Prioritization of m/z-values in mass spectrometry imaging profiles obtained using Uniform Manifold Approximation and Projection for dimensionality reduction

Tina Smets,[*,†] Etienne Waelkens,[‡] and Bart De Moor[†,¶]

†*STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, Department of Electrical Engineering (ESAT), KU Leuven, 3001 Leuven, Belgium.*

‡*Department of Cellular and Molecular Medicine, KU Leuven, 3001 Leuven, Belgium.*

¶*Fellow IEEE, SIAM*

E-mail: tina.smets@esat.kuleuven.be

Abstract

Mass spectrometry imaging (MSI) is a promising technique to assess the spatial distribution of molecules in a tissue sample. Non-linear dimensionality reduction methods such as Uniform Manifold Approximation and Projection (UMAP) can be very valuable for the visualization of the massive datasets produced by MSI. These visualizations can offer us good initial insights regarding the heterogeneity and variety of molecular patterns present in the data, but they do not discern which molecules might be driving these observations. To prioritize the m/z-values associated with these biochemical profiles, we apply a bidirectional dimensionality reduction approach taking into account both the spectral and spatial information. The results show that both sources of information are instrumental to get a more comprehensive view on the relevant m/z-values and can support the reliability of the results obtained using

UMAP. We illustrate our approach on heterogeneous pancreas tissues obtained from healthy mice.

Acknowledgements

This work was supported by KU Leuven: Research Fund (projects C16/15/059, C32/16/013, C24/18/022), Industrial Research Fund (Fellowship 13-0260) and several Leuven Research and Development bilateral industrial projects, Flemish Government Agencies: FWO (EOS Project no 30468160 (SeLMA), SBO project I013218N, PhD Grants (SB/1SA1319N, SB/1S93918, SB/151622)), EWI (PhD and postdoc grants Flanders AI Impulse Program), VLAIO (City of Things (COT.2018.018), PhD grants: Baekeland (HBC.20192204) and Innovation mandate (HBC.2019.2209), Industrial Projects (HBC.2018.0405)), European Commission (EU H2020-SC1-2016-2017 Grant Agreement No.727721: MIDAS), ICON MSIPad project (with acknowledgements to Gerard Griffioen (reMynd, Leuven, Belgium) and Arndt Asperger (Bruker Daltonik, Bremen Germany)).

Introduction

Mass spectrometry imaging (MSI) enables the untargeted measurement of biomolecular species and the visualization of their spatial distribution in a variety of tissue sections[1,2] . The combination of these elements gives rise to a powerful tool that allows us to dissect and characterise the biological composition of tissues both in health and disease. To discern the spatial pattern of molecules measured with MSI, their distribution is typically visualized in the form of ion images. These are visualizations that employ a pseudo-color scale to the mass spectral intensities associated with a particular m/z -value resulting in a heat map that is intuitive to interpret. However with the large number of features being measured, visualizing thousands of ions or m/z-values makes it infeasible to gain rapid insight into potential patterns present in the data.

For this reason we can rely on a number of dimensionality reduction and clustering techniques that improve the interpretability of the data through a comprehensive decomposition or visualization thereof. Methods such as principal component analysis (PCA), probabilistic latent semantic analysis (pLSA) and non-negative matrix factorization (NMF) have been particularly useful in this regard.[3] While PCA seeks to determine the orthogonal eigenvectors associated with the largest variance in the data,[4] NMF tries to resemble the original data matrix, X, as well as possible through iterative minimization of the residual squared distance between the products of the two factorized matrices, W and H ($X \approx WH$).[5,6] Yet another approach is taken by pLSA, which relies on a statistical mixture model to decompose the data of underlying latent variables via the iterative expectation-maximization (EM) method.[7] A variety of applications have also shown the value of a neural network approach in the form of Self - Organizing Maps (SOMs).[8–10] Non-linear dimensionality reduction methods such as t-distributed Stochastic Neighbour Embedding (t-SNE) on the other hand have gained popularity as they were shown to outperform methods such as PCA and SOM due to their strong visualization capabilties.[11,12] Uniform Manifold Approximation and Projection (UMAP) is an example of a recent non-linear dimensionality reduction method that is comparable to t-SNE but more scalable towards larger MSI datasets.[12,13] Like t-SNE, UMAP does not impose the strong assumption of a linear relationship between variables that is made by techniques such as Principal Component Analysis (PCA)[4], which is beneficial when working with biological models that are inherently non-linear.

While we have already shown the value of UMAP for the dimensionality reduction of MSI data in earlier work, we want to stress the fact that this method captures all features in the reduced space irrespectively of the number of components selected.[14] . This is an important aspect when visualizing data in two or three dimensions because the number of features embedded in the reduced space is not restricted by these two or three dimensions, even if more features are present in the data. This characteristic empowers strong data

visualizations and is absent from methods such as PCA or clustering approaches.

A downside of non-linear methods however is their limited explicability Although providing excellent and rapid insight into the major patterns or trends present in a tissue, they do not reveal which ions or m/z- values are driving these patterns and observations. In the end, it is through the identification of ions that are co-localized with certain regions of interest or differ in expression between samples of different conditions from which we can elucidate disease mechanisms or facilitate biomarker discovery.[15]

It is with this idea in mind that we , as illustrated in figure 1, propose a workflow starting from the hyperspectral images obtained using UMAP to identify those ions that are driving the different profiles as visualized by a color gradient. We do this by extracting these profiles according to their corresponding color of interest, followed by a spectral and spatial-driven prioritization of m/z-values. For the spectral part we extract the median peak spectral intensities from the data associated with the pattern of interest and apply peak picking to select those peaks that differ the most from the background profile (i.e. the residual tissue pixels). While the latter enables us to build a spectral prioritization of m/z-values, we are also interested in a spatial prioritization such that we can detect co-localized ions that are specific for the extracted profile. For this spatial prioritization we start from the two-dimensional embedded pixel space, which we cluster using the Hierarchical Density-based Spatial Clustering for Applications with Noise (HDBSCAN) algorithm[16] to group similar ion images. This enables us to correlate the selected profile to the average correlation per cluster and gain rapid insights into promising co-localization patterns. We specifically aim to include both the spectral and spatial information because this enables a more comprehensive characterization of the information present in the different profiles, which would not be possible when considering these sources of information individually.

