Hierarchical bi-clustering of mouse pancreas mass spectrometry imaging data using recursive rank-2 non-negative matrix factorization

Melanie Nijs*,†, Etienne Waelkens‡, 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, 3000 Leuven, Belgium

§Fellow IEEE, SIAM, IFAC

ABSTRACT: One of the main challenges in mass spectrometry imaging (MSI) data analysis remains the analysis of m/z-spectra displaying low signal-to-noise ratio (SNR) caused by their low abundance, sample preparation, matrix effects, fragmentation, and other artifacts. Additionally, we observe that molecules with a high abundance suppress those with lower intensities and misdirect classical tools for MSI data analysis, such as principal component analysis (PCA). As a result, the observed significance of a molecule may not always be directly related to its abundance. In this work, we present a recursive rank-2 non-negative matrix factorization (rr2-NMF) algorithm that automatically returns spectral and spatial visualization of co-localized molecules, both highly and lowly abundant. Using this hierarchical decomposition, our method finds spatial and spectral correlations on different levels of abundances. The quality of the analysis is evaluated on MALDI-TOF data of healthy mouse pancreatic tissue for the annotation of molecules of interest in the lower abundances. The results show interesting findings regarding the functioning and co-localization of certain molecules.

INTRODUCTION

Mass spectrometry imaging (MSI) has become an established tool for the untargeted analysis of thousands of molecules - including lipids, proteins, and metabolites - in a spatial context.[1] With the advantage of measuring all these molecules simultaneously, MSI allows for the exploratory analysis of tissue sections and aims to uncover novel biomarkers.[2]

While MSI provides a detailed molecular representation of the sample, the large number of unknown peaks imposes great challenges towards the full understanding of the biochemical processes taking place. In addition, analysis of MSI data is further complicated by phenomena as ion suppression and noise artifacts.

There are multiple factors causing noise artifacts on the true signal, e.g.: sample preparation, matrix effects, biological noise, refraction of molecules, adducts, isotopes, and so on.[3,4] To minimize artefacts resulting from sample preparation, the MSI field is working towards more standardized procedures.[5] Unfortunately, the other sources of noise are inherent to the data and its technology. This underscores the importance of data scientists to develop software which interprets these signals correctly by handling the noise and artefacts in a rigorous manner. For analysis of MSI data many approaches exist, depending on the problem at hand. This could be either exploratory and unsupervised, by clustering similar spectra and visualize the tissue accordingly, or supervised, using annotated data to classify, segment or predict MSI data in a more specific context.[6,7]

Not only noise artifacts but also high abundance of particular molecules may result in the fact that the intensity of a measured m/z-value does not always directly correlate to the value's importance. As a consequence, traditional data analytical approaches, such as principal component analysis (PCA), may be biased by the most dominant values and overlook nuances, or perhaps characterizing m/z-values, within the smaller intensities.

Hierarchical analysis of MSI data can overcome both challenges by grouping the m/z-values, split the data accordingly and repeat the analysis within each subgroup with more nuances. Methods such as hierarchical stochastic neighbor embedding (HSNE) and hierarchical density-based spatial clustering for applications with noise (HDBSCAN) have demonstrated the use of the hierarchical approach for MSI data, but are employed on just a sole dimension rather than both spectral and spatial.[8,9]

Our work focusses on the unsupervised clustering of matrix assisted laser desorption ionization time-of-flight (MALDI-TOF) MSI data for the annotation and identification of co-localized molecules exhibiting spatial and spectral correlations. We hereto present a recursive rank-2 non-negative matrix factorization (rr2-NMF) algorithm. By considering both the spectral and spatial dimension for the clustering, this algorithm can be seen as an unsupervised bi-clustering[10] method. Although bi-clustering has been demonstrated successfully on MSI data in earlier works[11,12], a recursive hierarchical approach based on non-negative matrix factorization (NMF) and tailored to MSI data has, as far as the author's knowledge, not yet been done.

The algorithm recursively performs a rank-2 NMF decomposition and extracts two subsets in the original space based on the spectral and spatial separation in these components. Depending on the level of hierarchy, this process is repeated for each subset and will automatically stop either when (1) a predefined maximum number of levels α is reached, or (2) the cosine similarity for all pixels within the cluster converges, i.e. the increment in similarity of the next subdivision is below a threshold β.

To evaluate its performance, the algorithm has been applied on three mouse pancreas MALDI-TOF datasets. Studying MSI data of the pancreas will lead to a better understanding of its normal functioning as well as insights in related diseases, such as diabetes and pancreatic cancer.[13–15] To this end, a detailed analysis of all samples has been conducted. The contributing m/z-values per cluster are inspected and mapped to the corresponding ion images. This shows novel information about the samples compared to earlier work[9,16,17] on the same three pancreas data sets. Finally, we investigate computational efficiency and provide an accompanying GitHub tutorial (including a multiprocessing option) for other researchers in the field who wish to apply rr2-NMF to their MSI data.

METHODS

MALDI data acquisition and preprocessing

For the evaluation of the method, we analyze three (M1, M2, and M3) healthy mouse pancreas samples obtained with a Bruker RaplifleX MALDI-TOF, as described previously by Smets et al. (2019).[18]

All three datasets have been preprocessed with a total ion current (TIC) normalization using the Python language, which was also used for all further coding. To demonstrate its robustness, no other preprocessing, e.g. peak-picking, has taken place before application of the rr2-NMF algorithm.

For the evaluation of the results, we compare the spatial clusters with the corresponding ion images. To prevent distortion of the ion images, resulting from artifical hot-spots caused by artifacts[19], we apply the contrast enhancement method CLAHE (Contrast Limited Adaptive Histogram Equalization), as provided by https://github.com/anntzer/clahe.

