KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Non-negative Matrix Factorization for the Analysis of Mass Spectrometry Imaging Data

with applications to other omics modalities

Melanie Nijs

Supervisors:
Prof. dr. ir. Bart De Moor
Prof. dr. Etienne Waelkens

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

July 2024

Non-negative Matrix Factorization for the Analysis of Mass Spectrometry Imaging Data

with applications to other omics modalities

Melanie NIJS

Examination committee:
Prof. dr. ir. Omer Van der Biest, chair
Prof. dr. ir. Bart De Moor, supervisor
Prof. dr. Etienne Waelkens, supervisor
Prof. dr. ir. Johan Suykens
Prof. dr. Rita Derua
Prof. dr. ir. Gamze Gürsoy
(Columbia University, USA)
Prof. dr. Eva Cuypers
(Maastricht University, NL)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

July 2024

© 2024 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Melanie Nijs, Kasteelpark Arenberg 10 box 2446, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

♡

dedicated to my three beautiful sisters
in loving memory of our parents

Preface

Pursuing a PhD gives you an opportunity to push boundaries, both scientifically and personally. For me, it has been a way to be creative with mathematics, to discover omics, and to allow my imagination to meet my passion for science. In this sense, I really did reach new horizons. Personally, however, it pushed me a bit too far. It feels surreal that I am actually writing this preface. And for that, I have many people to thank.

First and foremost, I want to express my gratitude to Prof. Etienne Waelkens. Etienne, without you, your patience to review my writing (to make sure it's written with the scientific caution the MSI field requires), your teaching, your contagious passion for science, and your support whenever I felt overwhelmed, I would not have been able to finish this work. For that, from the bottom of my heart, thank you! Also Prof. Bart De Moor, thank you for the opportunity to conduct this PhD.

Secondly, with an unbelievable timing of reaching out to me, I want to thank Prof. Gamze Gürsoy and her entire lab. The six months at the New York Genome Center allowed me to work in an amazing environment. It provided me the strength and the confidence to continue my research. It was such a blessed time, and left nothing but happy memories. Thank you for that, Gamze! I can't wait for my next visit to your lab.

Additionally, a big thank you to the Professors that have followed up on my work, took the time to read and improve this thesis, and offer support from the sidelines. Therefore, Prof. Johan Suykens, Prof. Rita Derua, thank you for your interest in my work, the fruitful discussions, and improving my research throughout the progress meetings. To prof. Eva Cuypers, thank you for the nice collaboration, the friendly face at the IMSIS conference, the tour at of the M4I lab and answering all the biomedical questions we have about the TBM Smoke project. Also, thank you to be a part of this jury and for the detailed feedback on my manuscript. Finally, Prof. Omer Van der Biest, thank you for

your time to act as the chair of my examination committee. Even though it was a last-minute change, your kindness during the preliminary defense was very much appreciated.

Not at the least I would also like to thank Ida, Aldona, Elsy, John and Maarten for the warmth at the coffee corner, the chats in the hallway and helping me out with all the administrative chaos a PhD brings along. I will miss you!

During these last few years, I've had the pleasure to work with the super smart and amazing colleagues. To everyone at STADIUS and everyone at the G2Lab: thank you!

Giulia, Lola, Naomi, my buddies throughout the whole PhD, thank you for being always full of love and support; Yamini, Zahra, Katrien, Wanqiu, let's keep on organizing our lunches and dinner parties; Thomas, Tina, to have had MSI buddies like you both has been an incredible support; Christof and Oliver, the TA activities and board game evenings were a highlight during the pandemic; but also everyone else: Mauricio, Fernando, Alexandre, Axel, Jaldert, Sem, Lukas, Sibren, Hans, Hannes, Kostas, Youri, Sarthak, Vincent, Willem, Benoît, Bob, Xi; thank you for all the fun moments at the office, and all the kindness you have showed me.

Also the people for New York, who have made my time inside and outside of NYGC so amazing: Paul and Abby, thank you for being the best desk buddies I could have wished for! I had so much fun getting to know you, and look forward to our adventures still to come! Carlos and Anoushka, thank you for giving me a crash course in genomics and being always so sweet. But also Jacob, Annie, Gedas, Conor, Ahmed, Kirill, Seungwan, Rashid, and of course Andrew. I truly had a blast at all the happy hours, the ping pong sessions, the boat tour, ... It was such a warm and happy environment, thank you!

Most importantly, I am so blessed with an incredible group of friends always there to support me, comfort me, dance with me, make me laugh, make me whole. It is what I cherish most in life. I can't even begin writing your names without tearing up, but I could write a whole new book thanking each and everyone of you. I love you all.

It is, however, a bittersweet feeling to be finishing my PhD, as it highlights the absence of those who cannot witness this moment, those I have lost too soon. There have been so many times I wanted to talk to you and get your advice. While the grief has turned into beautiful memories, I will always hear your voices in those difficult times.

As my *Papa* used to say: *"Doe wat dat ge denkt dat ge moet doen en geniet ervan"*, as my *Mama* used to say: *"Ge doet uw best en meer kunt ge niet doen"*,

as my *Bompa* used to say: *"Wat ge vandaag doet, moet ge morgen niet meer doen"*, and as my *Bonneke* used to say: *"C'est la vie d'artiste, c'est beau et c'est triste"*.

Not a day goes by without wishing you were still here, but thank you for giving me three amazing sisters. Zusjes, ik zie jullie graag.

Melanie

July 2024

Abstract

Molecular biology explores the flow of information from DNA to cell function and seeks to understand complex biological processes on a molecular level. Driven by huge advancements in the field of omics technologies over the last few decades, we can now study biomolecules such as DNA, RNA, proteins, and metabolites on a large scale. With applications in diagnostic tools, therapeutic interventions, genetically modified organisms, and personalized medicine, molecular biology and omics research contribute to our quality of life in numerous ways.

The word "omics" encompasses fields as genomics, transcriptomics, proteomics, metabolomics, ... depending on the specific type of biomolecule studied. Each omic modality relies on high-throughput infrastructure to generate large-scale data sets, and advanced bioinformatics to unravel the molecular mechanisms underlying physiological processes, pathogenesis, and environmental responses.

A single omics experiment results in a complex, large, high-dimensional, and noisy data set, but provides a wealth of information. As manual exploration of these data sets has become cumbersome, the challenge today lies in detecting the useful information using computational tools. Throughout this dissertation, we inspect mass spectrometry imaging (MSI), multiplexed immunohistochemistry (mIHC), and chromatin conformation capture (3C) data sets. For all of these modalities, computational analysis can be very complex due to low signal-to-noise ratios, complicated normalization approaches, and crucial information often being hidden in the smaller details.

Tackling these challenges of omics data analysis, we researched the use of non-negative matrix factorization (NMF) for the unsupervised dimensionality reduction and clustering of the data sets in an explainable manner. NMF is an optimization algorithm which aims to approximate a non-negative data set by a number of linearly additive components, represented as a product of two low rank matrices. It requires that both the data and the factor matrices only contain non-negative numbers, which is often the case for omics data.

In the first part of this thesis, we focus solely on the analysis of MSI data. MSI studies proteins and metabolites in their micro-environment. It is an untargeted, label-free technology that measures the spatial distribution of thousands of molecules present in a slice of tissue. This makes MSI an essential tool for biomarker discovery and the study of heterogeneous cancers. For the analysis of MSI data, we started with a mathematical comparison of the performance of four different NMF-related methods for the analysis of multiple MSI data sets. We provided detailed information on the mathematical implications of each method and demonstrated how NMF approaches that assume Poisson noise, a known property of MSI data, perform better. Secondly, we inspected the Mapper algorithm, a topological data analysis approach, to visualize MSI data of mice pancreas. Using the optimal parameter settings, we discovered a more detailed structure in the data compared to uniform manifold approximation and projection (UMAP). More specifically, we detected two clusters within the islets of Langerhans that were not detected with UMAP. In the final chapter of the first part, we introduced an extension of NMF tailored for the analysis of MSI data. Using a hierarchical implementation of NMF, we gained more information from the small and convoluted signals. This new approach allowed us to identify the two clusters in the islets and see differences in abundance between the two types of insulin correlating with these clusters.

In the second part of the thesis, we extend the knowledge on NMF collected from MSI data to data sets obtained with other omics modalities. In mIHC, multiple protein markers are measured in-situ using fluorescence signals. While this is a targeted approach with a limited number of molecules measured, mIHC offers a much higher spatial resolution than MSI up to (sub)cellular level. Here, we presented a downstream analysis pipeline based on the combination of NMF and a trained classifier for the fast clustering and automatic labelling of human melanoma samples obtained with MILAN technology. Due to NMF's ability to handle large data sets, the proposed pipeline obtained better computational efficiency than previous analyses of this data set. Finally, we move to genomics data analysis. More specifically, we inspected the three-dimensional structure of the chromatin in single cells obtained from developing mouse brain. We applied multiple non-negative matrix and tensor factorization approaches to find the correlation between cell-types and connected parts of the chromatin. Unique chromatin contacts for a certain cell-type could indicate an essential protein encoding region, which could then be linked to the corresponding transcripts. Unfortunately, more advanced tools are required to uncover the true signals, since the high presence of noise misdirects traditional analyses. To this end, we highlight how factorization techniques can improve results and can be extended to multi-omic analyses.

Beknopte samenvatting

Moleculaire biologie onderzoekt de informatiestroom van DNA naar celfunctie en streeft ernaar complexe biologische processen op moleculair niveau te begrijpen. Gedreven door enorme technologische vooruitgang op het gebied van de omics-technologieën in de afgelopen decennia, kunnen we nu biomoleculen zoals DNA, RNA, eiwitten en metabolieten op grote schaal bestuderen. Met toepassingen in diagnostische methodes, therapeutische behandelingen, genetisch gemodificeerde organismen en gepersonaliseerde geneeskunde dragen moleculaire biologie en omics-onderzoek op tal van manieren bij aan onze levenskwaliteit.

Het begrip "omics" omvat onderzoeksdomeinen zoals genomica, transcriptomica, proteomica, metabolomica, afhankelijk van het type molecule dat wordt bestudeerd. Elke omics-modaliteit vertrouwt op *high-throughput* infrastructuur om grootschalige datasets te genereren, en geavanceerde bio-informatica om de moleculaire mechanismen te ontrafelen die aan de basis liggen van fysiologische, pathologische en omgevingsprocessen.

Een enkel omics-experiment resulteert in een complexe, grote, hoog-dimensionale dataset, maar biedt een rijkdom aan informatie. Omdat het handmatig onderzoeken van deze grote datasets omslachtig is, ligt de uitdaging vandaag in het detecteren van nuttige informatie met behulp van computationele methodes. Doorheen dit proefschrift inspecteren we massaspectrometrie-beeldvorming (MSI), multiplex immunohistochemie (mIHC) en chromatine conformatie-capture (3C) datasets. Voor al deze modaliteiten kan de computationele analyse zeer complex zijn vanwege lage signaal-ruisverhoudingen, ingewikkelde normalisatiebenaderingen en cruciale informatie die vaak verborgen zit in de kleinere details.

