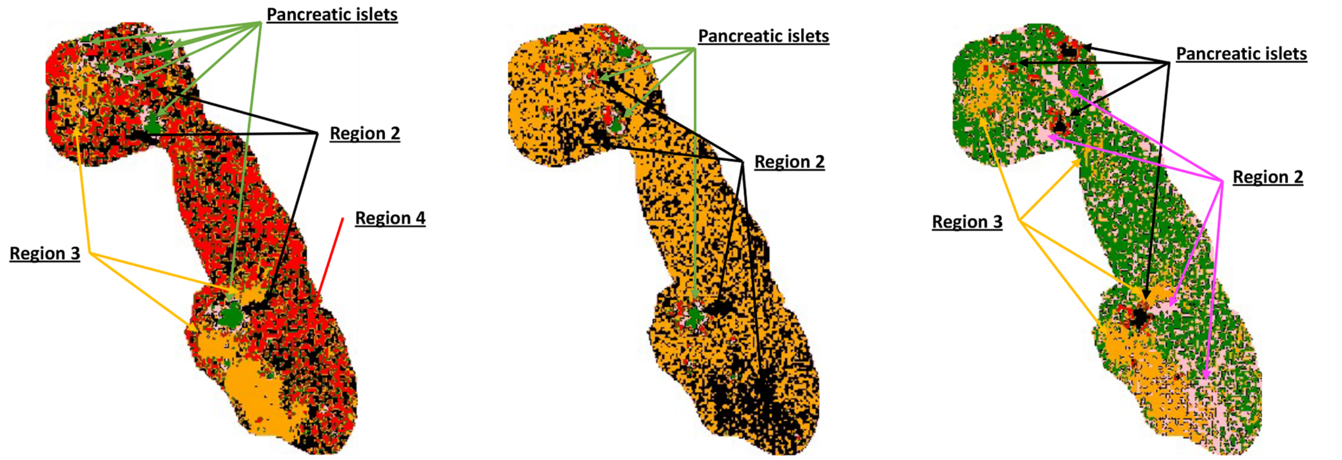
In order to cluster the data points from each interval, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method[40] was used. DBSCAN is a very flexible clustering method which does not assume underlying characteristics of the data except for the fact that clusters are dense regions separated by less dense regions. Unlike other clustering methods, such as k-means,[35,41] DBSCAN has no need for a predefined number of clusters. This is very useful for clustering inside the Mapper algorithm.

The filter function and the dimensionality of the reduced space have the biggest influence on the results. Because of this, we compared the results using UMAP and svd as filter functions. Afterward, a numerical and a visual comparison between our results and the UMAP clustering was done. For all results discussed below, fixed settings were used for some of the parameters. Within the DBSCAN clustering method, we used *min_samples = 20* because it gave a small number of large connected components. For all other parameters inside the DBSCAN method, the default values were used.[42] For the UMAP filter function, *n_neighbors = 15*, *min_dist = 0.1*, and *metric = cosine* were used. These were chosen based on Smets et al. (2020) and (2019).[3,31]

The number of intervals and the overlap percentage were determined using an extensive parameter search. A test interval for both parameters was determined based on an initial random search. Afterward, the algorithm was executed for all parameter combinations within these intervals. The final CH scores were compared, and some parameter settings were visually analyzed.

RESULTS AND DISCUSSION

Depending on the parameters chosen for the Mapper algorithm, different results were obtained. For most parameter combinations, the Mapper graph consisted of one or two components. When two components were obtained, one



(a) UMAP: 2 dimensions, 15 intervals, 25% overlap, 5 clusters CH score = 3023 (b) svd: 2 dimensions, 25 intervals, 85% overlap, 5 clusters CH score = 4655 (c) svd: 4 dimensions, 5 intervals, 55% overlap, 5 clusters CH score = 4724

Figure 3. Clusters obtained from the Mapper graph using different settings.

component represented the pancreatic islets and the other component contained the remaining parts of the pancreas; an example of these components can be seen in Figure 2.