A recursive rank-2 non-negative matrix factorization (rr2-NMF) algorithm for hierarchical bi-clustering of MSI data

Bi-clustering is defined as the simultaneous clustering of rows and columns, towards so-called bi-clusters, based on differentiating characteristics in both dimensions of the dataset.[20] In this work, we demonstrate a bi-clustering approach based on a hierarchical application of NMF. The key concept of NMF is to decompose a data set into its rank r main contributing factors. In a linearly additive manner, NMF aims to split the data such that datapoints are contributing as much as possible to a single component.

Mathematically, we state that NMF factorizes a data matrix $\boldsymbol{X} \in R_+^{M \times N}$, with M rows, the m/z-values, and N columns, the (vectorized) pixels, in two non-negative factor matrices: the spectral basis elements $\boldsymbol{W} \in R_+^{M \times K}$ and the weight matrix $\boldsymbol{H} \in R_+^{K \times N}$. In short, we write:

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}.$$

The solution is approximated by iteratively minimizing the Frobenius norm or the Kullback-Leibler divergence, depending on the expected distribution of the data, respectively Gaussian or Poisson. Known algorithms to do so are multiplicative update (MU) rules or alternating least squares (ALS). For more details on the theory and applications of NMF, we refer the reader to previous work and the book by N. Gillis.[17,21]

The usefulness of NMF and its related methods for the analysis of MSI data has been investigated in earlier work[17], where NMF is used for dimensionality reduction and visualization of the main trends in the samples. In continuation of that research, we have now tailored the NMF approach further for the specific analysis of MSI data exhibiting informative peaks in the lower abundances. To this end, we present a recursive rank-2 NMF algorithm.

Shown in pseudocode below, rr2-NMF starts with a rank-2 NMF on the full dataset. Based on the obtained spectral and spatial information of the two components, we sub-select from the original data two new datasets of co-localized peaks. This repeats on each subset, until either: (Algorithm 1) a predefined depth α is reached, or (Algorithm 2) the sub-cluster cosine similarity improves not more than a threshold β. The user can, based on the level of detail they want to obtain, select the most suited approach and the value for parameter α or β. No other user interaction is required. The generated output is a detailed descriptor of each subcluster, consisting of three elements: (1) the spectral distribution, (2) the spatial distribution, and (3) a table containing the m/z-values and intensities of the top n contributing elements. With this information, a more detailed analysis on the found metabolites can be done for the tissue under investigation.

Implementation details and code availability

All implementation has been done using Python and the scikit-learn[22] package for the NMF decomposition. Considering the ever-increasing quality of MSI data and the consequentially ever-increasing size of the data, we provide the option to run the code either sequential or in parallel using Python's multiprocessing library. Code will be made accessible on https://github.com/MelanieN/rr2-NMF.

Algorithm 1 Recursive rank-2 NMF with fixed depth stopping criterion

```
Define recursion depth α
function RR2-NMF(data, level)
    if level+1 = α then
        W, H ← NMF(data, rank = 2)
        PLOT_COMPONENTS(W,H)
    else
        W, H ← NMF(data, rank = 2)
        subset1 ← data(W1 > W2, H1 > H2)
        subset2 ← data(W1 <= W2, H1 <= H2)
        RR2-NMF(subset1, level+1)
        RR2-NMF(subset2, level+1)
    end if
end function
```

Algorithm 2 Recursive rank-2 NMF with cluster convergence stopping criterion

```
Define similarity increment threshold β
function RR2-NMF(data)
    W, H ← NMF(data, rank = 2)
    subset1 ← data(W1 > W2, H1 > H2)
    subset2 ← data(W1 <= W2, H1 <= H2)
    if SIMILARITY(data) − SIMILARITY(subset1) < β then
        PLOT_RESULT(subset1)
    else
        RR2-NMF(subset1)
    end if
    if SIMILARITY(data) − SIMILARITY(subset2) < β then
        PLOT_RESULT(subset2)
    else
        RR2-NMF(subset2)
    end if
end function
```

RESULTS AND DISCUSSION

Evaluation of rr2-NMF on three healthy mouse pancreas samples

To demonstrate the performance of rr2-NMF, a detailed analysis has been conducted on three healthy mouse pancreatic samples (M1, M2, and M3). In Figure 1, a graphical representation of the algorithm is shown: rr2-NMF applied on the M2 dataset with as a stopping criterion depth of two layers. Here, we see step-by-step how rr2-NMF iteratively splits the dataset based on similarities in both the spatial (upper graphic) and spectral (lower graphic) components. A similar overview for the alternate choice of the stopping criterion (the convergence of the cosine similarity) is shown in Figure 2. Immediately, rr2-NMF shows clear spatial regions with their corresponding molecular fingerprints – a valuable benefit compared to UMAP or Mapper. More specifically, we see low-intensity values, such as m/z = 3485.68 Da (glucagon) gaining more weight in these subclusters as opposed to non-hierarchical approaches. The same is the case for samples M1 and M3 (see Supporting Information S2 and S4).

The method confirms itself in a way that expected artefacts of one signal are clustered together, e.g.: multiple charge states, and adducts such as sodium and potassium. In Table 1, we list the m/z-peaks of cluster *"M2_convergence_2_1"*. Here we see the value corresponding to Ins2 (m/z = 5800.77 Da), with the two-charged state (m/z = 2901.45 Da), as well as the added potassium (m/z = 5838.09 Da), and both potassium and sodium (m/z = 5857.4 Da).