Om deze uitdagingen van omics-data-analyse aan te pakken, hebben we het gebruik van niet-negatieve matrixfactorisatie (NMF) onderzocht voor de *unsupervised* dimensionaliteitsreductie en clustering van de datasets op een verklaarbare manier. NMF is een optimalisatiemethode die als doel heeft een niet-

negatieve dataset te benaderen door een aantal lineair additieve componenten, voorgesteld als het product van twee lage-rang matrices. Het vereist dat zowel de data als de factormatrices niet-negatieve getallen bevatten, wat vaak het geval is voor omics-data.

In het eerste deel van dit proefschrift richten we ons uitsluitend op de analyse van MSI-data. MSI bestudeert eiwitten en metabolieten in hun micro-omgeving. Het is een ongerichte, labelvrije technologie die de spatiale verdeling meet voor duizenden moleculen aanwezig in een weefselbiopsie. Dit maakt MSI een essentieel instrument voor de identificatie van biomarkers en de studie van heterogene kankers. Voor de analyse van MSI-data zijn we begonnen met een wiskundige vergelijking van de prestaties van vier verschillende NMF-gerelateerde methoden voor de analyse van meerdere MSI-datasets. We hebben gedetailleerde informatie verstrekt over de wiskundige implicaties van elke methode en gedemonstreerd hoe NMF-benaderingen die uitgaan van Poisson-ruis, een bekende eigenschap van MSI-data, beter presteren. Ten tweede hebben we het Mapper-algoritme geïnspecteerd, een methode voor topologische data-analyse, om MSI-gegevens van muizenpancreas te visualiseren. Met behulp van de optimale parameters hebben we een gedetailleerdere structuur in de gegevens ontdekt in vergelijking met *uniform manifold approximation en projection* (UMAP). We detecteren specifiek twee clusters binnen de eilandjes van Langerhans gedetecteerd die niet werden gedetecteerd met UMAP. In het laatste hoofdstuk van het eerste deel hebben we een uitbreiding van NMF geïntroduceerd die is afgestemd op de analyse van MSI-data. Met behulp van een hiërarchische implementatie van NMF hebben we meer informatie verkregen uit de kleine en complexe signalen. Deze nieuwe benadering stelde ons in staat om de twee clusters in de eilandjes te identificeren en verschillen in aanwezigheid tussen de twee typen insuline te zien die correleren met deze clusters.

In het tweede deel van het proefschrift breiden we de kennis over NMF voor MSI data uit naar datasets verkregen met andere omics-modaliteiten. In mIHC worden meerdere eiwitmarkers in situ gemeten met fluorescentiesignalen. Hoewel dit een gerichte aanpak is met een beperkt aantal gemeten moleculen, biedt mIHC een veel hogere spatiale resolutie dan MSI tot (sub)cellulair niveau. Hier presenteerden we een *downstream*-analysemethode gebaseerd op de combinatie van NMF en een classificatiemodel voor het snel clusteren en automatisch labelen van menselijke melanoomstalen verkregen met MILAN-technologie. Dankzij het vermogen van NMF om grote datasets te verwerken, bereikt de voorgestelde methode een betere computationele efficiëntie dan eerdere analyses van deze dataset. Ten slotte gaan we over naar de analyse van genomische data. Meer specifiek hebben we de driedimensionale structuur van het chromatine in enkele cellen geïnspecteerd, verkregen uit ontwikkelende muizenhersenen. We pasten meerdere niet-negatieve matrix- en tensordecomposities toe om de correlatie

te vinden tussen celtypen en verbonden delen van het chromatine. Unieke chromatinecontacten voor een bepaald celtype kunnen wijzen op een essentieel coderingsgebied voor eiwitten, dat vervolgens kan worden gekoppeld aan de corresponderende transcripten. Helaas zijn meer geavanceerde tools nodig om de ware signalen te onthullen, aangezien de hoge aanwezigheid van ruis traditionele analyses misleidt. Om deze reden benadrukken we hoe factorisatietechnieken de huidige resultaten kunnen verbeteren en kunnen worden uitgebreid naar *multi-omics*-analyses.

List of Abbreviations

3C	chromatin conformation capture
AE	autoencoder
AF	autofluorescence
ALS	alternating least squares
ARI	adjusted Rand index
CH	Calinsky-Harabasz
CLAHE	contrast limited adaptive histogram equalization
CNN	convolutional neural network
CPD	canonical polyadic decomposition
Da	Dalton
DBSCAN	density-based spatial clustering of applications with noise
DESI	desorption electrospray ionization
DNA	dioxyribonucleic acid
E-P	enhancer-promoter
FFPE	formalin-fixed parfin-embedded
GMM	Gaussian mixture model
H&E	hematoxylin and eosin
HC	hierarchical clustering

HDBSCAN hierarchical density-based spatial clustering of applications with noise

HOSVD higher order singular value decomposition

HPC high power computing

ICA independent component analysis

IHC immunohistochemistry

Ins1 insulin 1

Ins2 insulin 2

KL Kullback-Leibler

KNN k nearest neighbours

LC-MS liquid chromatography mass spectrometry

LDA latent Dirichlet allocation or linear discriminant analysis

m/z mass-to-charge

MALDI matrix assisted laser desorption/ionization

mIHC multiplexed immunohistochemistry

MILAN multiple iterative labeling by antibody neodeposition

ML machine learning

MLP multi-layer perceptron

MNF maximum noise fraction

mRNA messenger ribonucleic acid

MS mass spectrometry

MS/MS tandem mass spectrometry

MSE mean squared error

MSI mass spectrometry imaging

MU multiplicative update

NMF non-negative matrix factorization

NN neural network

PCA	principal component analysis
PCR	polymerase chain reaction
PLS	partial least squares
PLSA	probilistic latent semantic analysis
PTM	post-translational modification
RF	random forest
RNA	ribonucleic acid
ROI	region of interest
rr2-NMF	recursive rank-2 non-negative matrix factorization
SA	sinapinic acid
sc-RNAseq	single cell RNA sequencing
scHi-C	single cell Hi-C data
SDF	structured data fusion
SIMS	secondary ion mass spectrometry
SNR	signal-to-noise ratio
SOM	self organizing map
SVD	singular value decomposition
SVM	support vector machine
t-SNE	t-distributed stochastic neighbor embedding
TAD	topological associated domain
TDA	topological data analysis
TIC	total ion current
TMA	tissue micro array
TOF	time-of-flight
UMAP	uniform manifold approximation and projection

Contents

Abstract vii

Beknopte samenvatting ix

List of Abbreviations xv

Contents xvii

List of Figures xxiii

List of Tables xxvii

Introduction 2

1 Introduction 3

1.1 Omics data: what, why and how? 4

1.2 Mass Spectrometry Imaging 10

1.2.1 The basics of MSI 10

1.2.2 MSI technology 13

1.2.3 Advantages and limitations of MSI 16

1.2.4 MSI data analysis 17

1.3 On the islets of Langerhans in mice pancreas 24

1.4 Thesis outline and main contributions 26

I Algorithms for Mass Spectrometry Imaging Data Analysis 31

2 Non-negative matrix factorization and related methods for MSI data analysis 33

2.1 Introduction . . . 35
2.2 Methods . . . 38
2.2.1 Non-negative Matrix Factorization (NMF) . . . 38
2.2.2 Kullback-Leibler Non-negative Matrix Factorization (KL-NMF) . . . 39
2.2.3 Probabilistic Latent Semantic Analysis (PLSA) . . . 40
2.2.4 Latent Dirichlet Allocation (LDA) . . . 42
2.2.5 Evaluation criteria . . . 43
2.3 Experimental setup . . . 45
2.3.1 Data . . . 45
2.3.2 Implementation . . . 45
2.4 Results . . . 46
2.4.1 Rank selection . . . 46
2.4.2 Pancreas data . . . 47
2.4.3 Lymph-node data . . . 51
2.4.4 Additional evaluation criteria . . . 53
2.5 Discussion . . . 55
2.6 Conclusion . . . 59
2.7 Acknowledgements . . . 60

3 Spatiochemical characterization of mouse pancreas MSI data using topological data analysis 63

3.1 Introduction . . . 65
3.2 Experimental section . . . 66
3.2.1 Data acquisition . . . 66
3.2.2 Mapper for clustering . . . 67
3.2.3 Mapper algorithm . . . 68
3.2.4 Parameter selection . . . 69

3.3 Results and discussion 70
3.3.1 Comparing filter functions 70
3.3.2 Quantitative comparison 72
3.3.3 Visual comparison 73
3.3.4 Matching ion images 74
3.3.5 Second data set 74
3.4 Conclusion 75
3.5 Acknowledgements 76

4 Hierarchical bi-clustering of mouse pancreas MSI data using recursive rank-2 NMF 77
4.1 Introduction 79
4.2 Methods 80
4.2.1 MALDI data acquisition and preprocessing 80
4.2.2 An rr2-NMF algorithm for hierarchical bi-clustering of MSI data 81
4.2.3 Implementation details and code availability 82
4.3 Results and discussion 84
4.3.1 Evaluation of rr2-NMF on three healthy mouse pancreas samples 84
4.3.2 A closer look at the islets of Langerhans in mice pancreas 84
4.3.3 Computational efficiency and limitations 91
4.4 Conclusion 92
4.5 Acknowledgements 92

II Applications to other omics modalities 93

5 An NMF-MLP pipeline for fast exploration and annotation of single cell MILAN data 95
5.1 Introduction 96
5.2 MILAN data of human melanoma samples 100
5.3 The NMF-MLP pipeline 101
5.4 Results & Discussion 103

5.4.1 Multi-sample analysis . . . 103
5.4.2 In-depth analysis of a single melanoma sample . . . 104
5.4.3 Computational benefits of the pipeline . . . 106
5.4.4 Future work . . . 107
5.5 Conclusion . . . 108
5.6 Acknowledgements . . . 109

6 Matrix and tensor factorization for cell type clustering of single cell Hi-C data 111
6.1 Introduction . . . 112
6.2 Single cell Hi-C genomics data of developing mouse brain . . . 116
6.3 scHi-C data analysis using matrix and tensor factorization . . . 118
6.3.1 Preprocessing . . . 119
6.3.2 Non-negative matrix factorization . . . 123
6.3.3 Non-negative tensor factorization . . . 128
6.4 Discussion . . . 131
6.4.1 Performance of the decompositions . . . 132
6.4.2 Future work . . . 132
6.5 Conclusion . . . 134
6.6 Acknowledgements . . . 134

Conclusion 138

7 Conclusion and future perspectives 139
7.1 Conclusion . . . 140
7.1.1 Algorithms for the analysis of MSI data . . . 140
7.1.2 Application to other modalities . . . 142
7.2 Remaining challenges . . . 143
7.3 Future perspectives . . . 145

Appendices 150

A Supporting Information for Non-Negative Matrix Factorization and related methods for MSI Data Analysis 151

B Supporting Information for Spatiochemical characterization of mouse pancreas MSI data using topological data analysis 159

C Supporting Information for Hierarchical bi-clustering of mouse pancreas MSI data using recursive rank-2 NMF 193

D Supporting Information for An NMF-MLP pipeline for fast exploration and annotation of single cell MILAN data 253

E Supporting Information for Matrix and tensor factorization of single cell Hi-C data 291