Comparing Filter Functions. When UMAP was used as a filter function, two connected components were often found. The first component represents the pancreatic islets, which are often further divided into an inner and outer part by our algorithm. This ring structure could be caused by artifacts, which could be introduced because of high dissimilarity between the pancreatic islets and the surrounding tissue. However, in mouse pancreatic islets the core consists of β-cells, while other types of cells are located at the periphery of the islets.[43−45] Further analysis is required to uncover whether the ring structure is based on a biological structure or caused by artifacts. Clustering the other connected component usually led to multiple different regions: *region 2*, which contains blood vessels and is shown in black in Figure 3a, *region 3* in yellow, and *region 4* in red.

Some drawbacks of using UMAP as a filter function are its higher run time than simpler filter functions and the fact that it is stochastic. The results will thus differ between runs, as will the run time.

Using svd as a filter function for dimensionality reduction to two or three dimensions usually led to one connected component. Clustering this component led to pancreatic islets, a cluster containing the blood vessels, and the remainder of the pancreas. The pancreatic islets were again divided into edges and centra. *Region 3*, which was previously found using UMAP, however, was not found here. An example can be seen in Figure 3b. In contrast to UMAP, svd is deterministic and much faster. More information about the run times can be found in the "Computational requirements" section of the Supporting Information.

Using different filter functions can lead to discovery of different tissue types. UMAP found *region 3*, and svd for two and three dimensions did not. *Region 2*, on the other hand, was better defined using svd. This makes the used method very well suited for explorative studies with the intention of discovering new tissue types.

Dimensionality reduction to four dimensions using svd as a filter function, however, led to interesting results. Now, both *region 2* and *region 3* are found, as can be seen in Figure 3c.

Quantitative Comparison. In order to compare the obtained results with the spatial clustering part of Smets et al. (2020),[31] the CH score was used to measure the fitness of the obtained clusterings. As was mentioned before, this benchmark CH score was 2300. Multiple clusterings obtained with the Mapper algorithm scored better. Since this scoring method uses the ratio of the in-cluster variance and the between-cluster variance, smaller clusters might be penalized. Because of this, a visual analysis was done for the parameter settings that obtained good scores. When only two clusters were obtained, namely the pancreatic islets and the remainder, a high clustering score was obtained of approximately 3200. Since these are the main different tissue types, this clustering is correct, but it does not give sufficiently fine grained information. This clustering will thus not be considered as very good in this work.

When a higher number of clusters resulted in a lower CH score, this clustering was accepted if the obtained clusters were biologically meaningful. The only clusterings that were visually analyzed, however, are the ones that obtained a score equal to or higher than 2300. For example, the experiment which produced Figure 3a resulted in CH scores starting from 3491 for four clusters, going down to 3023 for five clusters and 2357 for seven clusters. In the four cluster case, *region 3* is not clustered separately. Using more clusters thus gives more biological information in this case, even though a lower score is obtained. This shows the importance of the visual inspection.

When using UMAP as a filter function to reduce to two dimensions, the highest score that was obtained in our sparse parameter search was 3567 for 125 intervals, an overlap percentage of 90%, and four clusters. Reducing to three dimensions using UMAP as a filter function resulted in similar scores as reducing to two dimensions; more information can be found in the "Quantitative comparison" section of the Supporting Information.

Using svd as a filter function to map to two or three dimensions resulted in significantly higher CH scores. Multiple clusterings scored above 6000, with the highest observed CH score being 6802 for dimensionality reduction to three dimensions, using five intervals and an overlap percentage of 50%. Reducing the dimensionality to four dimensions using svd resulted in similar CH scores, although some remarkable changes could be seen in the visual inspection. In the Supporting Information, figures are shown corresponding to this analysis (Figures S1−S4), as well as a comparison of the CH scores. Additionally, an analysis of the run time for different parameter combinations can be found in the "Computational requirements" section of the Supporting Information. For example, when using svd, two dimensions, 5% overlap, and five intervals, the data set can be analyzed in less than 30 s on a MacBook Air M1 2020, which is faster than UMAP. When using higher overlap percentages and higher numbers of intervals, however, the algorithm becomes slower.