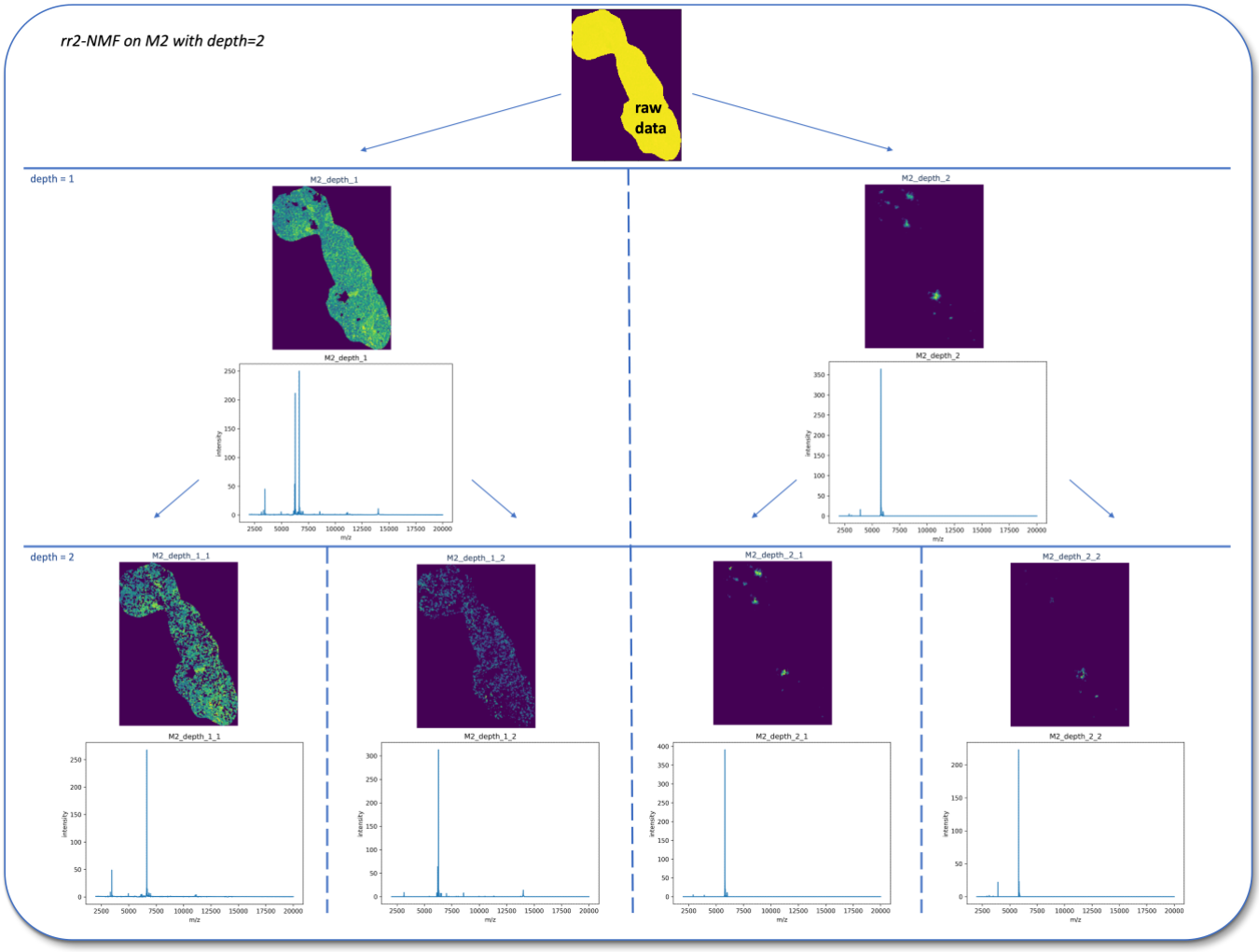

Figure 1. Graphical overview of the rr2-NMF algorithm of the M2 pancreatic tissue sample, simultaneous clustering of the spectral and spatial dimensions. Rr2-NMF iteratively splits the dataset based on similarities in both the spatial (upper graphic) and spectral (lower graphic) components. Here, we used as a stopping criterion: a fixed depth α of 2 levels. High resolution images are available in the Supporting Information (S3a).