Bibliography 305

List of publications 331

List of Figures

1.1 Overview omics 4
1.2 Example spatial multi-omics 9
1.3 Workflow MSI analysis 12
1.4 Overview MALDI 13
1.5 Types of MSI technologies 15
1.6 MSI data analysis workflow 18
1.7 Overview of popular algorithms used for the supervised and unsupervised analysis of MSI data, including references to recent literature. Figure inspired by [56]. 20
1.8 Graphical representation of non-negative matrix factorization. . 24
1.9 H&E staining of mouse pancreas 25
1.10 H&E staining of mouse pancreas islet of Langerhans 26
1.11 Mouse vs human pancreas anatomy 27

2.1 Graphical representation of non-negative matrix factorization. . 36
2.2 Rank selection of the pancreas sample M2 47
2.3 Rank selection of the pancreas sample M2 47
2.4 Visual results of the mouse pancreas sample M2 49
2.5 Visual results of the human-lymph-node sample P1 50
2.6 UMAP visualization of samples M2 and P1 54
2.7 Analysis of top five intensities of sample P1 58
2.8 Analysis of four values of interest of sample P1 59

3.1 Overview Mapper algorithm 68
3.2 Mapper clustering workflow 69
3.3 Mapper clusters using different settings 71
3.4 UMAP benchmark 73
3.5 Overlapping ion images 75
3.6 Mapper on second data set 76

4.1 rr2-NMF using fixed depth stopping criterion 86
4.2 rr2-NMF using convergence stopping criterion 87
4.3 Visualisation ion images of the detected m/z values of an end cluster for sample M2. 88
4.4 Zoomed-in islets of Langerhans of the M2 sample for Ins1 and Ins2. 90
4.5 Visualisation ion images of the detected m/z values of an end cluster for sample M1. 91

5.1 Principle of IHC 96
5.2 Overview MILAN staining 97
5.3 Overview mIHC pipeline 98
5.4 NMF-MLP pipeline for MILAN data. 101
5.5 Results melanoma multi-sample analysis 103
5.6 In-depth visualization of melanoma sample 104
5.7 Detailed fingerprints of melanoma sample 106

6.1 Overview DNA organization 112
6.2 scHi-C protocol and data format 115
6.3 Description scHi-C data obtained with Dip-C 117
6.4 Matrix and tensor representation for scHi-C data 119
6.5 BandNorm normalization example 120
6.6 Genomic distance 121
6.7 Imputation of a scHi-C contact map 122
6.8 Schematic non-negative matrix factorization 123
6.9 Semi-log plot of the singular values. 124

6.10 Visualization single NMF component 124
6.11 ARI scores NMF for different resolutions 125
6.12 ARI scores NMF for different preprocessing techniques 126
6.13 ARI scores KL-NMF for different resolutions 126
6.14 ARI scores KL-NMF for different preprocessing techniques . . . 127
6.15 ARI scores LDA for different resolutions 127
6.16 ARI scores LDA for different preprocessing techniques 128
6.17 Schematic Tucker decomposition of a three dimensional tensor 129
6.18 ARI scores Tucker for different resolutions 129
6.19 ARI scores Tucker for different preprocessing techniques 130
6.20 Schematic CPD decomposition of a three dimensional tensor . 130
6.21 ARI scores CPD for different resolutions 131
6.22 ARI scores CPD for different preprocessing techniques 131

List of Tables

2.1 Specifications data sets 48
2.2 Numerical analysis of pancreas samples 51
2.3 Numerical analysis of lymph samples 52

4.1 List of the m/z values corresponding to the peaks of cluster "M2_convergence_2_1". 85
4.2 List of the m/z values corresponding to the peaks of cluster "M2_convergence_2_2". 85

5.1 Properties melanoma data 100
5.2 Classification result of melanoma sample 105

Introduction

Chapter 1

Introduction

Mass Spectrometry Imaging, or MSI in short, allows for the untargeted, spatial visualisation of thousands of molecules present in a slice of tissue. It offers a goldmine in information, but the data's complex nature requires advanced mining techniques. The research presented in this thesis offers new insights towards the analysis of MSI data, mainly focussing on non-negative matrix factorization (NMF). Additionally, this work inspects how similar factorization approaches can benefit the analysis of other omics modalities.

This first chapter introduces some background information regarding omics technologies in general, and more specifically MSI, where we give a detailed overview on MSI data acquisition and analysis. Additionally, to ensure non-experts a fluent reading, we briefly describe the biology of mice pancreas. At the end of this chapter, a detailed outline of the thesis is presented, together with a short summary on the main contributions.

1.1 Omics data: what, why and how?

The Central Dogma: *"Once information has got into a protein it can't get out again." Information here means the sequence of the amino acid residues, or other sequences related to it.* - Francis Crick (1957) [43].

Near the end of the previous century, only a few decades ago, multiple technological innovations caused an unseen acceleration in the study of Crick's central dogma in molecular biology. These technologies, called omics, allow us to understand how information flows from our genes, through transcription and translation into proteins and metabolites, all the way to observable phenotypes in an organism, see Figure 1.1 [111].

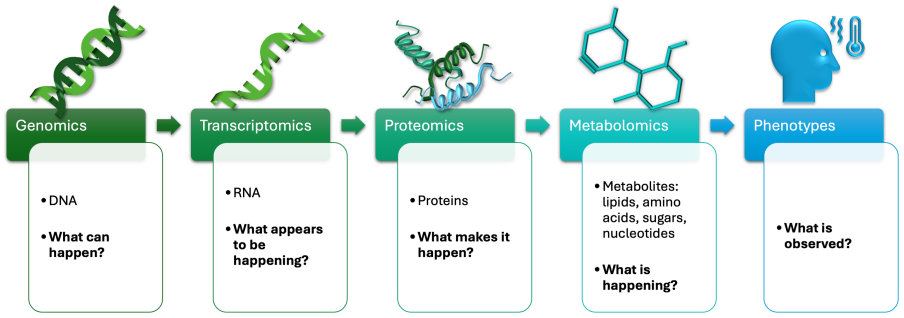


Figure 1.1: Overview of the omics cascade of information. DNA is transcribed into RNA, which can be translated into proteins. These proteins catalyse in turn metabolic interactions, which lead to observable phenotypes.

Given the huge number of genes, proteins, and cells in an organism, it is evident how omics tools generate data sets just as enormous, making it practically impossible to inspect them manually. For this reason, biochemical experts call upon the aid of bioinformatics to analyze and present the data in a comprehensive manner. To do so rigorously, a bioinformatician should always be aware of what the data represents, what the aim of the experiment is, and how the data acquisition can influence the true measurements. This first section gives a concise background on the what, why and how of omics data for the bioinformatics novice.

So, what is omics data? **Omics is an umbrella term to describe all technologies that provide us information of how a cell, tissue or organism functions** [128]. We can situate it in the field of molecular biology, but it encompasses a multidisciplinary collection of expertise, including: medicine, chemistry, pathology, pharmacology, physics, and bioinformatics. Due

to many technological advances, high-throughput tools have been developed to study various types of omics modalities efficiently. Here follows a short overview of the main different omics modalities: what they study, why they are useful, and how they are acquired. Note that this list is not exhaustive and we refer to the works by [128, 111, 96, 8, 14, 42, 109] for more information.

- **Genomics**

  The study of the function, expression and structure of all the genes in a cell or organism, including the interaction of those genes with each other and with the organism's environment. The genome is defined as the full set of DNA (deoxyribonucleic acid). Each DNA molecule consists of a double helix composed with the four nucleotide bases: adenine (A), thymine (T), guanine (G), and cytosine (C), with fixed pairing combinations A-T or C-G on the opposite strands of the helix. A gene then describes the set of nucleotides on the DNA strand that contains the instructions for the creation of one or more specific proteins.

  DNA sequencing [165] is used in genomics to read out the order of the nucleotide bases. For a DNA strand of interest, DNA polymerase is added to synthesize a new strand of which each of the bases have been chemically tagged with a different fluorescent label. Depending on which signal is detected, a different nucleotide can be identified. To expand this to longer pieces of DNA, multiple segments are sequenced. Based on the overlapping characteristics of each read, accuracy on the true signal increases and the sequence can be analyzed in context of the hypothesis under investigation. Comparing the obtained sequences allows us, for example, to detect genetic variations between cohorts of people that could be environmentally, evolutionary or disease related.

  In Chapter 6, the spatial organization of the genome is analyzed. This is measured using **chromatin conformation capture** technology, i.e. single cell Hi-C [145]. More details are provided in the introduction of Chapter 6. Looking at the spatial organization, we can infer which elements of the chromatin are in contact and interact with each other and detect genomic regions.

- **Transcriptomics**

  The study of the RNA (ribonucleic acid) within a cell or organism, also known as expression profiling. The mRNA or messenger RNA is the most studied class of RNA. It is created by an enzyme that copies the gene's protein-coding DNA, after which it travels out of the cell's nucleus and into the cytoplasm. Here, the mRNA is read and its information is then used to create proteins by placing amino acids in the order defined by the

nucleotide bases of the RNA/DNA. The transcriptome is a snapshot in time of the transcript under investigation, and contains aside from mRNA also the other types of RNA. With multiple transcriptomic data sets, one can understand how genes are regulated in different tissues, conditions, time points, and diseases. Additionally, it can be used to understand the functionality of unannotated genes.

Two main technologies to study the transcriptome are: RNA microarrays and RNA sequencing [147]. **RNA microarrays** were developed first and measure the abundance of a set of transcripts. They allow for the simultaneous assay of thousands of transcripts using fluorescence labels. It uses probes that are arrayed on a solid substrate such as glass, and measures the hybridisation to these probes. They require some prior knowledge, i.e. an annotated genome sequence or a library of expressed sequence tags, on the tissue of interest.

More recently, similarly to genomics, the mRNA can be measured using **RNA sequencing** (RNAseq). Here, without prior information required, the mRNA is sequenced to generate the coding DNA (cDNA), this is also called reverse transcription. Next, the cDNA is fragmented, sequenced into short reads and mapped to the reference genome. The abundance of mRNA transcripts is then determined by counting the reads mapped to these sequences, providing the expression levels of the different transcripts. With next generation sequencing, RNAseq provides a better coverage and resolution of the transcriptome compared to microarrays, but is higher in cost.

- **Proteomics**

  The study of the structure and function of proteins within a cell or organism. Proteins are essential components in all organisms, and are formed by sequences of several amino acids. In total, there are 20 amino acids that can serve as possible building blocks to make up a protein. Proteins are structural elements of all organs and body tissue, in control of the chemical processes and signalling between cells. They are responsible for the metabolic activities in a cell and their study allows us to understand, for example, how diseases manifest themselves. This is unfortunately a difficult task, due to their diversity and their complex and dynamic behaviour. As proteins are prone to modification, there exist many more than solely those the genome encodes for. With proteomics, scientists aim to answer the still many unknowns regarding protein identification, localization, abundance, production, interaction, modification, composition, and activity. It is through the unravelling of all these unknowns that we are able to not only understand gene function

better, but also cure diseases, find diagnostic markers, make vaccines, and so on.