Visual Comparison. The visual inspection of the results was done by comparing the resulting clusters to known biological regions. A reproduction of the spatial clustering using UMAP can be seen in Figure 4. The obtained clusters represent the pancreatic islets, *region* 3, and two different types of tissue, one of which seems to contain blood vessels. A first clear difference between this benchmark clustering and our results is the fact that the pancreatic islets were clustered internally in our results. Further on, the results when using UMAP as filter function did not differ substantially from our benchmark, as can be seen in Figure 3a.

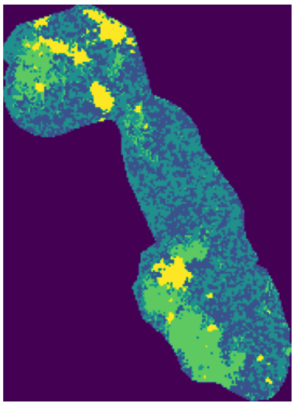


Figure 4. Benchmark clustering made using UMAP and k-means clustering, based on the spatial clustering part of Smets et al. (2020).[31]

When using svd as a filter function, on the other hand, *region* 2 was clustered differently, see Figures 3b and 3c. This region is indicated in black and pink, respectively. A tissue part at the bottom is clustered together with blood vessels running next to the islets. The main blood vessels are located close to the islets because the islets produce hormones, which have to be secreted into the blood flow.[46] Again, the islets were clustered internally to form a ring structure.

When using svd as a filter function to reduce to four dimensions, both *region* 2 and *region* 3 were found. This experiment obtained the most important clusters found by all tested settings. This can be seen in Figure 3c. In Figure S4, the evolution of clusters can be seen for this experiment. The CH score gradually decreases as the number of clusters increases.

Matching Ion Images. In a final experiment, conducted to check the correctness of the clusters, some ion images were analyzed. For each cluster we computed the mean mass spectrum. Using limited peak picking, we looked at the 60 highest peaks in these mean mass spectra in order to compare the ion images corresponding to these peaks with the obtained clusters. Naturally, there is not one m/z value solely responsible for the separation of a cluster. Another problem is the fact that the difference between some clusters is not always present in the highest peaks, and some m/z values are highly present in all tissue parts. Nevertheless, some interesting results can be shown. Here, the results for the core of the islets will be shown. More results are presented in the Supporting Information.

In Figure S16, it can be seen that the highest peak for this cluster is 5800.77 Da; this corresponds to insulin in our data set, and the corresponding ion image is shown in Figure S12. Insulin is a well-known biomarker for the pancreatic islets. This ion image clearly shows a combination of the cores and edges of the islets. Another interesting ion image, corresponding to the fifth highest peak, 6006.67 Da, shows a clearly higher intensity at the cores of the islets. The shape of the cores matches the high intensity parts of the ion image quite well. For most clusters, spatial mapping is difficult, but for this cluster it was done nevertheless; this can be seen in Figure 5.

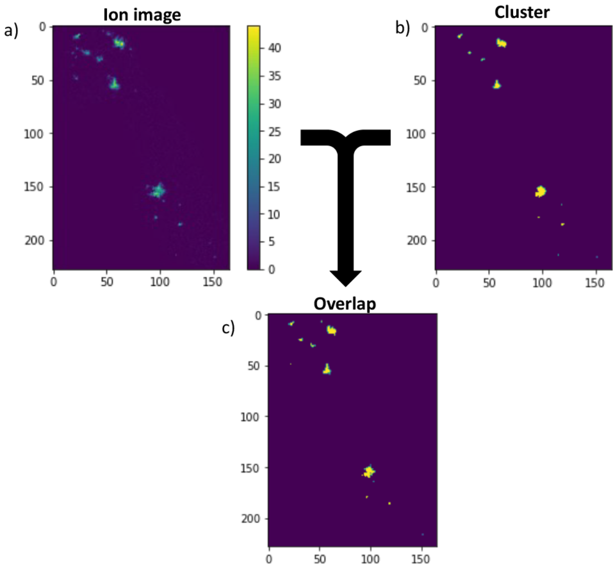


Figure 5. (a) Ion image for m/z value 6006.67. (b) Mask for the cluster representing the cores of the pancreatic islets. (c) Overlap between intensities higher than 15 from the ion image and the cluster.