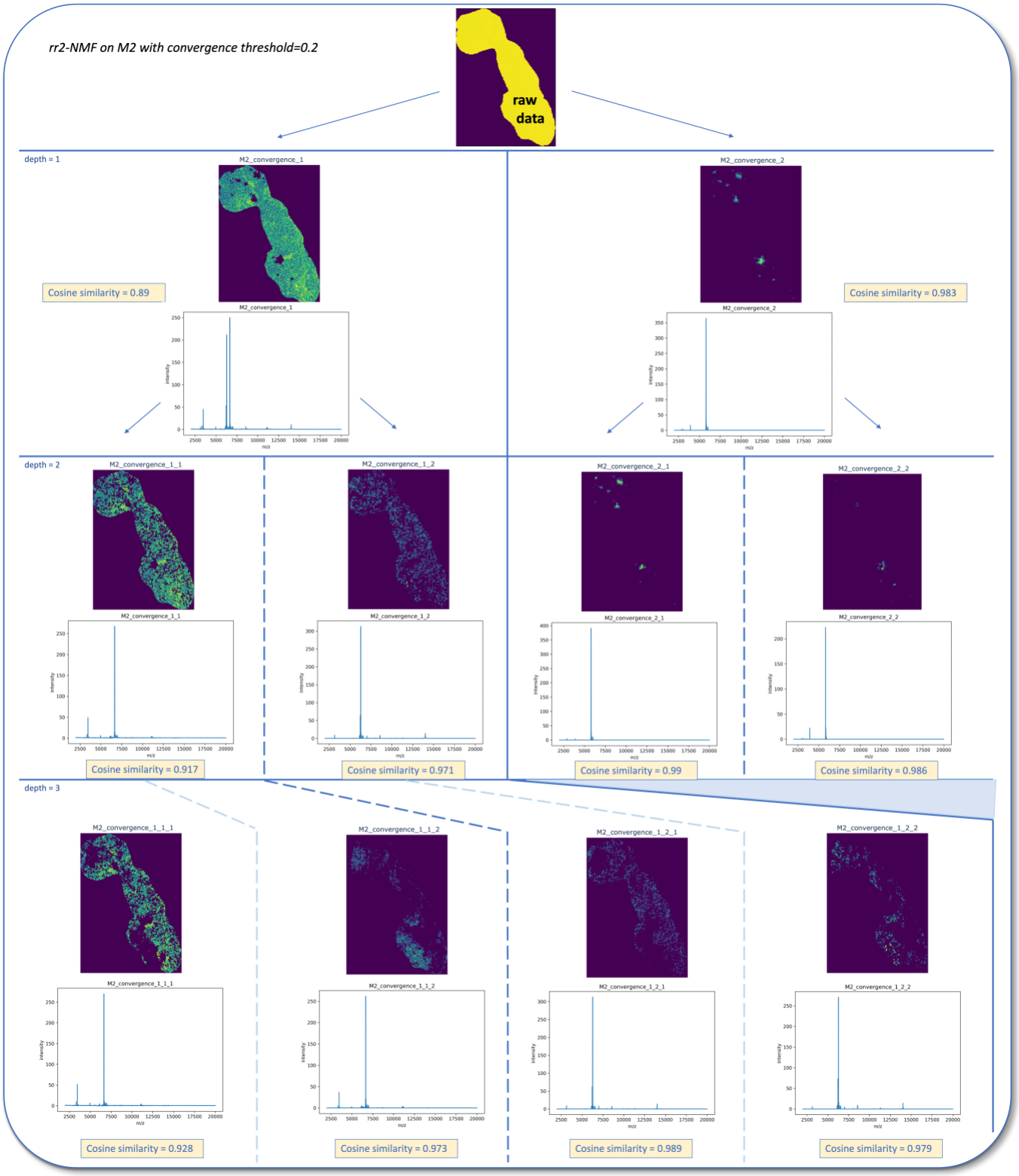

Figure 2. Graphical overview of the rr2-NMF algorithm of the M2 pancreatic tissue sample, simultaneous clustering of the spectral and spatial dimensions, now with stopping criterion: the convergence (increment $\beta < 0.02$) of the cosine similarity of the spectra within each cluster. High resolution images are available in the Supporting Information (S3b).

A closer look at the islets of Langerhans in mice pancreas

Using rr2-NMF, an interesting pattern arises in the islets of Langerhans. The islets play a vital role in our bodies as it is the mechanism monitoring our glucose levels: it produces insulin and releases it in our bloodstream when needed. The study of the islets, and the pancreas in general, has led to a better understanding of common diseases, such as diabetes.

Table 1. List of the m/z-values corresponding to the peaks of cluster *"M2_convergence_2_1".*

m/z-value [Da]	intensity	(possible) annotation
5800.77	391.0	$Ins2^{+}$
5814.93	43.0	Ins2 + **14.16**
5838.09	20.0	Ins2 + K^{+}
5782.76	11.0	Ins2 – H_2O
6005.39	11.0	
5857.4	10.0	
6023.4	8.0	
2901.45	5.0	$Ins2^{2+}$
3920.65	4.0	
5755.73	3.0	
5896.0	3.0	
5910.16	1.0	
5977.08	1.0	

Table 2. List of the m/z-values corresponding to the peaks of cluster *"M2_convergence_2_2".*

m/z-value [Da]	intensity	(possible) annotation
5808.5	223.0	$Ins1^{+}$
5843.24	23.0	
3924.51	22.0	
5822.65	17.0	Ins1 + **14.15**
5863.83	7.0	
5898.58	3.0	
3122.79	2.0	
5916.59	2.0	
2906.59	1.0	$Ins1^{2+}$
3485.69	1.0	glucagon