Technology-wise proteomics entails many different methods, each with their own benefits. To list them all would take us too far, but a comprehensive overview can be found in the work by Aslam et al. [14]. Proteomics can be analyzed by various techniques:

- **gel electrophoresis** [157]: separation of proteins using charged gel matrices,
- **chromatography** [166]: separation of proteins using chemical reactions in columns filled with specific substances, and
- **mass spectrometry** (MS) [7]: determination of molecule masses via ionization and movement in an electromagnetic field.

In this thesis, we will focus mostly on following spatial proteomics for exploratory analysis and localization of proteins: untargeted **mass spectrometry imaging**, as described in detail in Section 1.2, and targeted **immunohistochemistry** as explained in Chapter 5.

- **Metabolomics**

  The study of metabolite profiles - including lipids, drugs, and other small molecules - in a cell or organism under a given set of conditions. Metabolites are the final component in the interconnected pathway of the omics family. Metabolites influence observable phenotypes directly, making metabolomics a key technology in the evolution towards **precision medicine**. They help us characterise diseases, develop therapies and allow for the discovery of biomarkers used to diagnose or monitor patients. Computationally, analysis of metabolomics data is challenging because of the different characteristics of all molecules measured, i.e. polarity, solvability, acidity, and so on.

  Mass spectrometry [51] allows for the simultaneous detection and quantification of thousands of metabolites present in a sample. MS, in combination with **liquid chromatography** (LC-MS) [225], is an important metabolomics method and provides good molecular coverage and sensitivity. However, it lacks spatial information regarding in which cell of which part of the tissue the metabolite is present. Spatial metabolomics, such as **mass spectrometry imaging** [10, 138], provide more detailed information on the organization of metabolites in a biological sample.

- **Multi-omics**

  This term describes the study of a cell or organism using a combination of two or more omics modalities. This could be either using separate

technologies and merging the data computationally, or from novel technologies allowing for simultaneous measurement of multiple omics. Fuelled by the flow of information and inter-omics relations, multi-omics is emerging rapidly as one of the methods to watch. By using multiple modalities, the confidence increases on the detected signals. Additionally, it helps us map out the complex interactions expected to be present in organic tissue. An example of a spatial multi-omics is shown in Figure 1.2, for more details we refer to the work by Zhang and Spotbeen et al. [221].

Multi-omics are the future in the field of omics research, but it requires both technologically and computationally more advanced tools. With the combination of different measurements, additional challenges arise. First of all, the correct alignment of the modalities is a complex matter: spatially there could be a difference in resolutions or shape when subsequent tissue slices are being used, or molecularly the data sets could have different molecular ranges or sensitivities. Secondly, difference in artifacts and noise patterns across the data requires intelligent preprocessing techniques to clean the data and make it comparable for joint analysis, unifying scaling and normalization. Additionally, as single omics analysis often is already a computational burden, multi-omics is even more challenging and must rely on efficient coding and advanced hardware for both storage and analysis. To conclude, omics experiments require high-end technologies taking a substantial expense out of the budget of research labs. As no lab can study *all* the omics modalities, moving to multi-omics studies requires collaborations across labs, with clear communications regarding data handling, processing, transport, etcetera. The computational, infrastructural and logistic side of research all impose additional difficulties when conducting multi-omic research.

Aside from a better understanding of the biological processes in all living things, and a pathway to better medicine, omics data and good computational models provide us even more. It allows for **efficiency**: either in the lab or in the hospital, high-throughput technologies make it possible to measure more and faster. This means less waiting for results, applying omics in surgical settings, generalizing findings on bigger cohorts of data, and so on. The importance of omics in future research was highlighted with the recent COVID-19 pandemic. With the use of omics technologies and combined effort in labs worldwide, the virus was analyzed for its genomic fingerprint and vaccines were developed at a speed never seen before [16, 176].

Omics technologies allow us to think even bigger and tackle challenges we did not even believe possible. A well-known example being the Human Genome Project [79], an international collaboration completed in 2003 mapping out the

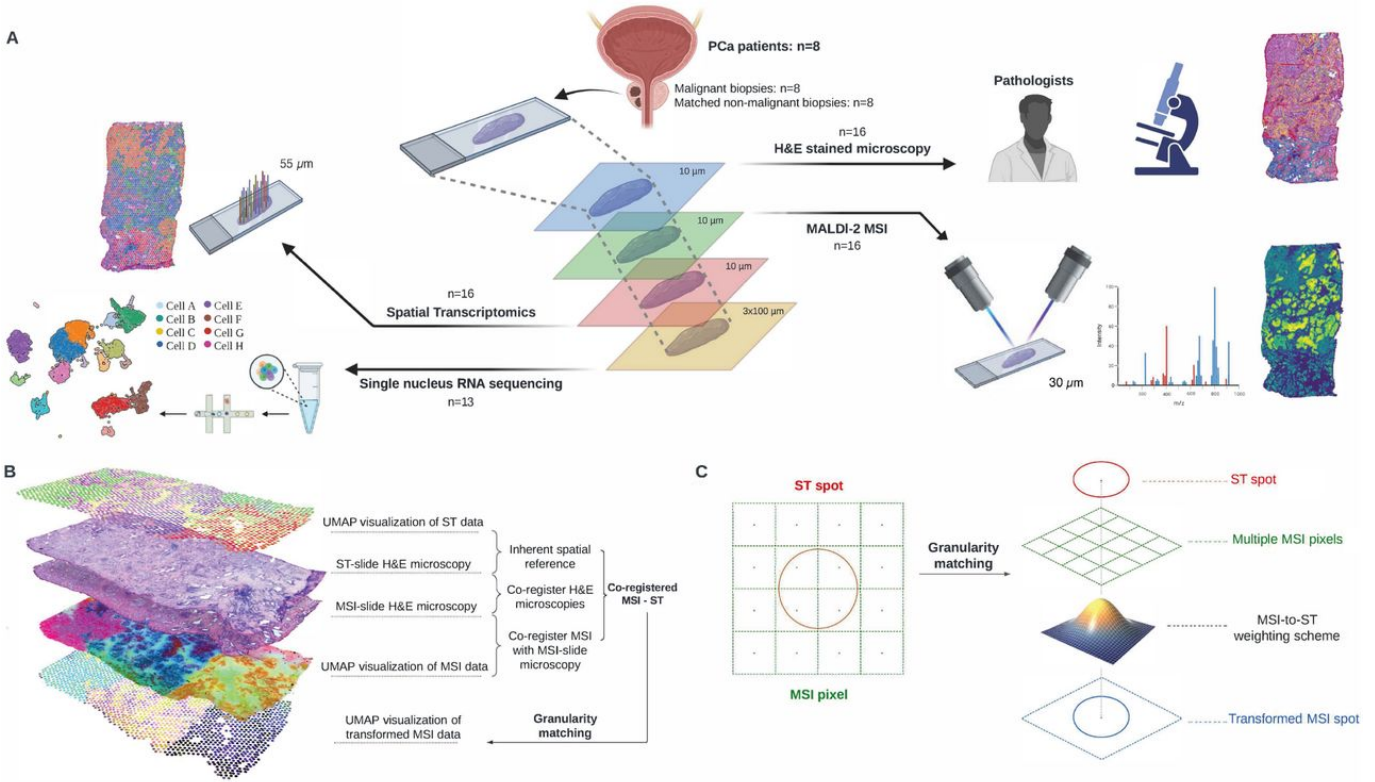


Figure 1.2: **Study design and overview of the spatial multi-omics integration pipeline (SMOx).** **A.** Schematic representation of the various data sets generated on prostate cancer samples and the applied integrated spatial multi-omics workflow. **B.** Strategy for co-registration and integration of different spatial multi-omics data layers generated on consecutive tissue sections. To link ST data with MSI data, UMAP visualizations of different modalities were co-registered in two steps based on the haematoxylin and eosin (H&E) stainings of the respective slides and MSI data, and a UMAP visualisation of transformed MSI data was generated after granularity matching. **C.** Overview of the granularity matching step to bridge the scale gap between MSI and ST data. Spatial readouts were aligned, and Gaussian weighting was applied to generate a transformed MSI spot. Reprinted from [221] with permission under a CC-BY-NC-ND 4.0 International license.

full DNA sequence of the human genome. Who knows what will be possible in the future using omics technologies - will cancer become a disease from the past?

Additionally, aside of the medical applications, omics can also be applied to material sciences and other biological research [17]. Think of agriculture where genetic modifications are used for healthier crops. All living things can be analyzed using omics technologies to understand their functioning or evolutions better.

As mentioned above, we need and will keep needing advanced data analytical tools to comprehend the data streams produced by omics technologies. It is **big data** and will become even bigger moving onto better resolutions and multi-omics, making it infeasible for manual inspection. Advanced algorithms offer more information, but also reduce the cost of creating more complex technologies. If a hypothesis can be answered computationally, it requires often less time, money and people than in the wet lab.

To conclude, omics data rely on advanced technology and algorithms. It is an interdisciplinary area of research designed to understand the flow of information from genes to phenotypes and is ever so crucial in our modern society.

1.2 Mass Spectrometry Imaging

The omics modality mainly used in this thesis is spatial proteomics, which is used to measure the spatial distribution of proteins, lipids or metabolites within a slice of tissue. This can then help us understand the biochemical processes taking place and lead to better understanding of disease progression and aid in the associated drug development. This work focusses on the development of analytical methods for **Mass Spectrometry Imaging (MSI)** as a tool for spatial proteomics, lipidomics and metabolomics. This section will elaborate more on the different types of MSI and discuss the various ways one could analyze the data, starting with a short recap on the basics of MSI.

1.2.1 The basics of MSI

In standard **mass spectrometry (MS)** [133, 181, 163], a molecular profile of a sample is obtained by ionization of the molecules present in the sample, often mixed with a solvent or matrix. The charged ions travel through an electromagnetic field where they are separated based on their **mass-to-charge (m/z) ratio** expressed in Dalton (Da). While MS has excellent molecular

coverage and sensitivity, it does not provide any information on the spatial abundances within the sample. **Knowing where the different molecules are localized in a biopsy is essential to understand how tumours and other cells behave in their micro-environment.**

MSI extends standard MS and measures both the spatial and molecular distributions by applying MS at multiple locations throughout the sample. To conduct an MSI experiment, multiple procedural steps have to be executed. Figure 1.3 gives an overview of a full MSI workflow.

Before MSI analysis, the sample has to be prepared in a rigorous manner, see Figure 1.3A. To minimize batch effects and other artifacts, it is important that samples are collected, frozen or embedded, and sectioned according to protocol. Similarly, matrix application or other sample applications for ionization should be applied according to the standards. This allows for comparable research between different experiments and across labs [82, 211].

Next, the tissue is analyzed by performing MS across the sample in a grid-like fashion, see Figure 1.3B. The distance between subsequent pixels in the grid determines the spatial resolution of the experiment. At each pixel, a laser beam or another ablation technique is applied to ionize the molecules and measure their m/z ratio. This results in a **mass spectrum** per pixel, showing the intensities of the full m/z range measured.