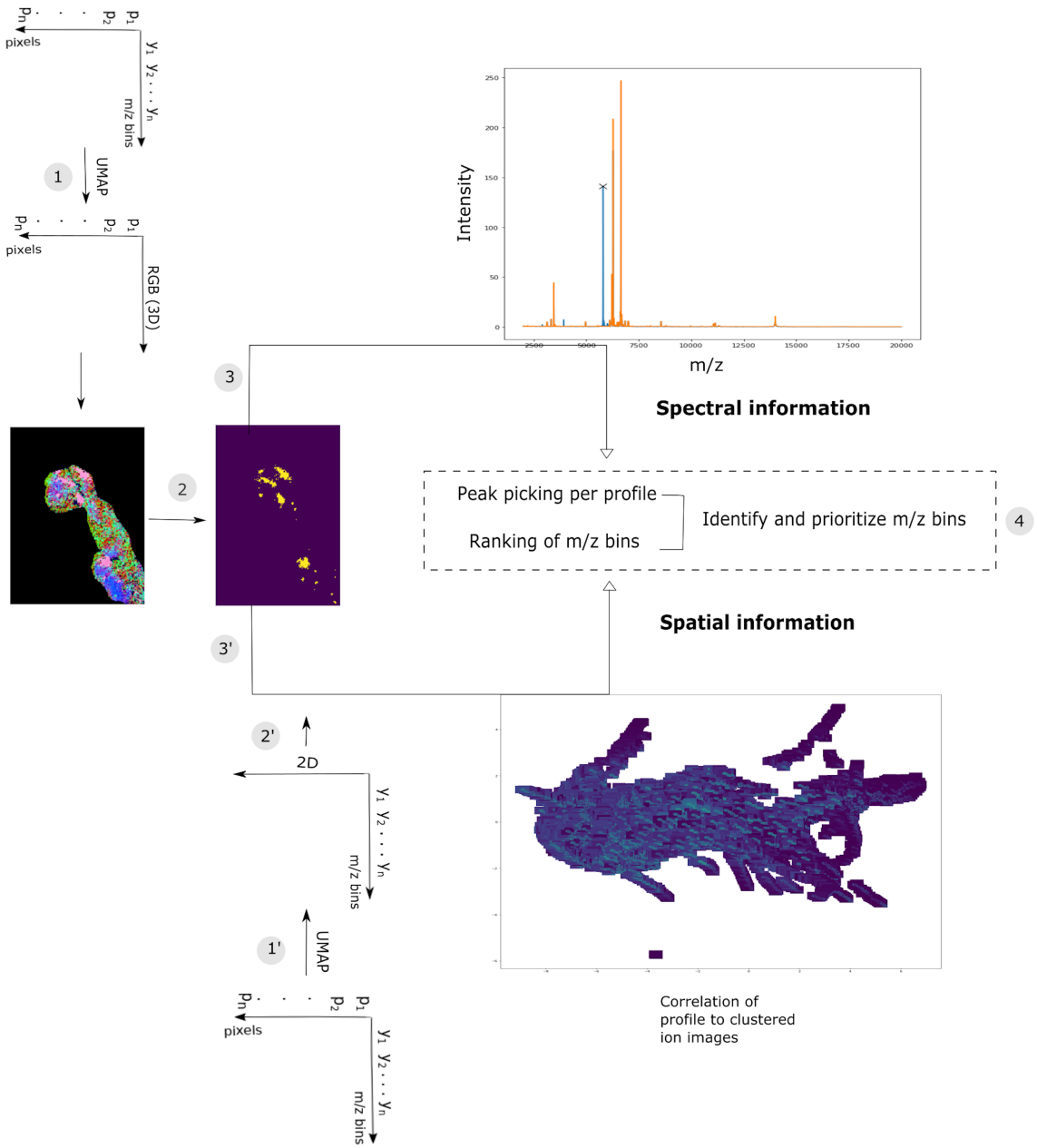

Figure 1: Method overview. We start by reducing the dimensionality of the MSI feature space (m/z bins) to three dimensions (1). This three-dimensional embedding is used for the hyperspectral visualization of the data which is then used to extract the different profiles according to their color representing similar biochemical content (2). The binary representation of a selected profile is then used (3) to identify peaks driving this profile in comparison to the overall tissue (spectral information). In parallel, this binary representation is used to match this pattern through correlation with clusters of similar ion images (3'). These are obtained through clustering (2') of the two-dimensional embedding of the pixel space rather than the m/z space (1'). By ranking the identified peaks obtained through the spectral and spatial information we are able to get a comprehensive view on the peaks that are driving the hyperspectral visualization obtained using UMAP (4). The combined ranking is obtained by combining the spectral and spatial ranking per tissue by giving a higher priority to those m/z-values that are i) prioritized both in a spatial and spectral manner within the same tissue and ii) observed in the same profile across multiple tissues

We demonstrate our method on MSI data collected from healthy mouse pancreas samples, because this tissue is heterogeneous while it contains endogenous peptides such as insulin, which avoids the need for any on-tissue digestion procedures . To the best of our knowledge, this is the first time that an approach is suggested to prioritize the m/z-values driving the embedded results obtained using UMAP.

Experimental Section

Mass Spectrometry Imaging (MSI) data acquisition and processing

MSI was performed on mouse pancreatic tissue. For all samples, MSI was done on a Bruker rapifleX MALDI-TOF mass spectrometer. Cryosections of 7 μm thickness were prepared and mounted on ITO glass slides. Sinapinic acid (SA) was used as matrix and applied using a Bruker ImagePrep. The pixel size was set to 50 μm, and the recorded m/z range was 2-20 kDa in positive linear mode. The acquisition speed was 9 pixels/s with 1000 lasershots/pixel and a laser repetition rate of 10 kHz. The dimensionality associated with the pancreas datasets is 10 606 (Sample 1), 14 791 (Sample 2) and 6937 (Sample 3) pixels by 14 000 m/z bins.

Data processing of Mouse Pancreas datasets

Data modeling and visualization. All data was normalized using total-ion count (TIC). The Python language was used for all data analyses. UMAP was used for dimensionality reduction to three dimensions, followed by hyperspectral visualization. By hyperspectral visualization we mean that based on their position in the obtained 3D embedding, the pixels were translated to RGB color coding by varying the red, green, and blue intensities linearly on the three independent axes, such that the minimum value on an axis is repre-

sented by a color intensity of 0, and the maximum value on an axis has an intensity of 255, which can be normalized to a scale of 0 to 1.[11] The UMAP mapping to three dimensions was performed using the Python implementation (https://github.com/lmcinnes/umap) with the default parameters (n_neighbors=15, gamma=1.0, n_epochs=None, alpha=1.0, init='spectral', spread=1.0, min_dist=0.1, a=None, b=None, random_state=None, metric_kwds=, verbose=True) except for the cosine and Chebyshev distance metric. An elaborate evaluation of the performance of UMAP and different distance metrics can be found in our previous work.[14]

Segmentation of biochemical profiles. Given that the hyperspectral image obtained using UMAP contains three channels (R,G,B), we can apply k-means clustering to this 3D space such that the cluster centroids would represent the dominant colors present in the image. Based on how many biochemical profiles or dominant colors we want to extract from the image we can determine k. For our experiments the Python library OpenCV v4.1.0 was used with k=10..