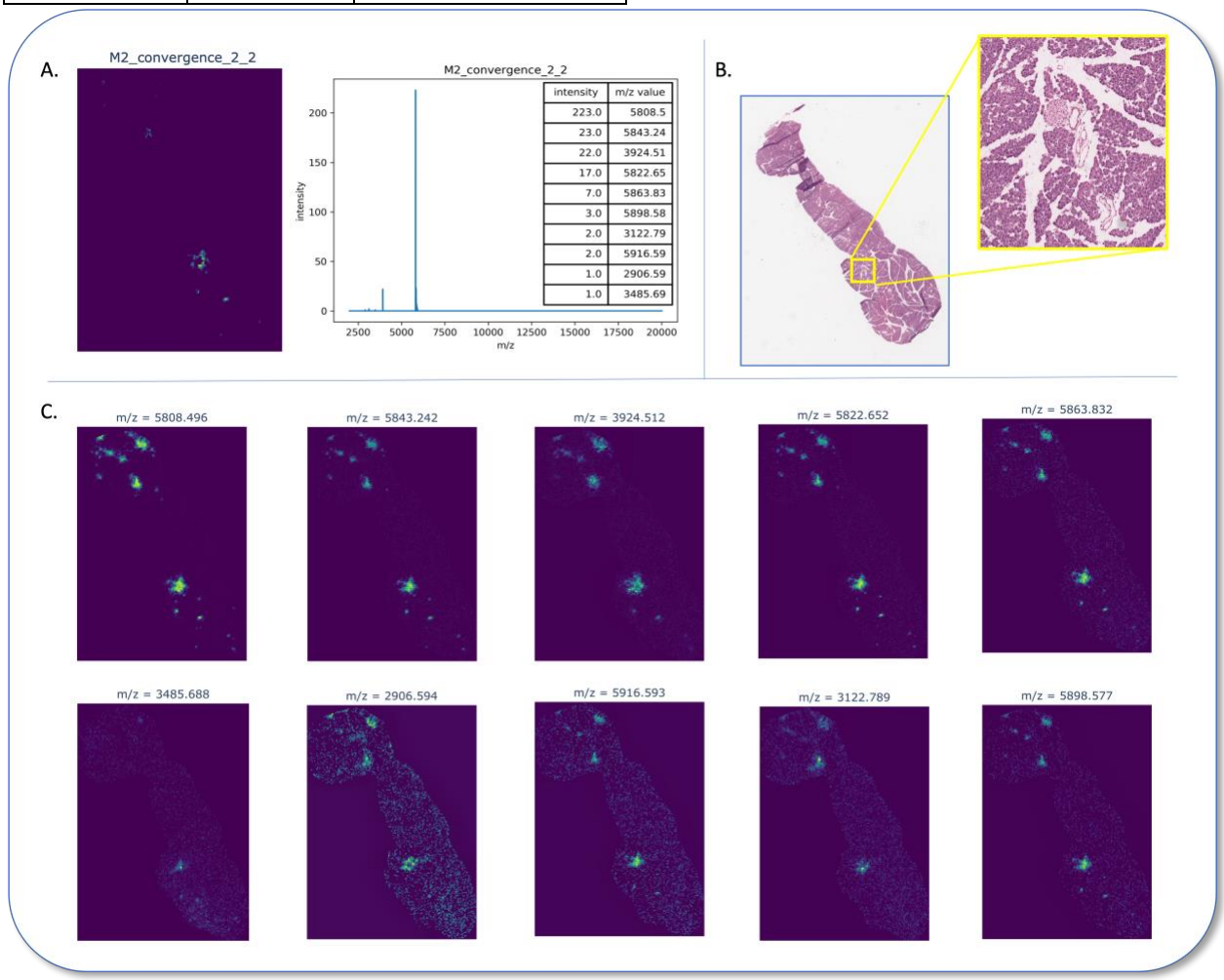

Figure 3. Spatial correspondence of cluster *"M2_convergence_2_2"* and the ion images of the corresponding m/z-peaks. The cluster shows relation between Ins1 (m/z = 5808.5 Da and m/z = 2906.59 Da) and the outer periphery of the islets of Langerhans in mice. (A) Result cluster obtained with rr2-NMF. (B) H&E staining of the M2 sample. (C) Ion images of the peak values corresponding to this cluster.

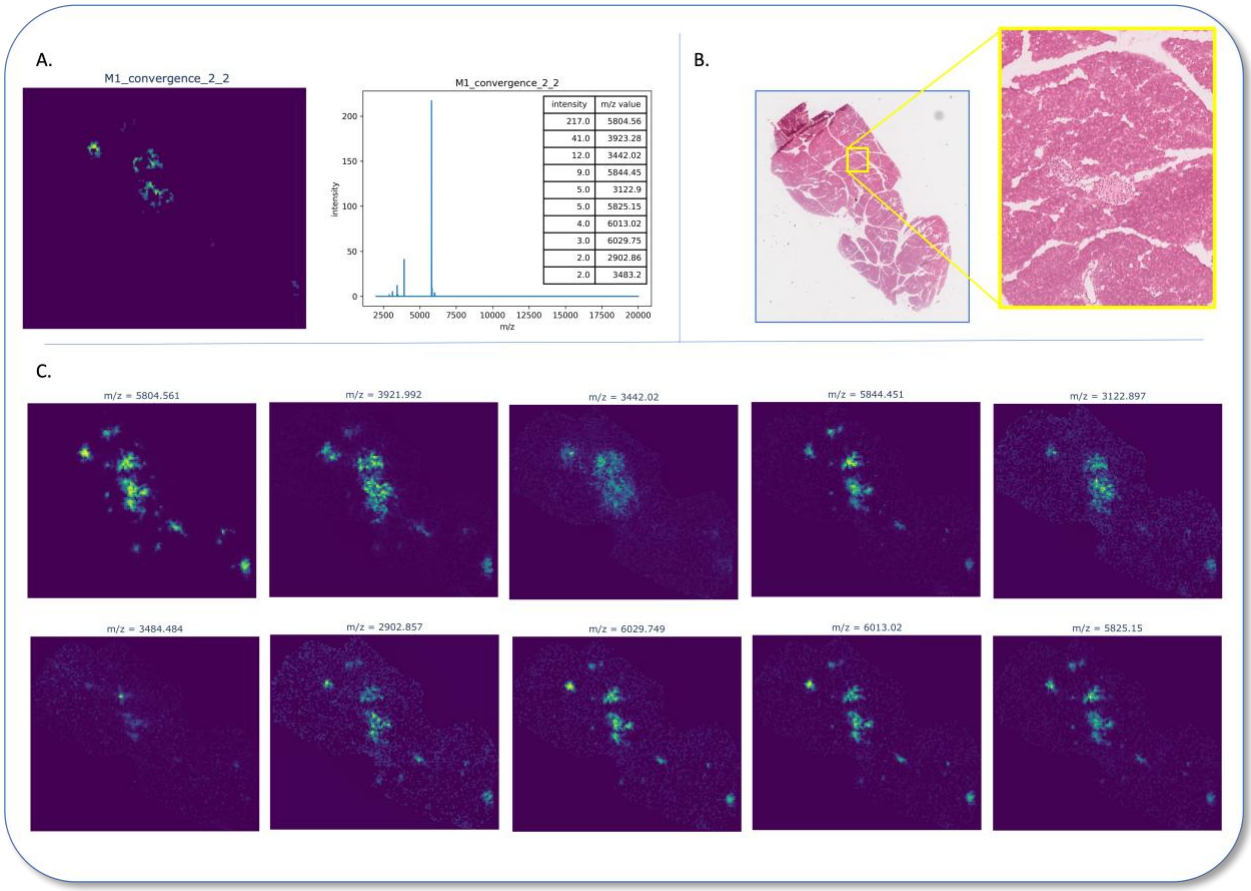

Figure 4. Spatial correspondence of cluster *"M1_convergence_2_2"* and the ion images of the corresponding m/z-peaks. The cluster shows the potentially higher presence of Ins1 in the outer periphery of the islets of Langerhans in this sample as well. (A) Result cluster obtained with rr2-NMF. (B) H&E staining of the M1 sample. (C) Ion images of the peak values corresponding to this cluster.