First of all, a mask was made for the ion image; each pixel with intensity higher than 15 was given a label of 1, and each pixel with lower intensity was given label 0. For this mapping the Dice score was measured. The Dice score can be calculated as two times the number of data points in the overlap of the pictures divided by the number of data points in both pictures summed. $\text{DICE} = \frac{2|A \| B|}{|A| + |B|}$. The resulting Dice score was 0.89, which is close to 1, meaning a high overlap.[47]

Second Data Set. Here, a small analysis is done for the second data set. A more extensive analysis can be found in the corresponding section in the Supporting Information. For the second data set, the same parameters were tested as for the first data set, but only the svd filter was tested. When doing dimensionality reduction to two dimensions, the highest obtained CH score is 7081.16; these scores cannot be directly compared to the score of the first data set. When doing dimensionality reduction to three dimensions, the highest CH score was 6824.95; the clusters for three dimensions look very similar to those for two dimensions. When doing dimensionality reduction to four dimensions, the highest obtained CH score was 7591.59. Just as for the first data set an extra tissue part, *region 3*, is discovered here, as can be seen in Figure 6. As can be seen in Figures S19−S21 in the Supporting Information, the pancreatic islets are again divided into two or three parts. The CH score = 5402.23.

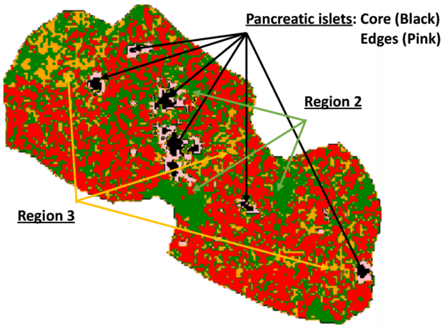


Figure 6. Clusters obtained from the Mapper graph using svd as a filter function for the second data set with the following parameters: 4 dimensions, 15 intervals, 70% overlap, 5 clusters.

■ CONCLUSION

In this work, we provided evidence that the Mapper technique is able to discover the same tissue types as UMAP, which is one of the current state-of-the-art methods for analyzing MSI data. In addition, it was able to find extra clusters. An interesting separation of the pancreatic islets was obtained, together with a more refined cluster containing blood vessels.

The Mapper technique requires multiple changeable parameters, such as the filter function and the clustering method. Even though this is often seen as a disadvantage, it also makes the Mapper technique flexible and suited for explorative studies. Changing the filter function to better fit the application can result in new insights compared to standard filter functions. In future work, it might be worthwhile to look into more specialized filter functions. Also, analyzing the use of different clustering algorithms might be useful.

In conclusion, the Mapper technique seems to be a very promising tool for analyzing biological data and more specifically MSI data.

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.2c05606.

Additional resulting figures from different experiments mentioned in the text, as well as additional CH score oversight and a section on computational requirements (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Helena Derwae − *Department of Electrical Engineering (ESAT), KU Leuven, 3001 Leuven, Belgium;* orcid.org/0000-0002-6807-2057; Email: helena.derwae@gmail.com

Authors

Melanie Nijs − *Department of Electrical Engineering (ESAT), KU Leuven, 3001 Leuven, Belgium*

Axel Geysels − *Department of Electrical Engineering (ESAT), KU Leuven, 3001 Leuven, Belgium*

Etienne Waelkens − *Department of Cellular and Molecular Medicine, KU Leuven, 3001 Leuven, Belgium*

Bart De Moor − *Department of Electrical Engineering (ESAT), KU Leuven, 3001 Leuven, Belgium; Fellow IEEE, SIAM at STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, 3001 Leuven, Belgium*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.2c05606