Spectral information. Based on the biochemical profiles corresponding to certain colors , as selected using the k-means procedure, we determine the location median peak intensities of these profile-specific pixels. The peak finding and prominence measures give insight into the difference between profile-specific peaks in comparison to the other regions of the tissue. For this purpose, the find_peaks and peak _prominences functionality as provided in the Python SciPy package and signal processing library were used.[19] The prominence of a peak measures how much a peak stands out from the surrounding baseline of the signal and is defined as the vertical distance between the peak and its lowest contour line, whereas the find_peaks function finds all local maxima by simple comparison of neighbouring values..

Spatial information. For the biochemical profiles corresponding to certain color gradients, as selected using the k-means procedure, we also construct a binary image for this profile. This binary image is then used to rank all ion or m/z-images according to their Spearman correlation to this specific profile of interest. For this comparison, we first reduce the pixel-space to a two-dimensional embedding with UMAP (default parameters and cosine distance metric) which we then subject to the HDBSCAN clustering algorithm[16] to group similar ion or m/z-images. For the clusters obtained we calculate the average correlation value per cluster to the selected profile and we return the top N ranked ion images per cluster.

HDBSCAN is as the name suggests a hierarchical clustering variant of the DBSCAN algorithm.[16,20] HDBSCAN aims to identify the presence of dense regions by using sliding windows that move towards the high density points. Important advantages of HDBSCAN over other clustering algorithms are i) the ability to identify clusters of data with varying shape ii) robustness to clusters of different densities and iii) the ability to perform density-based clustering, which eliminates the need to specify the number of clusters desired. The only hyperparameter that needs to be specified upfront is the minimum cluster size. This parameter is the primary parameter to effect the resulting clustering and intuitive to interpret since it refers to the smallest size grouping that should be considered as a cluster[16]. Our intention here is to apply clustering such that when correlating a certain profile or pattern of interest to all available ion images, we can identify as many trends as possible by grouping together similar ion images, for which each cluster can still be evaluated in depth if desired. We therefore prefer a small minimal cluster size over a large one. To run the HDBSCAN algorithm we used the scikit-learn implementation (https://github.com/scikit-learn-contrib/hdbscan) using the default parameters and a minimal cluster size of 5.

The combined ranking is obtained by combining the spectral and spatial ranking per tissue by giving a higher priority to those m/z-values that are i) prioritized both in a spatial and

spectral manner within the same tissue and ii) observed in the same profile across multiple tissues

Histogram equalization for contrast enhancement of ion or m/z-images. An ion image represents the mass spectra insensities for a particular m/z- value. These ion images are typically visualized using a pseudo-color scale wherein gradually changing colors are assigned to the intensities. Using this approach can lead to hot spots of pixels with artificially high intensities, distorting the pseudo-color scale such that other pixels will lack contrast. As illustrated in previous work[21] advanced contrast-enhancing procedures like histogram equalization are useful to alleviate this problem. We therefore applied Contrast Limited Adaptive Histogram Equalization (CLAHE)[22] to the ion images prior to visualization for contrast enhancement. To this end we relied on the CLAHE implementation in the scikit-image library for Python (https://scikit-image.org/).[23]

Results and Discussion

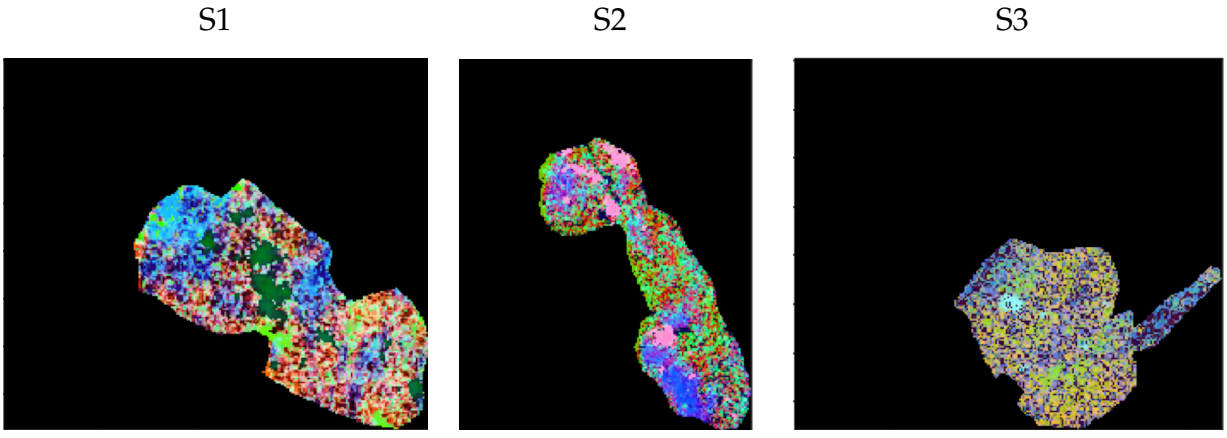

Figure 2: Overview of the hyperspectral visualizations for the three healthy pancreas tissue samples. From left to right are shown samples S1, S2 and S3. The different colors in the image are a representation of each pixel's location within the embedded space. As such similar colors or RGB values represent similar spectra or biochemical patterns within one sample. This is however not necessarily the case for different tissues hence the green profile in S1 is completely independent from the green profile in S2 or S3.