Extrapolating knowledge from the study of mice to humans, however, is not always straight forward. For one thing, we know that rodents have two types of insulin, i.e. Ins1 and Ins2, whereas in humans, we only have one (resembling mouse Ins2).[23] As known for rodents (and not the case for humans), we expect to find a circular organization in their islets of Langerhans, with alpha cells on the outer periphery and beta, gamma, and delta cells in the core.[24]

Confirming this pattern, rr2-NMF additionally allows to inspect all molecules following this orientation by looking at the ion images of the found m/z-values for this cluster. In Figure 3, we show correspondence of known and unknown molecules present in this outer periphery of the islets. With rr2-NMF, we now see a connection between this known circular orientation and the spatial patterns of additional molecules, such as Ins1 (m/z = 5808.5 Da). By using the protein atlas UniProt[25], Table 2 shows the possible molecular annotations of m/z-values of interest in the organization of the islets corresponding to cluster *"M2_convergence_2_2"* as shown in Figure 3. Note that the results of rr2-NMF are detected bi-clusters and not true ion images or spectra. They show patterns that arise on a higher level of detail.

Although lowly abundant, it is not a surprise that m/z = 3485.69 Da could be found in the outer border of the islets as this mass corresponds to the mass of glucagon, produced by the alpha cells.

More interestingly, when looking at the ion image of Ins1 (m/z = 5808.5 Da), it suggests an increased intensity along the outer border of the islets. The same is true for the two-charged state of Ins1 (m/z = 2906.59 Da). Here, because of the lower intensity and lower m/z-value of the molecule, the SNR becomes smaller, hence the presence of noise throughout the sample. Yet, the signal is detected because of the hierarchical bi-clustering approach. This potential higher presence of Ins1 in the outer border of the islets was further investigated (see below).

The study of the spatial distrubution and localization of Ins1 and Ins2 is not straightforward as the masses of 5800 Da (Ins2) and 5808.5 Da (Ins1) have partially overlapping isotope distributions. This explains mainly why earlier approaches sometimes could not distinct the two. In order to compare the spatial distributions of both molecules, we have analysed the ion images of the isotopes at – 5 Da (Ins2) and + 5 Da (Ins1) of their peak values, thereby increasing the ability to distinct both insulin isoforms. The result is shown in Figure 5, suggesting that Ins2 and Ins1 are both present in the center as well as in the border of the pancreatic islets. However, the result also suggests a higher correlation of Ins1 with the outer periphery of the islets, this while the intensity of the respective isotope peak of Ins2 (118 for m/z = 5795 Da) is clearly higher than the respective isotope peak of Ins1 (59 for m/z = 5813 Da).

This is suggestive for a difference in expression level of Ins2 and Ins1 in the center and in the border of the islets. Although,

a distinct regulation of the expression levels of Ins1 and Ins2 have been reported earlier[26,27], a different spatial-related expression in normal mouse pancreatic islets has not been studied yet and opens new horizons for further biomedical research.

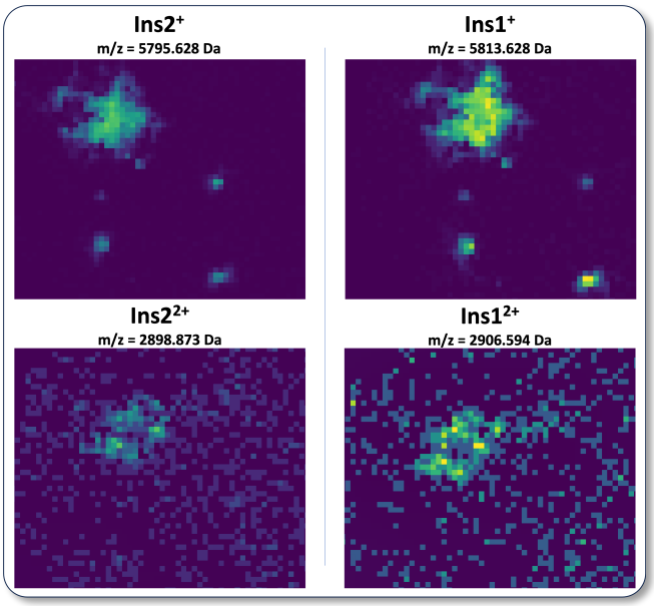

Figure 5. Zoomed-in islets of Langerhans of the M2 sample for Ins1 and Ins2 of the isotopes at – 5 Da (Ins2) and + 5 Da (Ins1) of their peak values in single charged state, and – 2.5 Da (Ins2) and + 2.5 Da (Ins1) for the double charged states. We see higher correlation of Ins1 with the outer border of the islets than Ins2.

Another observation is the presence of peptides displaying a mass shift of 14 Da for both the Ins1 and Ins2 full length protein. It is tempting to attribute this mass shift to a mono-methylation. However, it is clear that further research is required to investigate whether this is a new and previously not discovered PTM modification or an artifact or results from another explaination. Further exploration of this observation is certainly advisable as some understudied PTMs of insulin itself, such as citrullination and deamidation, have been reported to be linked with insulin secretion.[28]

Different ways exist to confirm our findings regarding the annotations of m/z -values of interest. We could search a confirmation in other, but similar samples, or perform a targeted molecular profiling of one of the samples. Here, we currently limit ourselves to the former. Figure 4 shows similar clustering for the M1 sample, and displays the ion images of the m/z-values contributing to the *"M1_convergence_2_2"* cluster. Supporting Information S5 shows the results on the third, M3, sample.