With the m/z spectra obtained for each pixel, heat maps or **ion images** can be produced showing the spatial distribution of a specific m/z value throughout the tissue. These ion images are very informative, but since there are often several thousands of m/z values measured, checking each image manually is time consuming. To this end, computational methods will be applied to reduce the dimensionality while capturing as much information as possible (see Figure 1.3C). After the necessary preprocessing steps, several methods can be applied towards visualization, clustering, registration, and identification of the MSI data. Section 1.2.4 will elaborate more on the main trends in computational analysis of MSI data.

To further identify a molecule, either its detected mass is compared to an atlas of known molecules or a subsequent MS (tandem MS or MS/MS) experiment is performed [127, 91]. When breaking down a protein into its amino acid building blocks, its sequence can be derived. The protein's identification will then be determined by matching this sequence to existing libraries, such as UniProt [187].

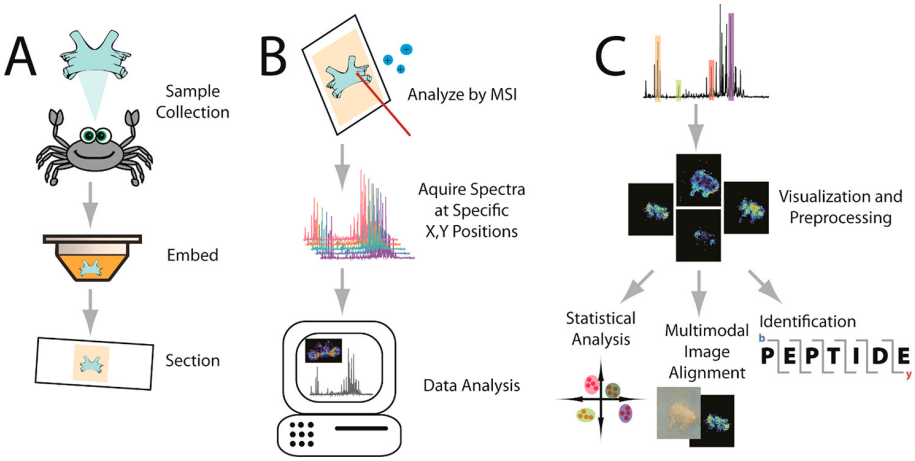


Figure 1.3: **Visual workflow for the MSI analysis.** A crustacean's brain is used as an example tissue for this workflow. **(A) Sample preparation**. After collection from the animal, the sample is embedded in a supporting medium (microscopy) or frozen (MSI) for sectioning onto slides. Other sample processing, such as applications of enzymes, matrix, or derivatization agents, may be performed depending on the molecular species of interest or the instrument being used. **(B) Sample analysis**. After acquiring a spectrum at each (x,y) grid point on the tissue, sophisticated software tools are used to process and visualize the data. A laser is used to ionize molecules as depicted, although several nonlaser based methods are also used. **(C) Data processing**. After preprocessing the data (e.g., baseline correction), the distribution of selected molecules can be visualized. From there, identification of the m/z values and statistical analysis between different images or image coregistration with other image modalities can occur. Reprinted with permission from [29]. Copyright 2018, American Chemical Society

1.2.2 MSI technology

MSI technology knew a difficult start since its first introduction in 1967 [121] up until the late 1990s. Scientist were sceptical in the beginning and challenging data analysis, lack of standardization and the limitations of the first MSI devices resulted in a slow adoption of MSI [210]. It is due to the simultaneous boom of MS technology, computing resources and bioinformatics in the last decades that MSI has gained popularity in the field of omics research [10]. The now ever improving specs of MSI tools, such as timing and resolution, allow MSI to provide even more detailed information for molecular analysis and to be of assistance in surgical settings. Being listed as one of Nature's methods to watch in 2018 [57] and spatial (multi-)omics as technology to watch 2022 and 2023 [65, 66, 185], MSI is as relevant as ever.

MALDI-TOF MSI

The type of MSI used throughout this dissertation, is **matrix assisted laser desorption-ionization time-of-flight** (MALDI-TOF) MSI, first introduced in 1997 by Caprioli et al. [36] and most commonly used in the field [54]. Here, we will go over the technical details of MALDI-TOF MSI technology, as shown in Figure 1.4.

MALDI refers to the **ionization technique**. First, a thin chemical matrix, consisting of crystallized molecules, is applied to the tissue section. This can

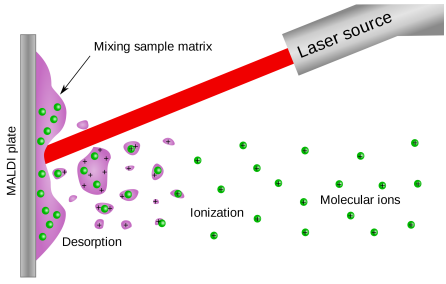


Figure 1.4: Schematic overview of MALDI ionization of a slice of tissue. After application of a chemical matrix, a laser (often UV) irradiates the mixture. This causes ablation of both matrix and sample molecules, and is followed by desorption and ionization of the sample by charged after which the charged molecules will be sucked into the electromagnetic field of the mass analyzer. Copyright Mikayé, CC BY-SA 3.0, via Wikimedia Commons.

be done by for example pneumatic sprays or sublimation [192]. Some popular matrices used are: alpha-cyano-4-hydroxycinnamic acid (CHCA), sinapinic acid (SA), or 2,5-dihydroxybenzoic acid (DHB) [117]. Next, a laser beam will irradiate the matrix covered tissue section. This causes desorption and ionization of the sample by the matrix crystals and will result in the charged molecules to be detected by the mass analyzer for measurement of the m/z ratio. To improve the sensitivity, recent developments use a secondary ionization, called MALDI-2. In terms of resolution, MALDI typically spans tens to hundreds of micrometers (μm). However, leveraging transmission mode geometry has propelled its capabilities, now enabling resolutions as fine as 600 nanometers (nm) [139].

TOF refers to the **mass analyzer** used. After ionization, the charged molecules are led through an electromagnetic field with at the end a detector that will record the time between the ionization and the arrival of the molecule. The TOF mass analyzer will separate ions according to their m/z ratio, implied by their *time of flight* in the analyzer: big molecules will travel slower than small ones and double charged ions will travel faster than single charged ones. However, it is possible for different molecules to have the same m/z ratio. In such cases further identification using MS/MS might resolve the ambiguity. To further improve the separation, TOF analyzers also exist in reflective mode to prolong the distance to the detector.

Other types of MSI

Aside from MALDI-TOF, a wide range of ionization techniques and mass analyzers are available. Figure 1.5 shows an overview of the different ionization techniques for MSI, and compares them across the main criteria for MSI technologies: sensitivity, spatial resolution and mass range. Sensitivity, expressed in attomoles (amol), depicts the capability of MSI technologies to detect and measure extremely low concentrations of analytes, often down to the scale of 10^{-18} moles. Well-known ionisation techniques, aside from MALDI, are DESI (desorption electrospray ionization), SIMS (secondary ion mass spectrometry), LESA (liquid extraction surface analysis), and LAESI (laser ablation electrospray ionization) [213].

In **DESI**, an electrically charged solvent mist is sprayed on the tissue causing ionization. While it lacks good spatial resolutions due to the spray size, it can be applied at ambient pressure and does not require the use of a laser. There is thus no need for matrix application, resulting in little to no sample preparation. Recently developed nano-DESI improves the spatial resolution by using finer probes, yet the method remains at micrometer resolution [213].

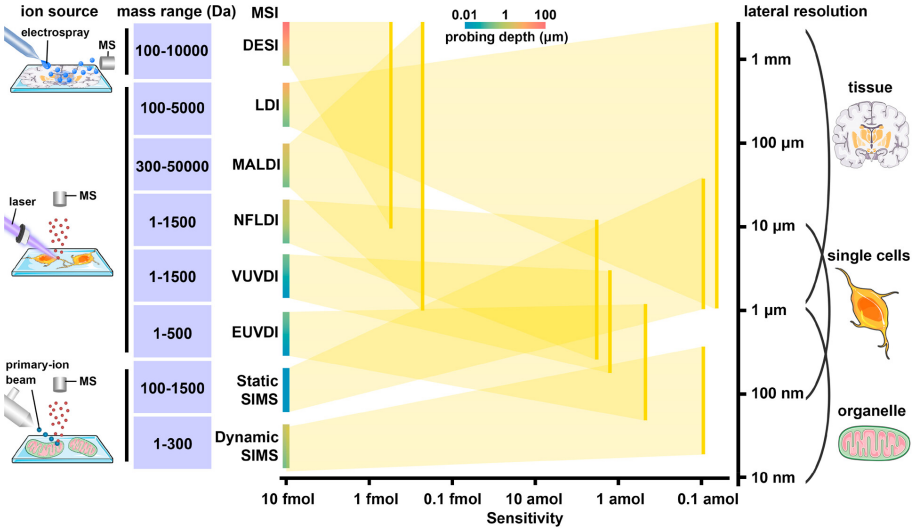


Figure 1.5: Overview of the different types of MSI technologies and how they differ in terms of ion source, mass range, sensitivity, probing depth and spatial resolution. Copyright 2023 Xing et al., Co-published by Nanjing University and American Chemical Society, CC-BY-NC-ND 4.0 [213].

SIMS, being the oldest of the MSI technologies, provides the best spatial resolutions, but again, the advantage comes at a cost. In SIMS, a focused ion beam is released on the tissue, causing a release of ionized molecules from the sample (a secondary ion beam) towards the mass analyzer. The high energy ion beam causes fragmentation of larger molecules and thus limits the mass range for which it is suitable. Yet, the spatial resolution of SIMS is so high that it can analyze samples on a subcellular level down to 50 nm.

LESA, omits the need for a laser by using a solvent and an electrospray capillary to extract the molecules from the surface in ambient atmosphere. It is widely applicable and sensitive, but has poor spatial resolutions, typically around 1 mm. New extensions such as μLESA now allow up to 400 μm.

LAESI involves the direct desorption and ionization of molecules from the surface of a sample using a laser. Without the application of a matrix, ionized molecules ablate under impact of the laser and are intercepted by an electrospray plume, where the ions are solvated. Also here, little to no sample preparation is required. While LAESI reaches similar spatial resolutions as MALDI, it remains less popular in the field of MSI.

Aside from the different ion sources, there are also different types of mass

analyzers and even combined or tandem analyzers. Commonly used alternatives to the TOF analyzer are ion trap, Fourier transform ion cyclotron resonance (FT-ICR), and quadrupole mass analyzers. All of these aim for the best sensitivity, but selecting the most suited analyzer depends on the ion source and type of problem. More details can be found in the extensive review by Li et al. [119].

1.2.3 Advantages and limitations of MSI

This section provides a concise summary of the main advantages and limitations encountered when using MSI.

Advantages

Spatial information

Compared to standard mass spectrometry, MSI provides additional information about the spatial distribution of the analytes within a slice of tissue. This makes it possible to map the molecular organization and get a better understanding of the biochemical processes at play.

Non-destructive

Both application of the matrix and the ionization technique cause only minor alterations to the tissue [63]. This allows for multiplexing, i.e. simultaneous measurement of multiple analytes in a single experiment. After performing MSI, other omics techniques can be applied on the tissue or it can be stored in case further analysis is required at a later time.

Untargeted

MSI does not rely on prior information to detect the molecules. If they ionize, they are measured. MSI has no restrictions except the predefined m/z range for the experiment. It is thus an excellent tool for exploratory analysis and biomarker discovery.