Spatial and spectral prioritization based on the hyperspectral visualization obtained using UMAP enables quality control and molecular insights

The hyperspectral visualizations for these samples are the starting point of our analysis and are shown in figure 2. To illustrate our approach, we show the results for two segmentations obtained from these images. The first example, as shown in figure 3, correlates with endocrine tissue more specifically, the anatomical location of the pancreatic islets or the Islets of Langerhans. The (binary) segmentations in panels D-F correspond to the pink, dark green and light blue regions of the hyperspectral visualizations of samples 1,2 and 3 respectively. These regions clearly overlap with the pancreatic islets, which are annotated in the H&E stainings shown in panels A-C. The combined rankings are presented below each segmentation. These rankings are obtained by combining the spectral and spatial ranking per tissue by giving a higher priority to those m/z-values that are i) prioritized both in a spatial and spectral manner within the same tissue and ii) observed in the same profile across multiple tissues. The top three m/z-values are the same for all samples, while the top four is completely the same for samples 1 and 3. In all tissues the highest ranked ion is the m/z- value around 5800 Da, which likely corresponds to the full-length active insulin molecule after cleavage of the c-peptide from pro-insulin. This peak with a major abundance around 5800 Da, is observed to lie within the anticipated m/z-deviations, as can be expected from MALDI measurements, across all tissues. Moreover, the region-specific localization that corresponds to the islets of Langerhans, as shown also in the H&E stainings, further strenghtens this association (figures 3 and 4). For the values around 5840 and 5821 Da a putative association could be made to ATP5E (ATP synthase subunit epsilon) and VIP (Vasoactive Intestinal Peptide) respectively. Previous research regarding the mouse pancreatic islet proteome has highlighted the enrichment of proteins that play a role in oxidative phosphorylation[24] which might explain the detection of ATP5E. In

addition, the stimulatory effect of VIP on insulin has previously been reported.[25,26] The ubiquitously distributed m/z- value of 6276.22 detected in the third sample is likely to correspond to a part of the mouse kallikrein protein. This protein is known to play a role in the pancreas metabolism and has been observed both in the endocrine and exocrine pancreas. .

Another key-player involved in glucose metabolism is glucagon which could putatively correspond to the m/z values around 3484. This molecule is observed in all samples with a distribution around the pancreatic islets (Figure S27) and ranked around position 15 for the three tissues which can be well explained given its lower abundance. This indicates that the results are likely to have a biological meaning in addition to the confirmed spatial co-localization as shown in the corresponding ion images. Moreover using this method a good and robust correspondence between tissues is observed as illustrated by the combined ranking. The particular ion images associated with the top three m/z-values for sample 1 are shown in figure 4 and are available for all samples in the supplementary information (S1-S25). Figure 4 shows additional confirmation regarding the spatial co-localization of the identified ions with the dark green segment, which is also supported by the overlay between the H&E and hyperspectral visualization. For a larger representation of the H&E stainings we refer to the supplementary information (figures S28-S30).

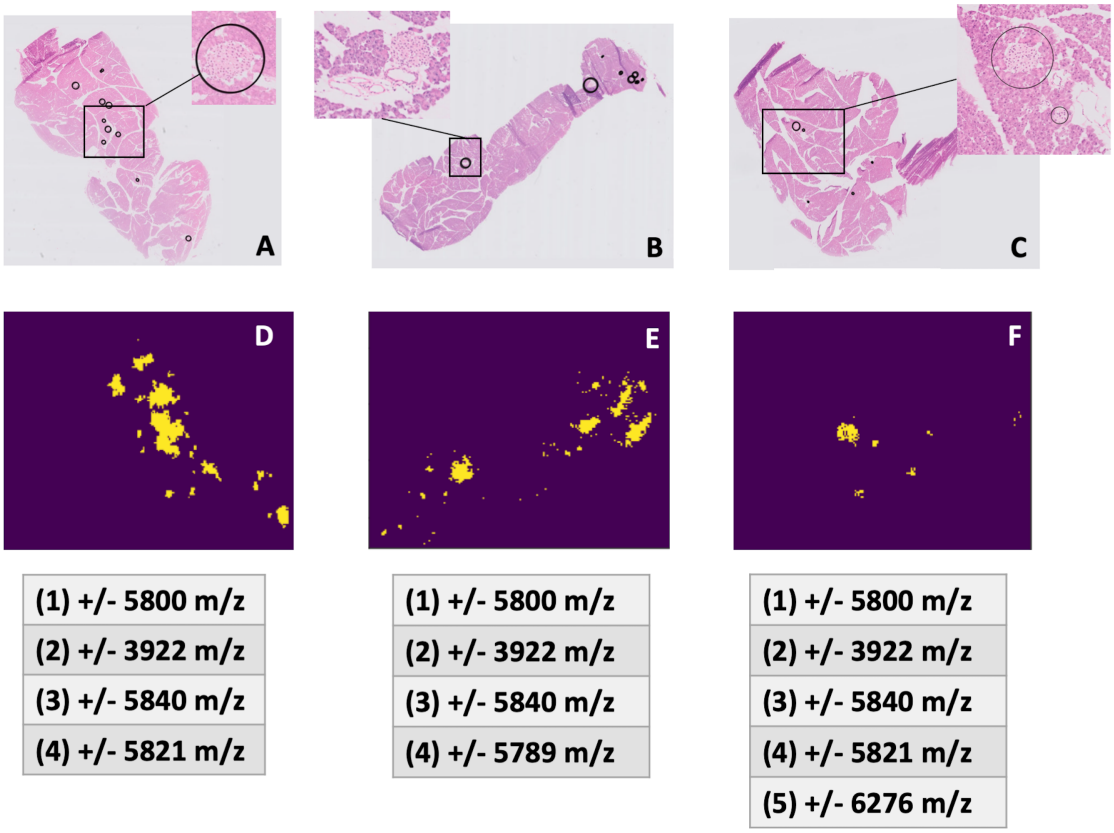

Figure 3: In panels, D-F, a segmented profile associated with samples 1,2, and 3 respectively, is shown. These (binary) segmentations correspond to the pink, dark green and light blue regions of the associated hyperspectral visualizations in figure 2. This segmentation correlates with endocrine tissue, more precisely the pancreatic islets, as indicated in the H&E stainings (panels A-C). The rankings presented below each segmentation, are obtained by combining the spectral and spatial ranking per tissue by giving a higher priority to those m/z-values that are i) prioritized both in a spatial and spectral manner within the same tissue and ii) observed in the same profile across multiple tissues. The close similarity of the prioritized m/z-values shows the value of this method to obtain intra-and inter-sample knowledge regarding the molecular composition supported by the visualizations obtained using UMAP.