Computational efficiency

The gain in information compared to some traditional data-analytical tools comes at the small cost of more computational resources. This, however, is neglectable. On a 2019 MacBook Pro with 2,6 GHz 6-Core Intel Core i7 and 16 GB RAM, all pancreatic samples ran in less than 10 minutes wall time, see Supporting Information (S1). For more recent MALDI-TOF data sets, spatial resolutions are often much higher, and to this end the provided parallel implementation can significantly improve computation time.

CONCLUSION

With rr2-NMF, we present a novel approach, based on NMF, and tailored to the characteristics of MSI data. Applied to three mouse pancreas MALDI-TOF proteomics data sets, we demonstrate its capabilities to detect bi-clusters with an increased level of sensitivity for lower abundant signals due to the hierarchical nature of the method. The clusters now also include low intensities (see Table 1, Table 2 and the intensities tables in Figure 3 and 4).

Furthermore, the application of the method to the mouse pancreatic MSI datasets demonstrated identical spatial localisation of Ins1 and Ins2 but suggested differences in expression levels. Also a mass shift of 14 Da was observed in the clusters containing Ins1 and Ins2. The results were confirmed in in each sample of the three samples and demonstrate that the presented rr2-NMF approach should be regarded as a valuable alternative for processing MSI data. Moreover, with respect to the low computational time, the method could be particularly important when applying MSI in certain medical contexts, e.g. support of surgical decisions. Finally, the open-source tool allows other researchers to perform rr2-NMF on their own data.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website.

- Runtimes; Isotope selection; Results rr2-NMF on samples M1, M2, M3 (PDF)

AUTHOR INFORMATION

Corresponding Author

*Melanie Nijs − STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, Department of Electrical Engineering (ESAT), KU Leuven, 3001 Leuven, Belgium; orcid.org/ 0000-0001-9154-7543; Email: melanie.nijs@esat.kuleuven.be

Author Contributions

All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by KU Leuven Research Fund (projectsi BOF/23/064, 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; PhD Grant (SB/1SA1319N), EWI: the Flanders AI Research Program, VLAIO: CSBO (HBC.2021.0076) Baekeland PhD (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).

REFERENCES

(1) Caprioli, R. M.; Farmer, T. B.; Gile, J. Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. *Anal. Chem.* **1997**, *69* (23), 4751–4760. https://doi.org/10.1021/ac970888i.

(2) Alexandrov, T. Spatial Metabolomics and Imaging Mass Spectrometry in the Age of Artificial Intelligence. *Annual review of biomedical data science* **2020**, *3* (1), 61–87. https://doi.org/10.1146/annurev-biodatasci-011420-031537.

(3) Baquer, G.; Sementé, L.; Mahamdi, T.; Correig, X.; Ràfols, P.; García-Altares, M. What Are We Imaging? Software Tools and Experimental Strategies for Annotation and Identification of Small Molecules in Mass Spectrometry Imaging. *Mass Spectrometry Reviews* **2023**, *42* (5), 1927–1964. https://doi.org/10.1002/mas.21794.

(4) Buchberger, A. R.; DeLaney, K.; Johnson, J.; Li, L. Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights. *Analytical Chemistry* **2018**, *90* (1), 240–265. https://doi.org/10.1021/acs.analchem.7b04733.

(5) Ucal, Y.; Coskun, A.; Ozpinar, A. Quality Will Determine the Future of Mass Spectrometry Imaging in Clinical Laboratories: The Need for Standardization. *Expert Review of Proteomics* **2019**, *16* (6), 521–532. https://doi.org/10.1080/14789450.2019.1624165.

(6) Hu, H.; Laskin, J. Emerging Computational Methods in Mass Spectrometry Imaging. *Advanced Science* **2022**, *9* (34), 2203339. https://doi.org/10.1002/advs.202203339.

(7) Verbeeck, N.; Caprioli, R. M.; Van de Plas, R. Unsupervised Machine Learning for Exploratory Data Analysis in Imaging Mass Spectrometry. *Mass Spectrometry Reviews* **2020**, *39* (3), 245–291. https://doi.org/10.1002/mas.21602.

(8) Abdelmoula, W. M.; Pezzotti, N.; Hölt, T.; Dijkstra, J.; Vilanova, A.; McDonnell, L. A.; Lelieveldt, B. P. F. Interactive Visual Exploration of 3D Mass Spectrometry Imaging Data Using Hierarchical Stochastic Neighbor Embedding Reveals Spatiomolecular Structures at Full Data Resolution. *J. Proteome Res.* **2018**, *17* (3), 1054–1064. https://doi.org/10.1021/acs.jproteome.7b00725.

(9) Smets, T.; Waelkens, E.; De Moor, B. Prioritization of m/ z-Values in Mass Spectrometry Imaging Profiles Obtained Using Uniform Manifold Approximation and Projection for Dimensionality Reduction. *Analytical Chemistry* **2020**, *92* (7), 5240–5248. https://doi.org/10.1021/acs.analchem.9b05764.

(10) Mirkin, B. *Mathematical Classification and Clustering*; Pardalos, P., Horst, R., Series Eds.; Nonconvex Optimization and Its Applications; Springer US: Boston, MA, 1996; Vol. 11. https://doi.org/10.1007/978-1-4613-0457-9.

(11) López-Fernández, H.; Reboiro-Jato, M.; Madeira, S. C.; López-Cortés, R.; Nunes-Miranda, J. D.; Santos, H. M.; Fdez-Riverola, F.; Glez-Peña, D. A Workflow for the Application of Biclustering to Mass Spectrometry Data. In *7th International Conference on Practical Applications of Computational Biology & Bioinformatics*; Mohamad, M. S., Nanni, L., Rocha, M. P., Fdez-Riverola, F., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing: Heidelberg, 2013; pp 145–153. https://doi.org/10.1007/978-3-319-00578-2_19.