Limitations

Spatio-molecular resolution trade-off

When the spatial resolution improves, the ablation area (the part of the tissues that becomes ionized) will decrease. This smaller surface will have fewer molecules which will result in a lower sensitivity or signal-to-noise ratio (SNR). Additionally, there is a trade-off with the speed of the acquisition: a faster experiment might come at the cost of a lower spatial

resolution, a lower SNR or a smaller molecular resolution or range. In an ideal world, all three parameters are tackled to enhance performance, but unfortunately this will lead to additional challenges in terms of data storage and analysis [29, 139].

Lack of standardization

The lack of standardized procedures across labs, experiments and processing of the data has lead to a suboptimal level of reproducibility, a limitation well-known within the field [82, 211]. This is why experts are **calling for a unified approach in MSI experimentation**. There are numerous benefits of a better standard in MSI, e.g.: comparable research, higher quality AI models, less bias, approval of MSI by health authorities, user friendly software, explainable results, integration with other omics modalities, etcetera.

Data storage and analysis

Due to the untargeted approach, the obtained MSI data sets are often very large. The writing, storing and calling of the measurement data, has a big influence on both acquisition and computational timing. Aside from the need for high-end computer hardware, algorithms for data analysis should be able to cope with the **curse of dimensionality**. Additionally, MSI data can contain a lot of noise and artifacts. This can originate from dust in sample preparation, detector noise, or biological noise such as chemical isotopes highly present in the bigger molecules. This raises the need for tailored approaches in preprocessing and analysis: able to handle low SNR and small nuances in the data. This way, we could learn even more from the mountain of information provided by a single MSI experiment [83].

1.2.4 MSI data analysis

Because of the high number of molecules measured in every pixel of the sample, an MSI experiment can easily generate gigabytes to terabytes of data. As mentioned above, efficient data storage and analysis is required to fully exploit all the information MSI provides. In this section, we will discuss the different steps in MSI data analysis, as shown in Figure 1.6. This includes the main trends in preprocessing and analysis using both supervised and unsupervised machine learning.

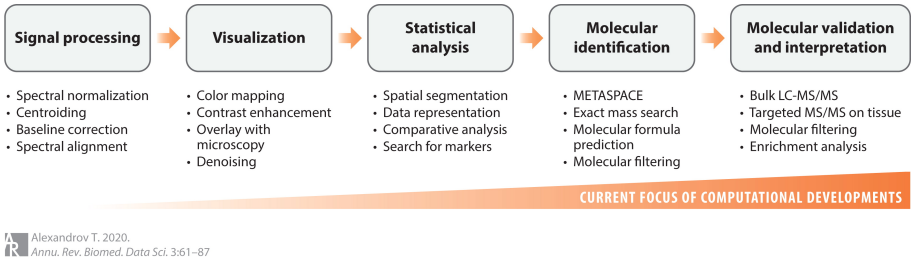


Figure 1.6: Steps of a typical data analysis workflow in MSI: preprocessing, visualization, statistical analysis, molecular identification, and molecular validation and interpretation. The main focus in MSI data analysis is moving more and more towards identification and validation of the molecular profiles. Figure reprinted with permission from [10]. Copyright (2020) Annual Reviews.

Preprocessing

Preprocessing MSI data mostly refers to the spectral domain, and includes: baseline correction, binning, mass alignment, peak picking, and normalization. With good preprocessing techniques, artifacts from lab specific sample preparation can be reduced. This allows comparison between different labs and provides a more robust starting point for further downstream analysis. More on the preprocessing of MSI data can be found in the reviews by Rafols et al. [158] and Hu and Laskin [98]. Here, we will elaborate on normalization and peak picking of MSI data.

- **Normalization**

  Normalization is the first and foremost step in MSI data preprocessing. It aims to correct the data for artifacts resulting from matrix effects or experimental variations in the data acquisition, both within one and between multiple experimental setups [98]. The normalized spectra depict scaled intensities which are more comparable across all pixels. The most used normalization approach in MSI are total ion current (TIC) and root mean square (RMS) normalization. In TIC normalization, each intensity value within a spectrum (pixel) is divided by the sum of all intensities of that spectrum (pixel). Similarly, using RMS normalization each intensity value within a spectrum (pixel) is divided by the root mean square of all intensities of that spectrum (pixel). In recent years, however, the discussion arises for a new normalization standard [46]. Different new approaches have been proposed, such as median fold change (MFC) normalization [24] and intensity profile normalization (IPN) [204], and

normalization of MSI is still actively being researched. Of note, the KU Leuven has an international patent (Van de Plas R., De Moor B., Waelkens E.) on intensity normalization of MS data where peaks demonstrating high variabilities in intensity are omitted. Unfortunately, due to differences in sample preparation across labs and the lack of a ground truth, objective comparison through benchmarking is extremely challenging. This makes that the shift towards a better normalization standard is advancing rather slowly.

- **Peak picking**

  Depending on which mass analyzer is used, the mass resolution can range from 10^4 for TOF to up to 10^6 for the most recent FT analyzers. To reduce the spectral dimension, and thus also the data size, peak picking or spectral centroiding can be applied. Here, associated m/z values are reduced to their central peak (the centroid). Popular approaches are Gaussian functions and wavelets [98]. After this the range of m/z values is reduced to those selected by the peak picking algorithm. This will reduce computational cost and storage while preserving the important features of the data. However, for exploratory analysis, peak picking may cause potential values of interest to be over looked [144].

After preprocessing, the data is often more compact and comparable. Nowadays, successful and generalizable preprocessing strategies are incorporated in vendor software. However, **more complex phenomena, categorized as technical or chemical noise, are still challenging to overcome in MSI data processing**. This makes MSI preprocessing still an active area of research and requires further analytic approaches to be able to cope with data impurity and sample preparation effects (technical noise), and MSI related characteristics (chemical noise: isotopes, two charged states, matrix effects, adducts).

Supervised vs. unsupervised analysis

In the earlier years, downstream analysis of MSI was mostly limited to basic statistics and principal component analysis (PCA) [195, 103]. Since then, the world has seen an amazing evolution in terms of computational possibilities. Specifically for MSI data analysis, Figure 1.7 shows a general overview of the most used algorithms together with example applications in recent papers.

Whether to perform supervised or unsupervised analysis depends on the availability of prior information, such as pathological annotations. When labels are available, supervised analysis is often used for the discovery of new biomarkers. While this is a key goal of performing MSI technology, expert

annotations are time consuming and scarce. This is why biomarker detection is also done by an exploratory - unsupervised - analysis for the detection of m/z values of interest. This requires, however, a subsequent, targeted analysis of the m/z values to confirm the findings. For both supervised and unsupervised methods, the choice of algorithm relies on preference, the type of problem and the available data set, with new approaches emerging continuously.

- **Supervised MSI data analysis**

  Supervised analysis of MSI data utilizes **a priori annotated labels** to differentiate between different biological regions within a sample or different pathologies across samples. We differentiate between **regression**, a method used to examine the relationship between one dependent variable and one or more independent variables, and **classification**, used to assign items into predefined categories based on their features. When a model is trained for classification, new samples can be labelled in **assistance of patient diagnosis**. Additionally, further analysis of the distinct features amongst the different classes can lead to **biomarker discovery**. Where as classification aims to differentiate between two distinct cases (e.g. tumour vs normal tissue), the to be determined values in regression analysis are of continuous nature (e.g. the degree of the disease) [98].

  Despite their good performances in the specific pathological settings, the **generalization of supervised MSI data analysis is challenging** due to the lack of ground truth and limited amount of expert annotations. Even when labels are available, it is difficult to create balanced training, test and validation sets, as the pixels in MSI data sets exhibit spatial correlation. Furthermore, the newer deep learning methods are black box and require secondary approaches to detect the m/z values of interest for further biomarker discovery.

  Below, we discuss regression and classification, the main approaches for supervised MSI data analysis and provide examples from recent literature. Note that some methods could be used for both regression and classification. For more details, we refer to the mentioned articles and the review by Hu and Laskin [98] and Jetybayeva [102].

  Regression models can be used to model MSI samples on a continuous scale. Regression of the m/z spectra can be used both for diagnostics as in certain preprocessing schemes [98]. Popular for class prediction are regression approaches such as: Lasso regression [205], ridge regression [205], and partial least squares (PLS) [209, 101] method, sometimes utilized in combination with discriminant analysis [53]. With these models, statistically significant m/z peaks are detected and used for further classification.

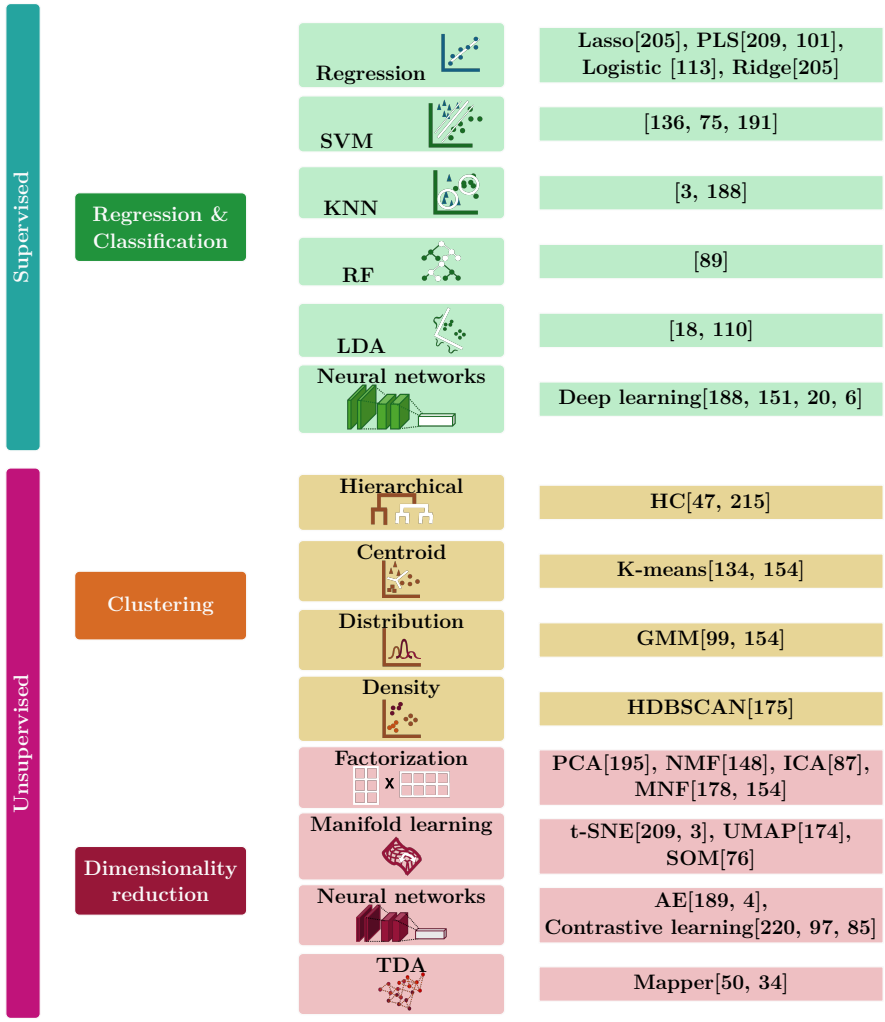


Figure 1.7: Overview of popular algorithms used for the supervised and unsupervised analysis of MSI data, including references to recent literature. Figure inspired by [56].