For the second segmentation, which corresponds to the light and dark blue regions of the hyperspectral visualizations, we observe a similar prioritization across the different tissues corresponding to exocrine pancreas. As shown in figure 5, the top three of m/z-values is almost identical for the different tissue samples.. The assumption that this profile represents exocrine tissue is supported by the fact that the m/z -value around 6651 Da has been associated with the acini in earlier work.[27] Moreover, the MASCOT engine identifies

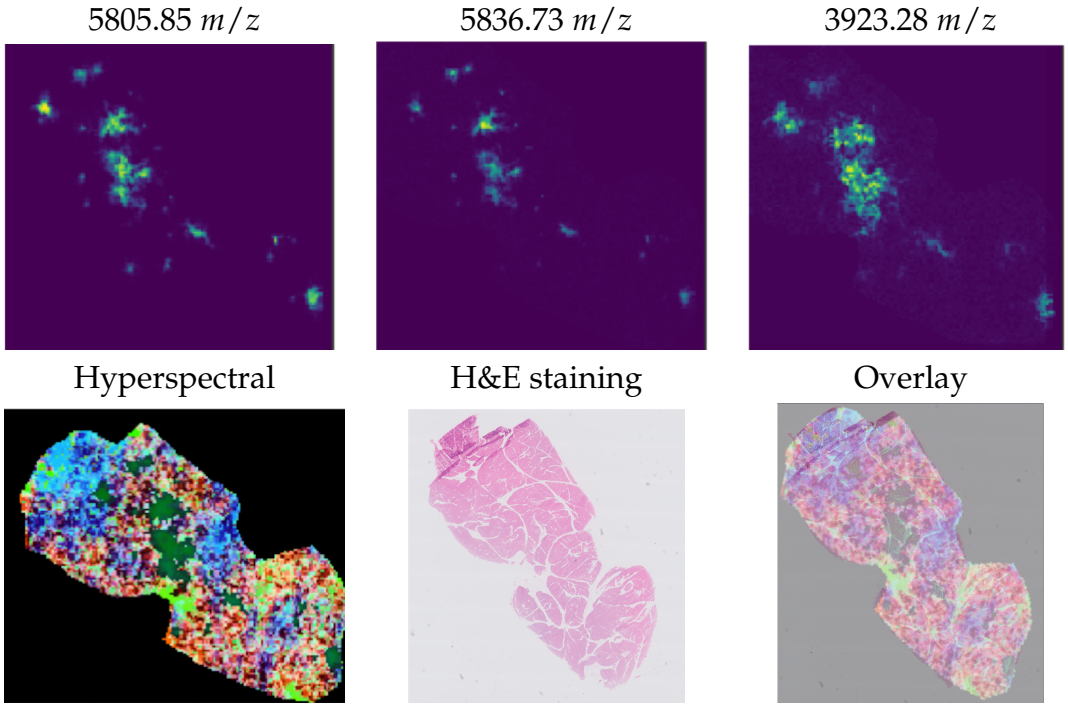

Figure 4: The m/z-value of 5805.85 Da, shown in panel A, likely corresponds to the full-length active insulin molecule after cleavage of the c-peptide from pro-insulin. For m/z 5821.29 Da, a putative association could be suggested to ATP5E (ATP synthase subunit epsilon), which plays a role in the oxidative phosphorylation process and VIP (Vasoactive Intestinal Peptide), a known player regarding insulin metabolism, is suggested to be associated with the m/z-value of 5836.73 Da. All prioritized ion images show a clear association with endocrine tissue and more specifically a distribution around the pancreatic islets (figure 2).

a match with Selenocysteine lyase (6643 Da), a protein associated with high expression in the exocrine pancreas according to the Human Protein Atlas (HPA). For the value around 6691 Da, we could identify a putative association to Single-pass membrane protein with coiled-coil domains 4 (SMCO4), a protein for which also a high expression in the exocrine pancreas is noted in the HPA. In addition to the biological relationship, the ion images also show a similar spatial distribution to the segmentation obtained from the hyperspectral visualizations, as shown in figure 6 for sample 1 and in the supplementary information for the other samples (S1-S25).

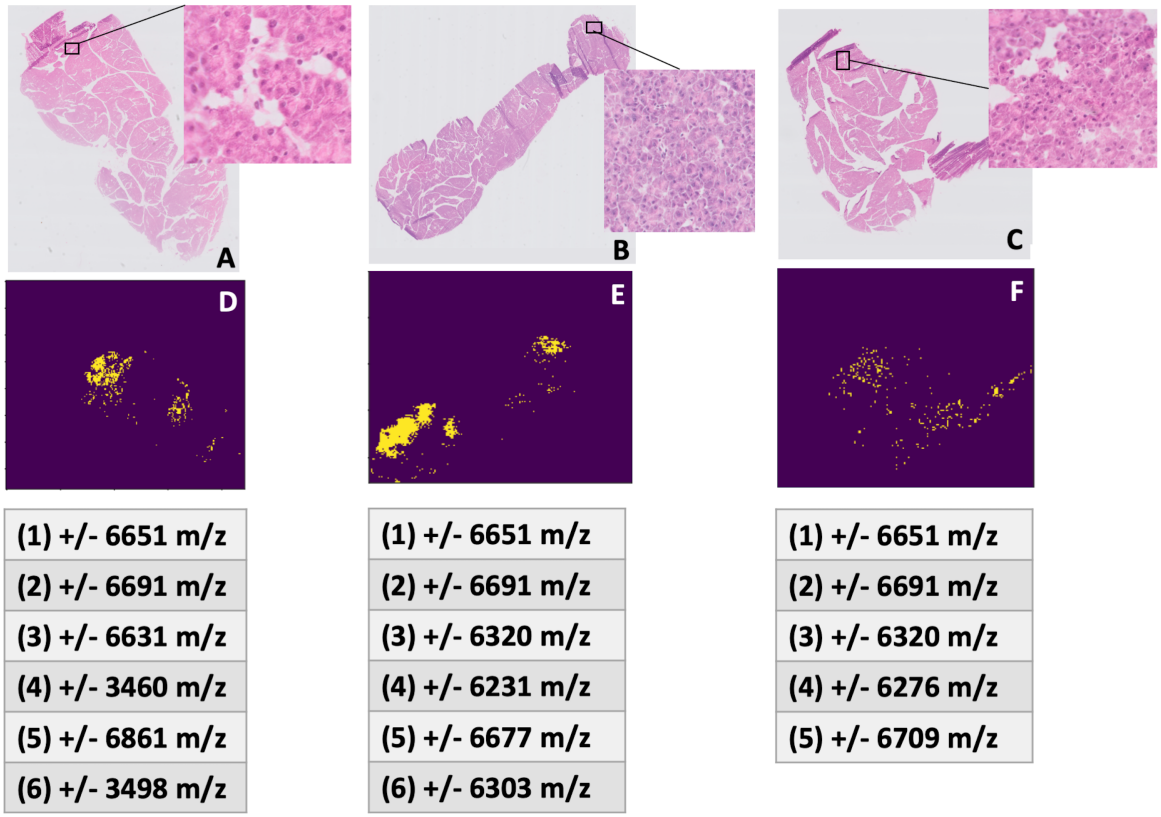

Figure 5: In panels D-F, a segmented profile associated with samples 1,2, and 3 respectively is shown. This profile corresponds to exocrine tissue as indicated in the H&E stainings. The rankings are obtained by combining the spectral and spatial ranking per tissue by giving a higher priority to those m/z-values that are i) prioritized both in a spatial and spectral manner within the same tissue and ii) observed in the same profile across multiple tissues. The close similarity of the prioritized m/z-values shows the value of this method to obtain intra-and inter-sample knowledge regarding the molecular composition supported by the visualizations obtained using UMAP. The particular ion images for sample 1 are shown in figure 5 and are available for all samples in the supplementary information (S1-S25).