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by KU Leuven: Research Fund (projects C16/15/059, C3/19/053, C24/18/022, C3/20/117, C3I-21-00316), Industrial Research Fund (Fellowships 13-0260, IOFm/16/004), and several Leuven Research and Development bilateral industrial projects; Flemish Government Agencies: FWO: EOS Project no G0F6718N (SeLMA), SBO project S005319N, Infrastructure project I013218N, TBM Project T001919N; Ph.D. Grant (SB/1SA1319N), EWI: the Flanders AI Research Program, VLAIO: CSBO (HBC.2021.0076) Baekeland Ph.D. (HBC.20192204), European Commission: European Research Council under the European Union's Horizon 2020 research and innovation programme (ERC Adv. Grant grant agreement No 885682); other funding: Foundation "Kom op tegen Kanker", CM (Christelijke Mutualiteit). The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation-Flanders (FWO) and the Flemish Government.

■ REFERENCES

(1) Buchberger, A. R.; DeLaney, K.; Johnson, J.; Li, L. *Anal. Chem.* **2018**, *90*, 240−265.

(2) Murayama, C.; Kimura, Y.; Setou, M. *Biophys. Rev.* **2009**, *1*, 131−139.

(3) Smets, T.; Verbeeck, N.; Claesen, M.; Asperger, A.; Griffioen, G.; Tousseyn, T.; Waelput, W.; Waelkens, E.; De Moor, B. *Anal. Chem.* **2019**, *91*, 5706−5714.

(4) Fonville, J. M.; Carter, C. L.; Pizarro, L.; Steven, R. T.; Palmer, A. D.; Griffiths, R. L.; Lalor, P. F.; Lindon, J. C.; Nicholson, J. K.; Holmes, E.; Bunch, J. *Anal. Chem.* **2013**, *85*, 1415−1423.

(5) van der Maaten, L.; Hinton, G. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.

(6) Nijs, M.; Smets, T.; Waelkens, E.; De Moor, B. *Rapid Commun. Mass Spectrom.* **2021**, *35*, No. e9181.

(7) Alexandrov, T. *BMC Bioinform* **2012**, *13*, S11.

(8) Skaf, Y.; Laubenbacher, R. *J. Biomed. Inform.* **2022**, *130*, 104082.

(9) Bukkuri, A.; Andor, N.; Darcy, I. K. *Front. Artif. Intell.* **2021**, *4*, 4.

(10) Rabadan, R.; Blumberg, A. *Topological Data Analysis for Genomics and Evolution: Topology in Biology*; Cambridge University Press: Cambridge, 2019.

(11) Singh, Y.; Farrelly, C. M.; Hathaway, Q. A.; Leiner, T.; Jagtap, J.; Carlsson, G. E.; Erickson, B. J. *Insights Imaging* **2023**, *14*, 58.

(12) Amézquita, E. J.; Quigley, M. Y.; Ophelders, T.; Munch, E.; Chitwood, D. H. *Dev. Dyn.* **2020**, *249*, 816−833.

(13) Munch, E. *J. Learn. Anal.* **2017**, *4*, 47−61.

(14) Saul, N.; Tralie, C. *Scikit-TDA: Topological Data Analysis for Python*; 2019; DOI: 10.5281/zenodo.2533369 (accessed: September 16, 2021).

(15) Lum, P. Y.; Singh, G.; Lehman, A.; Ishkanov, T.; Vejdemo-Johansson, M.; Alagappan, M.; Carlsson, J.; Carlsson, G. *Sci. Rep.* **2013**, *3*, 1236.

(16) Offroy, M.; Duponchel, L. *Anal. Chim. Acta* **2016**, *910*, 1−11.

(17) Vejdemo-Johansson, M.; Skraba, P. *Topology, Big Data and Optimization*; Springer: Cham, Switzerland, 2016.

(18) Chazal, F.; Michel, B. *Front. Artif. Intell.* **2021**, *4*, 4.

(19) Frosini, P. *Bull. Aust. Math. Soc.* **1990**, *42*, 407−415.

(20) Frosini, P. *JCISS* **1992**, *17*, 232−250.

(21) Frosini, P. *Proc. SPIE* **1992**, *1607*, 122−133.

(22) Verri, A.; Uras, C.; Frosini, P.; Ferri, M. *Biol. Cybern.* **1993**, *70*, 99−107.