Classification models aid MSI data analysis towards the annotations of new samples and the detection of class specific features or biomarkers (e.g. using SHAP values [190]). Popular classification approaches include: linear discriminant analysis (LDA) [18, 110], logistic regression [113], recursive maximum margin criterion (RMMC) [191], support vector machines (SVM) [136, 75], random forests (RF) [89], k-nearest neighbours (kNN) [3, 188], gradient tree boosting [227], and other (deep) neural networks [188, 151, 20, 6]. Often these classification approaches are used to determine pathologies of unseen samples, based on previously annotated data of a specific type of tissue.

- **Unsupervised MSI data analysis**

  In the case of unsupervised analysis, no labels are used. Here, **dimensionality reduction** methods can be used to compress the data set for storage or further computations. Another aim could be the visualisation or spatial segmentation of the sample by **clustering the pixels based on their spectra**. Providing a false color image indicating the different spatial regions in the sample could be useful in surgical settings (in line with classical pathology) or for multimodal analysis. Alternatively, one could be more interested in the molecular fingerprints in the sample and perform **clustering of the ion images or *m/z* values**.

  Here, we will go over the most used algorithms for unsupervised MSI data analysis, but for more information we refer the reader to recent review papers [201, 98, 102].

  Clustering MSI data, as mentioned above, towards a spatial segmentation is useful in pathology and multimodal alignment. As for manifold learning approaches, the different clusters can be used to create a false color image of the sample. Also here, different frameworks exist. Often used for MSI data analysis are: k-means [134, 154] (**centroid** based), HDBSCAN [175] (**density** based), Gaussian mixture models or GMM [99, 154] (**distribution** based) and **hierarchical** clustering or HC [47, 215]. While these clustering approaches give good results, they are often less scalable. To accommodate for this, a dimensionality reduction is done before applying the clustering algorithm. Doing so, however, might already cause the clustering results to miss certain features [98].

  Dimensionality reduction, as shown in Figure 1.7, encompasses a variety of algorithms, e.g. factorization methods, manifold learning, deep learning, and graph learning.

- **Factorization** methods are a popular technique for compression, visualisation and clustering of MSI data. Well-known examples are: principal component analysis (PCA) [195], non-negative matrix factorization (NMF) [148], minimum noise fraction (MNF) [178, 154] and independent component analysis (ICA) [87]. The purpose of factorization methods is to capture the main trends in the data by decomposing it into a product of two (or more) low rank matrices or tensors. The resulting factors can exhibit specific characteristics by adding different constraints, such as: high variance, non-negativity, optimal SNR or independently distributed, etc. A drawback here is the prior selection of the number of components or rank r that gives the best representation of the data.
- MSI data is inherently non-linear. **Manifold learning** exploits this non-linear nature by projecting the structure of the data in their high dimension onto a lower dimensional manifold. This often results in a better representation of the local relationships between the datapoints. Some well-known examples are: t-distributed stochastic neighbour embedding (t-SNE) [209, 3], uniform manifold approximation and projection (UMAP) [174], and self-organizing maps (SOMs) [76]. To improve results, spatially aware clustering incorporates spatial correlations of the pixels [11] While these methods provide good hyperspectral imaging results, it is more difficult to determine which molecular characteristics contributed to the clustering of the different pixels.
- Aside from traditional machine learning approaches, increasingly popular are (unsupervised) **neural networks** and deep learning architectures. By compressing the data step by step in multiple layers, different features are extracted each time, resulting in a compressed latent space describing the data. Applied to MSI data analysis, (variational) autoencoders (AE) [189, 4], self-supervised learning and contrastive learning [220, 97, 85] provide excellent results for the compression of MSI data, peak learning, and detecting the co-localization of m/z values.
- Newer in the field are graph [52, 122] and **topological data analysis** (TDA) [50, 34] approaches. The key concept of TDA is to look at the data in a higher dimensional graph and use topological properties, such as shapes and patterns, to cluster the data points rather than distances. A popular approach here is the use of Mapper, which projects overlapping bins of connected components in the high dimensional space into a lower dimensional graph representation by applying filters. In this

graph, the nodes are the clusters and the edges are shared data points connecting the clusters.

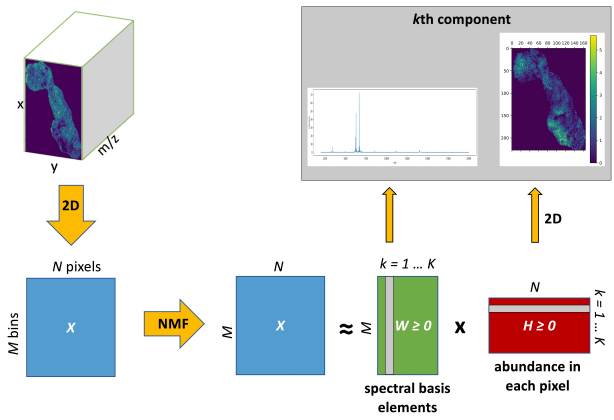


Figure 1.8: Graphical representation of non-negative matrix factorization. More details in Chapter 2.

This thesis focusses on the application of NMF to various omics modalities. We utilize this unsupervised dimensionality reduction technique for the fast and understandable analysis of these data sets, exploiting the non-negative nature of omics data. NMF offers a fast and scalable framework for the visualisation, low dimensional representation and clustering of the features. It allows for interpretable factor matrices and provides useful insights in the co-localization of m/z values, and detection of novel m/z values of interest. A schematic representation is given in Figure 1.8 and more details on NMF for MSI data analysis can be found in Chapter 2.

1.3 On the islets of Langerhans in mice pancreas

The first part of this dissertation (Chapters 2, 3 and 4) describes several algorithms for the analysis of MALDI-TOF MSI data of healthy mice pancreas. To ensure non-experts a fluent reading, this section briefly introduces some biological background on the islets of Langerhans in mice pancreas.

Before MSI existed, and still a gold standard in pathology today, **hematoxylin and eosin (H&E) staining** allowed for the spatial analysis of a biopsy, often used in cancer diagnosis. The H&E staining colours the tissue in such a way that

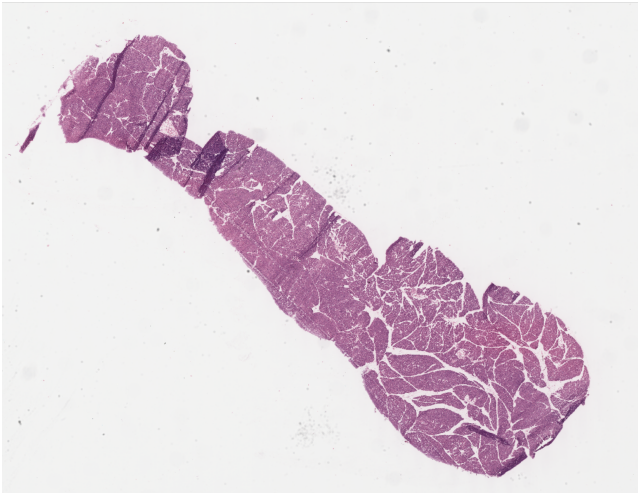


Figure 1.9: H&E staining of healthy mouse pancreatic tissue.

it enhances contrast between the different anatomical elements. Hematoxylin gives the cell nuclei a purple colour, while the eosin reacts with the extracellular matrix and the cytoplasm, dying it different shades of pink [71, 38].

After staining, the slice of tissue is analyzed using a microscope. Through patterns in the colouring, cell structures become visible and allow for a pathologist to identify the anatomical composition on a subcellular level. An example of a stained microscopic image of a healthy mouse pancreas is given in Figure 1.9.

The islets of Langerhans represent the endocrine part of the pancreas and are responsible for the production of insulin, which regulates the blood glucose levels. Figure 1.10 shows a zoomed-in section of the same H&E staining containing an islet. It displays the typical circular organization of the islets found in rodents and the difference in cell-shape compared to the surrounding region.

The islets in mice are composed of an inner core containing the insulin-producing β-cells and an outer mantle where the glucagon-producing α-cells reside, see Figure 1.11F [179, 44]. Aside from insulin and glucagon, other hormones such as somatostatin (δ-cells), ghrelin and Polypetide P (PP-cells) are secreted by the islets as well.

The exocrine region of the pancreas produces enzymes responsible for the digestion of carbohydrates, nucleic acids, proteins and fatty acids upon entering the digestive tract through the duodenum [104]. Although there is no direct

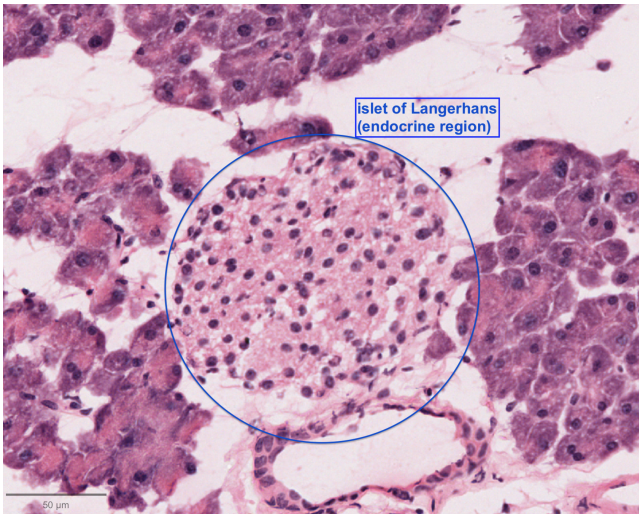


Figure 1.10: H&E staining of healthy mouse pancreatic tissue zoomed-in (scale $50\mu m$) on an islet of Langerhans (endocrine region) showing a clear circular shape and different cell structures than the exocrine region surrounding the islets.

interaction between the two, the exocrine and endocrine regions are our bodies' first response to the food entering our intestines and essential for our understanding of diseases such as diabetes and pancreatic cancer.

The pancreas is mostly studied in rodent models, but to extrapolate the obtained knowledge directly to the human system would be prone to inaccuracies. There are numerous differences between humans and rodents regarding pancreas structure and functioning [58]. First of all, there is a notable difference in islet organization, where in humans there is no distinct outer periphery (Figure 1.11E-F). Secondly, rodents have two types of insulin, encoded in two different genes, whereas humans have only one. To translate findings from analysis of rodent pancreas tissue to that of humans, one should be aware of the known and possible unknown similarities and differences between both species.