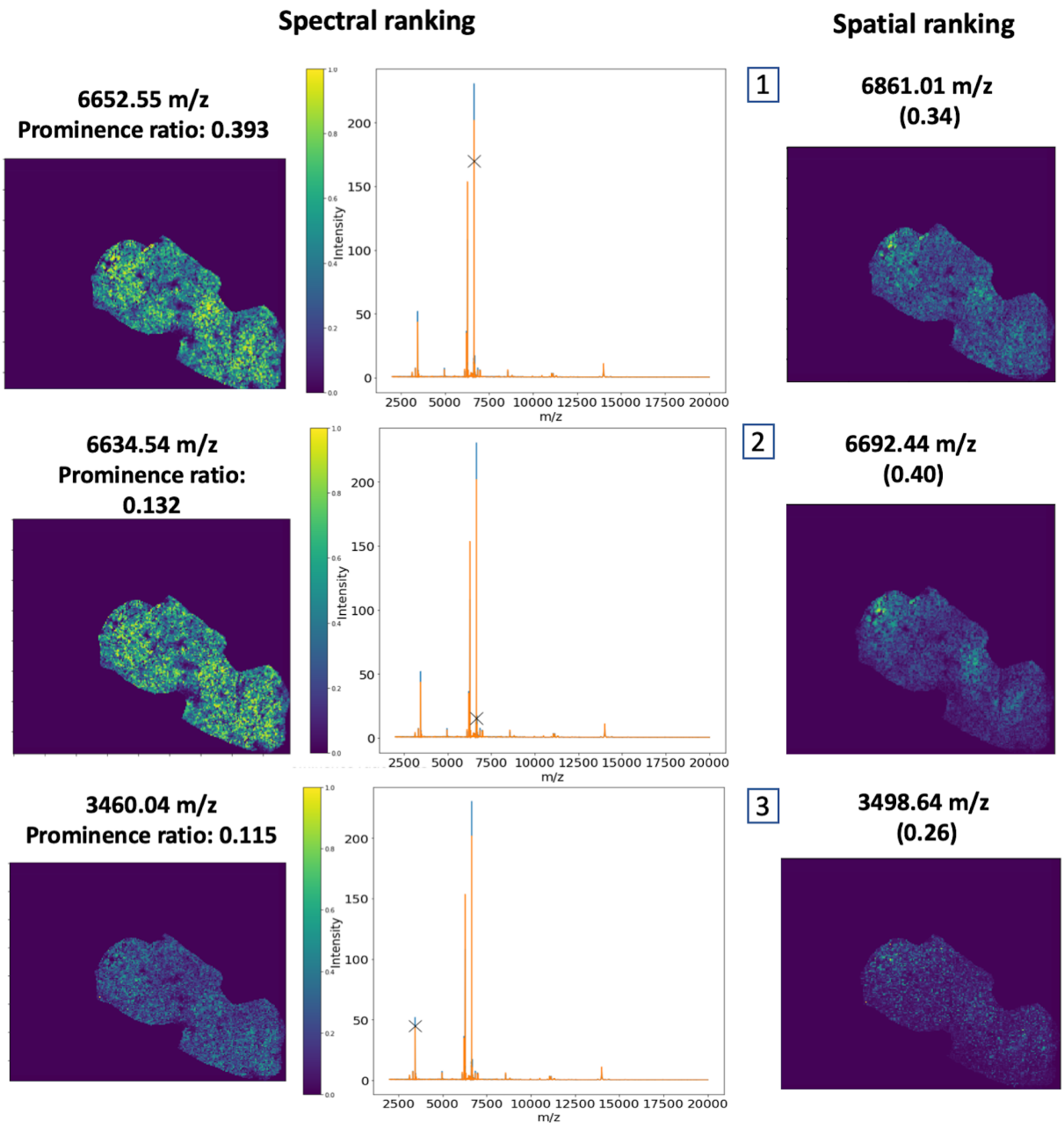

Figure 6: Top three of spectral (left) and spatial (right) results ranked vertically from top to bottom. For the spectral results the particular ion image corresponding to the highlighted peak is shown together with the median spectra intensities associated with the extracted profile (in blue) against the median overall spectra intensities (orange). For the spatial results the particular ion images are shown together with their correlation values for the highest ranked ion image per cluster which were also ranked according to their highest average correlation to the extracted profile. A clear association is visible to the extracted pattern which is assumed to be exocrine tissue. The m/z- value around 6650 Da has previously been associated with the acini.

These kind of observations can be very insightful, not only to study similar molecular

compositions across different samples, but also to assess the quality of the hyperspectral visualizations obtained using UMAP. An example of the latter use case is shown in figures 7-8 where instead of the cosine distance metric, the Chebyshev distance metric was used for constructing the embeddings. In grey, we can discern an additional profile within the blue region . Because we are not able to detect any ion images corresponding to this grey profile,, we can assume that the cosine metric achieves a better representation of the underlying data. Moreover using this approach one can guide the level of detail that is desired in two ways i) by selecting the number of colors one wants to extract and ii) by selecting the number of ion images to be returned per cluster. This is not only important to give us more confidence in the visualizations obtained using UMAP but can also strongly support us in finding biological meaning, in particular when dealing with heterogeneous tissue samples..