(23) Moroni, D.; Pascali, M. A. *Pattern Recognition. ICPR International Workshops and Challenges*; Lecture Notes in Computer Science; 2021; pp 211−226.

(24) Bubenik, P. *J. Mach. Learn. Res.* **2015**, *16*, 77−102.

(25) Takahashi, K.; Abe, K.; Kubota, S. I.; Fukatsu, N.; Morishita, Y.; Yoshimatsu, Y.; Hirakawa, S.; Kubota, Y.; Watabe, T.; Ehata, S.; Ueda, H. R.; Shimamura, T.; Miyazono, K. *Nat. Commun.* **2022**, *13*, 5239.

(26) Klaila, G.; Vutov, V.; Stefanou, A. *Supervised topological data analysis for MALDI imaging applications*. arXiv:2302.13948, 2023; https://arxiv.org/abs/2302.13948.

(27) Singh, G.; Memoli, F.; Carlsson, G. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. *Eurographics Symposium on Point-Based Graphics*, 2007.

(28) Hoef, L. V.; Adams, H.; King, E. J.; Ebert-Uphoff, I. *A Primer on Topological Data Analysis to Support Image Analysis Tasks in Environmental Science*. 2022; https://arxiv.org/abs/2207.10552.

(29) Minervino, M. *Topological Data Analysis with Mapper*. https://www.quantmetry.com/blog/topological-data-analysis-with-mapper/ (accessed: December 3, 2021).

(30) Duponchel, L. *J. Spectr. Imaging* **2018**, *7*, 1−10.

(31) Smets, T.; Waelkens, E.; De Moor, B. *Anal. Chem.* **2020**, *92*, 5240−5248.

(32) McInnes, L.; Healy, J.; Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2018; https://arxiv.org/abs/1802.03426.

(33) Tauzin, G.; Lupo, U.; Tunstall, L.; Pérez, J. B.; Caorsi, M.; Medina-Mardones, A.; Dassatti, A.; Hess, K. *giotto-tda: A Topological Data Analysis Toolkit for Machine Learning and Data Exploration*; 2020.

(34) Spectral clustering. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html (accessed: Septemer 28, 2022).

(35) Pedregosa, F.; et al. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(36) von Luxburg, U. *Stat. and Comput.* **2007**, *17*, 395−416.

(37) Aljarah, I.; Habib, M.; Nujoom, R.; Faris, H.; Mirjalili, S. A Comprehensive Review of Evaluation and Fitness Measures for Evolutionary Data Clustering. In *Evolutionary Data Clustering: Algorithms and Applications. Algorithms for Intelligent Systems*; Springer: Singapore, 2021; pp 23−71.

(38) Calinski Harabasz score. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.calinski_harabasz_score.html?highlight=calinsky (accessed: September 28, 2022).

(39) Wall, M. E.; Rechtsteiner, A.; Rocha, L. M. *arXiv: Bio. Ph.* **2003**, *5*, 91−109.

(40) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*,1996; pp 226−231.

(41) Clustering: kmeans. https://scikit-learn.org/stable/modules/clustering.html#k-means (accessed: November 21, 2021).

(42) https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html (accessed: September 28, 2022).

(43) Da Silva Xavier, G. *J. Clin. Med.* **2018**, *7*, 54.

(44) Brissova, M.; Fowler, M. J.; Nicholson, W. E.; Chu, A.; Hirshberg, B.; Harlan, D. M.; Powers, A. C. *J. Histochem. Cytochem.* **2005**, *53*, 1087−1097.

(45) Cabrera, O.; Berman, D. M.; Kenyon, N. S.; Ricordi, C.; Berggren, P.-O.; Caicedo, A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 2334−2339.

(46) Campbell, J. E.; Newgard, C. B. *Nat. Rev. Mol. Cell. Biol.* **2021**, *22*, 142−158.

(47) Carass, A.; Roy, S.; Gherman, A.; Reinhold, J. C.; Jesson, A.; Arbel, T.; Maier, O.; Handels, H.; Ghafoorian, M.; Platel, B.; et al. *Sci. Rep.* **2020**, *10*, 8242.