(12) Ovchinnikova, K.; Kovalev, V.; Stuart, L.; Alexandrov, T. OffsampleAI: Artificial Intelligence Approach to Recognize off-Sample Mass Spectrometry Images. *BMC Bioinformatics* **2020**, *21* (1), 129. https://doi.org/10.1186/s12859-020-3425-x.

(13) Prentice, B. M.; Hart, N. J.; Phillips, N.; Haliyur, R.; Judd, A.; Armandala, R.; Spraggins, J. M.; Lowe, C. L.; Boyd, K. L.; Stein, R. W.; Wright, C. V.; Norris, J. L.; Powers, A. C.; Brissova, M.; Caprioli, R. M. Imaging Mass Spectrometry Enables Molecular Profiling of Mouse and Human Pancreatic Tissue. *Diabetologia* **2019**, *62* (6), 1036–1047. https://doi.org/10.1007/s00125-019-4855-8.

(14) Grüner, B. M.; Hahne, H.; Mazur, P. K.; Trajkovic-Arsic, M.; Maier, S.; Esposito, I.; Kalideris, E.; Michalski, C. W.; Kleeff, J.; Rauser, S.; Schmid, R. M.; Küster, B.; Walch, A.; Siveke, J. T. MALDI Imaging Mass Spectrometry for In Situ Proteomic Analysis of Preneoplastic Lesions in Pancreatic Cancer. *PLOS ONE* **2012**, *7* (6), e39424. https://doi.org/10.1371/journal.pone.0039424.

(15) Eberlin, L. S.; Margulis, K.; Planell-Mendez, I.; Zare, R. N.; Tibshirani, R.; Longacre, T. A.; Jalali, M.; Norton, J. A.; Poultsides, G. A. Pancreatic Cancer Surgical Resection Margins: Molecular Assessment by Mass Spectrometry Imaging. *PLOS Medicine* **2016**, *13* (8), e1002108. https://doi.org/10.1371/journal.pmed.1002108.

(16) Derwae, H.; Nijs, M.; Geysels, A.; Waelkens, E.; De Moor, B. Spatiochemical Characterization of the Pancreas Using Mass Spectrometry Imaging and Topological Data Analysis. *Anal. Chem.* **2023**, *95* (28), 10550–10556. https://doi.org/10.1021/acs.analchem.2c05606.

(17) Nijs, M.; Smets, T.; Waelkens, E.; De Moor, B. A Mathematical Comparison of Non-Negative Matrix Factorization Related Methods with Practical Implications for the Analysis of Mass Spectrometry Imaging Data. *Rapid Communications in Mass Spectrometry* **2021**, *35* (21), e9181. https://doi.org/10.1002/RCM.9181.

(18) Smets, T.; Verbeeck, N.; Claesen, M.; Asperger, A.; Griffioen, G.; Tousseyn, T.; Waelput, W.; Waelkens, E.; De Moor, B. Evaluation of Distance Metrics and Spatial Autocorrelation in Uniform Manifold Approximation and Projection Applied to Mass Spectrometry Imaging Data. *Analytical Chemistry* **2019**, *91* (9), 5706–5714. https://doi.org/10.1021/acs.analchem.8b05827.

(19) Alexandrov, T. MALDI Imaging Mass Spectrometry: Statistical Data Analysis and Current Computational Challenges. *BMC Bioinformatics* **2012**, *13* (16), S11. https://doi.org/10.1186/1471-2105-13-S16-S11.

(20) Madeira, S. C.; Oliveira, A. L. Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2004**, *1* (1), 24–45. https://doi.org/10.1109/TCBB.2004.2.

(21) Gillis, N. *Nonnegative Matrix Factorization*; Society for Industrial and Applied Mathematics: Philadelphia, PA, 2020. https://doi.org/10.1137/1.9781611976410.

(22) Pedregosa, F.; Weiss, R.; Brucher, M.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12* (85), 2825–2830.

(23) Shiao, M.-S.; Liao, B.-Y.; Long, M.; Yu, H.-T. Adaptive Evolution of the Insulin Two-Gene System in Mouse. *Genetics* **2008**, *178* (3), 1683–1691. https://doi.org/10.1534/genetics.108.087023.

(24) Campbell, J. E.; Newgard, C. B. Mechanisms Controlling Pancreatic Islet Cell Function in Insulin Secretion. *Nat Rev Mol Cell Biol* **2021**, *22* (2), 142–158. https://doi.org/10.1038/s41580-020-00317-7.

(25) The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **2023**, *51* (D1), D523–D531. https://doi.org/10.1093/nar/gkac1052.

(26) Deltour, L.; Leduque, P.; Blume, N.; Madsen, O.; Dubois, P.; Jami, J.; Bucchini, D. Differential Expression of the Two Nonallelic Proinsulin Genes in the Developing Mouse Embryo. *Proceedings of the National Academy of Sciences* **1993**, *90* (2), 527–531. https://doi.org/10.1073/pnas.90.2.527.

(27) Roderigo-Milne, H.; Hauge-Evans, A. C.; Persaud, S. J.; Jones, P. M. Differential Expression of Insulin Genes 1 and 2 in MIN6 Cells and Pseudoislets. *Biochemical and Biophysical Research Communications* **2002**, *296* (3), 589–595. https://doi.org/10.1016/S0006-291X(02)00913-0.

(28) Yang, C.; Wei, M.; Zhao, Y.; Yang, Z.; Song, M.; Mi, J.; Yang, X.; Tian, G. Regulation of Insulin Secretion by the Post-Translational Modifications. *Frontiers in Cell and Developmental Biology* **2023**, *11*.

For Table of Contents Only