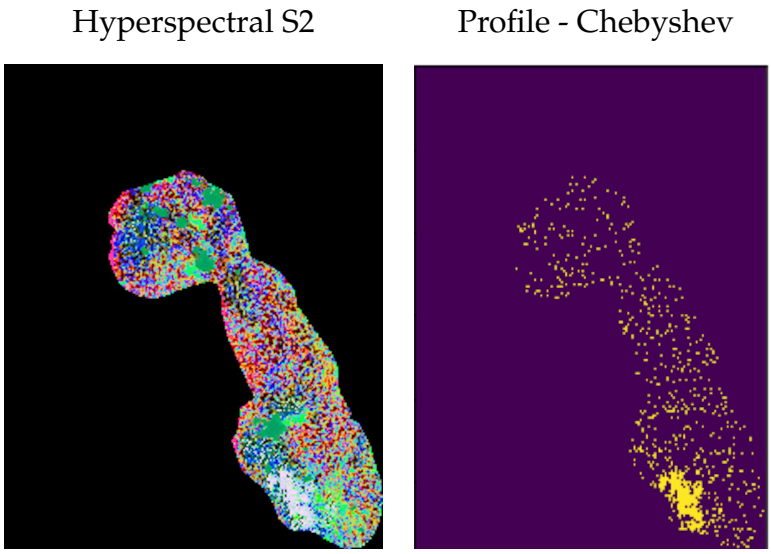

Figure 7: On the left the hyperspectral visualization for sample 2 obtained from a healthy mouse pancreas is shown with on the right the extracted profile corresponding to the grey region. This hyperspectral visualization was obtained by using the Chebyshev distance metric instead of the cosine distance metric.

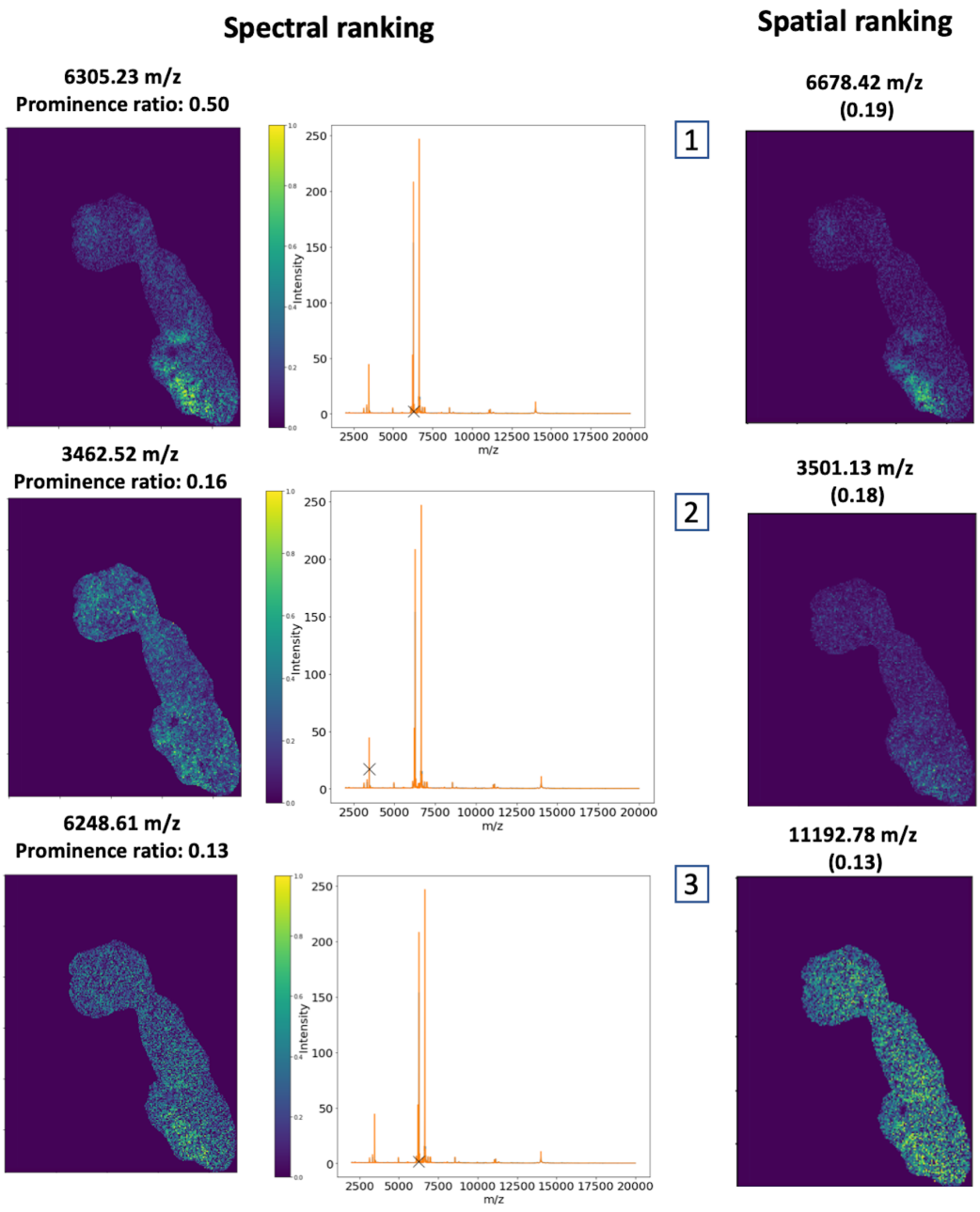

Figure 8: Top three of spectral (left) and spatial (right) results ranked vertically from top to bottom. For the spectral results the particular ion image corresponding to the highlighted peak is shown together with the median spectra intensities associated with the extracted profile (in blue) against the median overall spectra intensities (orange). For the spatial results the particular ion images are shown together with their correlation values for the highest ranked ion image per cluster which were also ranked according to their highest average correlation to the extracted profile. In comparison to the results obtained using the cosine distance metric, there is an additional grey region present for which we extracted the corresponding profile shown on the left. We are not able to detect this region in the spectral and spatial results. We can therefore assume that the observation of this region is probably a consequence of the distance metric rather than a reflection of the underlying data.

To the best of our knowledge it is the first time this approach is applied to the results obtained using UMAP. This approach combines the strong visualization and dimensionality reduction performance of UMAP with identifying m/z-values for a specific spatial region. In earlier work, the limitation of PCA in this regard has been highlighted. In particular it was noted that ion images found using PCA differ sometimes from the corresponding score images of the first principal components. At the same time, the added value of spatial segmentation maps with regard to clustering results to identify co-localized m/z- values was noted.[21,28] Pancreatic tissue samples have the advantage that various endogenous peptides with key metabolic functions are within the mass range of MALDI-MS imaging experiments. For insulin, the proposed identification is considered realistic because of its high abundance and the striking match with the localization of the pancreatic islets. This is a fine example underscoring the importance to consider both spatial information and spectral information in mass spectrometry imaging. The identity of the other co-localized m/z-values is highly speculative and more advanced identification tools such as immunohistochemistry are essential to further strengthen the proposed identifications, but the fact that we can obtain a robust ranking across tissues shows the value of our method for explorative analysis. Furthermore, in our earlier work, we have shown the enrichment of UMAP in comparison to methods such as for example PCA because it can take the non-linear nature of biological phenomena into account. Moreover, methods such as UMAP are able to capture the complete feature space into two or three dimensions, an attribute that other methods such as PCA and clustering methods lack and which facilitates strong data visualizations.[14] Interestingly, UMAP already alleviated a big limitation inherent to non-linear dimensionality reduction methods being their computational speed. With improvements at this level now being available through GPU support for UMAP in the NVIDIA RAPIDS implementation, we expect the application of UMAP in the analysis of large datasets such as MSI data to take even more importance.[29]

Conclusion

We have shown that it is possible to obtain insight into the molecular patterns that are discernible in the hyperspectral visualizations obtained using UMAP. Starting from a bidirectional dimensionality reduction we are able to rank m/z- values according to their profile taking into account both the available spatial and spectral information. We believe that this is a valuable approach not only to gain biological insight, but also to support the results obtained using UMAP. The recent inclusion of the UMAP algorithm into the RAPIDS library and the associated GPU acceleration makes this approach even more attractive due to fast turn-around times. Moreover combining insights based on strong visualizations and taking into account both the spectral and spatial information will enable us to construct a more comprehensive picture of the underlying biological phenomena at play.

References

(1) Caprioli, R. M.; Farmer, T. B.; Gile, J. *Analytical Chemistry* **1997**, *69*, 4751–4760, PMID: 9406525.

(2) Römpp, A.; Spengler, B. *Histochemistry and Cell Biology* **2013**, *139*, 759–783.

(3) Verbeeck, N.; Caprioli, R. M.; Van de Plas, R. *Mass spectrometry reviews*

(4) Ma, S.; Dai, Y. *Briefings in Bioinformatics* **2011**, *12*, 714–722.

(5) Siy, P. W.; Moffitt, R. A.; Parry, R. M.; Chen, Y.; Liu, Y.; Sullards, M. C.; Merrill, A. H.; Wang, M. D. Matrix factorization techniques for analysis of imaging mass spectrometry data. Proceedings of the 8th IEEE International Conference on Bioinformatics and Bioengineering, BIBE 2008, October 8-10, 2008, Athens, Greece. 2008; pp 1–6.

(6) Trindade, G.; Abel, M.-L.; Watts, J. *Chemometrics and Intelligent Laboratory Systems* **2017**, *163*, 76–85.

(7) Hanselmann, M.; Kirchner, M.; Renard, B. Y.; Amstalden, E. R.; Glunde, K.; Heeren, R. M. A.; Hamprecht, F. A. *Analytical Chemistry* **2008**, *80*, 9649–9658, PMID: 18989936.

(8) Kohonen, T. *Proceedings of the IEEE* **1990**, *78*, 1464–1480.

(9) Gorzolka, K.; Kölling, J.; Nattkemper, T. W.; Niehaus, K. *PLoS One* **2016**, *11*.

(10) Kölling, J.; Langenkämper, D.; Abouna, S.; Khan, M.; Nattkemper, T. W. *Bioinformatics* **2012**, *28*, 1143–1150.

(11) 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. *Analytical Chemistry* **2013**, *85*, 1415–1423, PMID: 23249247.

(12) van der Maaten, L.; Hinton, G. *Journal of Machine Learning Research* **2008**, *9*, 2579–2605.

(13) McInnes, L.; Healy, J. *ArXiv e-prints* **2018**,

(14) Smets, T.; Verbeeck, N.; Claesen, M.; Asperger, A.; Griffioen, G.; Tousseyn, T.; Waelput, W.; Waelkens, E.; De Moor, B. *Analytical chemistry* **2019**,

(15) Minerva, L.; Ceulemans, A.; Baggerman, G.; Arckens, L. *Proteomics. Clinical applications* **2012**, *6*.

(16) McInnes, L.; Healy, J.; Astels, S. *The Journal of Open Source Software* **2017**, *2*.

(17) Buza, E.; Akagic, A.; Omanovic, S. Skin detection based on image color segmentation with histogram and K-means clustering. 2017 10th International Conference on Electrical and Electronics Engineering (ELECO). 2017; pp 1181–1186.

(18) Itseez, Open Source Computer Vision Library. 2015; `https://github.com/itseez/opencv`.

(19) Jones E, P. P., Oliphant E SciPy: Open Source Scientific Tools for Python. 2001; http://www.scipy.org/.

(20) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-based Algorithm for Discovering Clusters 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.

(21) Alexandrov, T. *BMC Bioinformatics* **2012**, *13*, S11.

(22) Zuiderveld, K. In *Graphics Gems IV*; Heckbert, P. S., Ed.; Academic Press Professional, Inc.: San Diego, CA, USA, 1994; Chapter Contrast Limited Adaptive Histogram Equalization, pp 474–485.

(23) van der Walt, S.; Schönberger, J. L.; Nunez-Iglesias, J.; Boulogne, F.; Warner, J. D.; Yager, N.; Gouillart, E.; Yu, T.; the scikit-image contributors, *PeerJ* **2014**, *2*, e453.

(24) Petyuk, V. A.; Qian, W.-J.; Hinault, C.; Gritsenko, M. A.; Singhal, M.; Monroe, M. E.; Camp, D. G.; Kulkarni, R. N.; Smith, R. D. *Journal of proteome research* **2008**, *7*, 3114–3126.

(25) Ahrén, B.; Wierup, N.; Sundler, F. *Diabetes* **2006**, *55*, S98–S107.

(26) Bishop, A.; Polak, J.; Green, I.; Bryant, M.; Bloom, S. *Diabetologia* **1980**, *18*, 73–78.

(27) Minerva, L.; Clerens, S.; Baggerman, G.; Arckens, L. *PROTEOMICS* **2008**, *8*, 3763–3774.

(28) Deininger, S.-O.; Becker, M.; Suckau, D. *Methods in molecular biology (Clifton, N.J.)* **2010**, *656*, 385–403.

(29) NVIDIA, Open Source GPU Data Science, RAPIDS. 2019; https://github.com/rapidsai.

Graphical TOC Entry