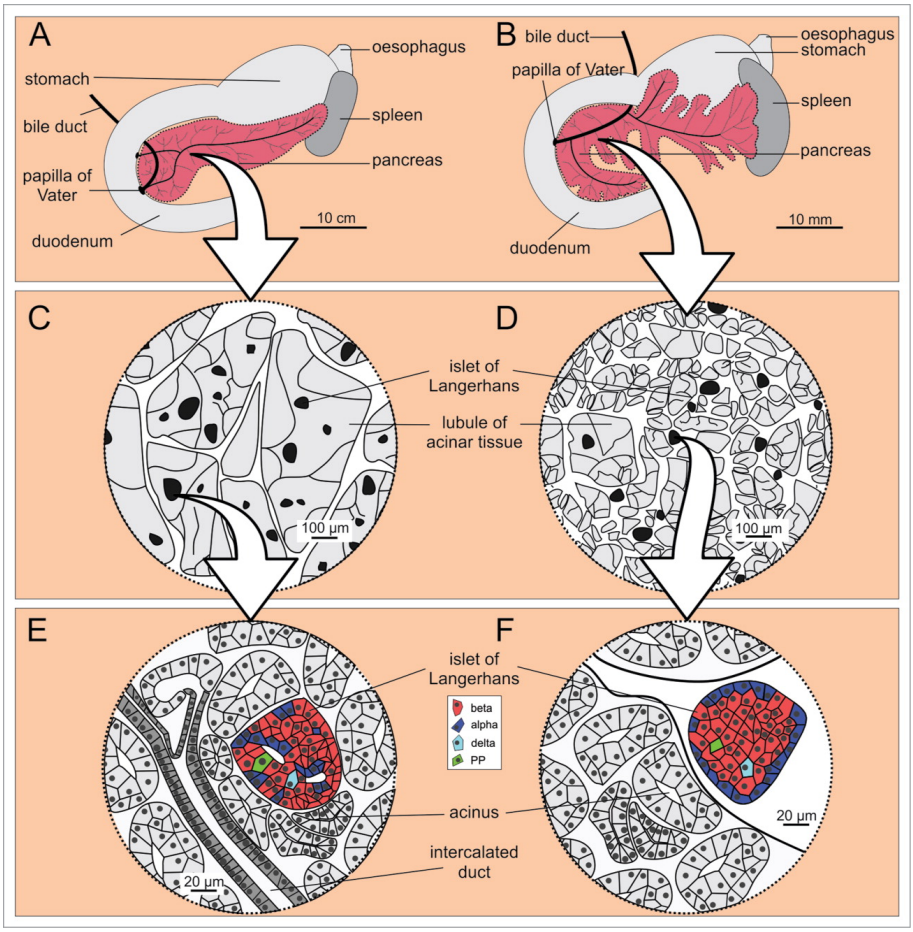


Figure 1.11: **Microscopic anatomy of the human (A, C, and E) and the mouse (B, D, and F) pancreas.** (A and B) Macroscopic anatomy of the human and the mouse pancreas, respectively. (C and D) Magnifying a portion of pancreas reveals larger lobules in humans when compared to mouse, whereas the islets of Langerhans are of fairly comparable size in humans and mice. (E and F) Cell composition and location of the islets of Langerhans within the pancreas are markedly different in the 2 species. Note the more diffusely distributed endocrine cells in humans (E) and the mantle-core pattern in mice (F). [58] Copyright 2015, reprinted by permission of Informa UK Limited, trading as Taylor & Taylor & Francis Group, http://www.tandfonline.com.

1.4 Thesis outline and main contributions

Part I

The first part of the thesis focusses on computational approaches for the analysis of MSI data, with a focus on mice pancreatic tissue.

Chapter 2 discusses a mathematical comparison of four NMF related methods for the analysis of MSI data. The methods under investigation are standard NMF, Kullback-Leibler NMF, probabilistic latent semantic analysis and latent Dirichlet allocation. It aims to demonstrate how the different methods rely on different assumptions of the data and how they use different algorithms to solve the factorization. By comparing all four methods on three proteomics and four lipidomics data sets, it allows for a comparison of the four methods. With multiple evaluation criteria, we demonstrate the benefit of KL-NMF for MSI data analysis, as this was not yet done in previous literature. The key take-away is how each method's framework already imposes a difference on the result and that this should always be taken into account when applying them.

Chapter 3 differs from the others as it applies topological data analysis using Mapper rather than NMF. More specifically, we demonstrate how the Mapper algorithm allows for a more refined hyperspectral visualisation for mouse pancreas MSI data sets compared to UMAP. An extensive parameter tuning is done to find the best result. When performing Mapper with the optimal settings, a ring structure around the islets of Langerhans is detected. This might be related to the islets' known circular orientation in rodents.

Chapter 4 presents a novel MSI hierarchical bi-clustering technique, based on a recursive rank-2 NMF (rr2-NMF) decomposition. Using NMF's linearly additive property, the data is systematically split in subsets, allowing for a more localized and detailed analysis - both spatially and spectrally. When the stopping criterion is reached, the end clusters show interesting patterns in both the pixel and m/z domain. Applied to the same healthy mouse pancreatic samples, again a circular pattern in the outer periphery of the islets arises. Exploiting the benefit of using an NMF based approach, the spatial patterns are now accompanied by a spectral signature allowing for new insights regarding the biochemical processes at play. Most interestingly, a different concentration pattern is detected between the two types of insulin present in rodents.

Part II

The second part extends the usefulness of NMF to other omics modalities, such as: mIHC spatial proteomics, and single cell genomics..

Chapter 5 applies an NMF-autoencoder framework for the fast and automatic clustering and annotation of MILAN immunohistochemistry data (targeted spatial proteomics). Applied to human melanoma samples, it shows to be computationally efficient to a (sub-)cellular level, omitting the need for sampling and extrapolation as done in earlier work.

Chapter 6 investigates how NMF and its higher order variant, non-negative tensor factorization, can benefit the analysis of single cell genomics data. More specifically, we apply factorization techniques to annotated chromosome conformation capture data in order to detect relationships between the three-dimensional organization of the chromosome and the cell-type it belongs to - a remaining open problem in genomics research.

Conclusion

Chapter 7 concludes this work with a comprehensive discussion of the presented research and an outlook on future perspectives.

Part I

Algorithms for Mass Spectrometry Imaging Data Analysis

Chapter 2

Non-negative matrix factorization and related methods for MSI data analysis

This chapter has been previously published as:

Nijs, M., Smets, T., Waelkens, E., and 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 35*, 21 (Nov. 2021), e9181. Publisher: John Wiley & Sons, Ltd

Abstract

Rationale Non-negative matrix factorization (NMF) has been used extensively for the analysis of mass spectrometry imaging (MSI) data, visualizing simultaneously the spatial and spectral distributions present in a slice of tissue. The statistical framework offers two related NMF methods: probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA), which is a generative model. This work offers a mathematical comparison between NMF, PLSA, and LDA, and includes a detailed evaluation of Kullback-Leibler NMF (KL-NMF) for MSI for the first time. We will inspect their results for MSI data analysis as these different mathematical approaches impose different characteristics on the data and the resulting decomposition.

Methods The four methods (NMF, KL-NMF, PLSA, and LDA) are compared on seven different samples: three originated from mice pancreas and four from human-lymph-node tissues, all obtained using MALDI-TOF MS.

Results Where matrix factorization methods are often used for the analysis of MSI data, we find that each method has different implications on the exactness and interpretability of the results. We have discovered promising results using KL-NMF, which has only rarely been used for MSI so far, improving both NMF and PLSA, and have shown that the hitherto stated equivalent KL-NMF and PLSA algorithms do differ in case of MSI data analysis. LDA, assumed to be the better method in the field of text-mining, is shown to be outperformed by PLSA in the setting of MALDI-MSI. Additionally, the molecular results of the human-lymph-node data have been thoroughly analyzed for better assessment of the methods under investigation.

Conclusions This paper offers an in-depth comparison of multiple NMF related factorization methods for MSI. We aim to provide fellow researchers in the field of MSI a clear understanding of the mathematical implications using each of these analysis techniques, which might affect the exactness and interpretation of the results.

Keywords: Non-negative matrix factorization, Mass Spectrometry Imaging, Kullback-Leibler, probabilistic latent semantic analysis, latent Dirichlet allocation

2.1 Introduction

Mass Spectrometry Imaging (MSI) is an exploratory research technique aiming to visualize the spatial distributions of biomolecules present in a slice of tissue [36]. Well-known MSI analysis techniques include clustering and dimensionality reduction methods, which entail amongst others factorization methods and non-linear dimensionality reduction (NLDR) techniques. Clustering, such as k-means[125], identifies similar pixels based on the distance between their spectra. NLDR methods, such as t-distributed stochastic neighbour embedding (t-SNE)[198] and uniform manifold approximation (UMAP)[135], embed the data set on a high-dimensional non-linear manifold. In contrast to these techniques, factorization methods reduce the traditionally very large MSI matrices to a product of two low-rank matrices, containing the spectral and spatial distributions of the data. Each column-row combination of this product can be seen as an individual component describing an important section of the data. Because of the simultaneous extraction of the spectral and spatial elements of MSI data and their straight-forward interpretability, factorization methods have been extensively proven useful over the last decade [172].

A well-known example of matrix factorization is principal component analysis (PCA) [131]. PCA projects the data orthogonally and linearly onto a number K of orthogonal principal components, i.e. the eigenvectors corresponding to the K largest eigenvalues of the matrix. The data is factored in a score matrix and a loadings matrix. The score matrix, which maximizes variation along each principal component, tries to capture as much of the data as possible on the reduced rank-K subspace. The loadings matrix describes the weights of the original data elements in the new space. PCA is an unconstrained factorization method and does not take into account the non-negative nature of MSI data. This can be resolved using non-negative matrix factorization (NMF) instead.

NMF, as shown in Figure 2.1, factorizes the matrix into a spatial and spectral component by minimization of the Frobenius norm whilst constraining all values to be greater than or equal to zero. NMF has been used extensively for the analysis of MSI data and extensions and variations of this algorithm is still actively being researched [214, 118].

Besides non-negativity as an important characteristic of MSI data, MSI data is known to follow a Poisson distribution because of the count nature of the mass spectrometer. With MALDI-MSI, the tissue under investigation is ionized by means of a laser beam, after which the ions are accelerated through the mass spectrometer until they reach a sensor that measures the number of hits as well as the time between ionization and reaching the sensor. Due to the heterogeneous distribution of molecules (proteins) within the tissue, the counts

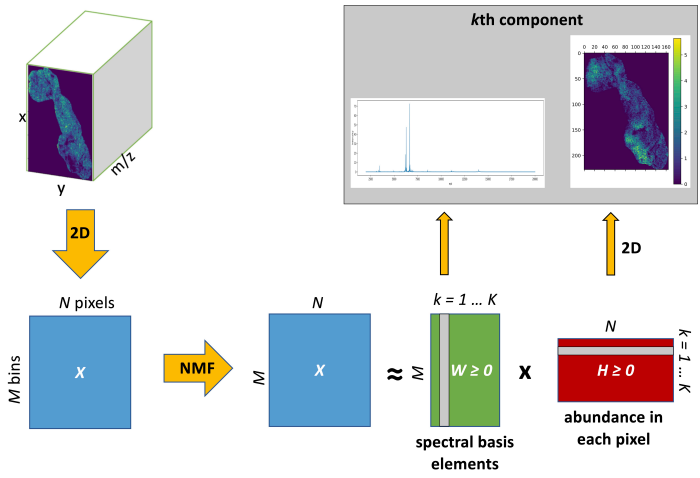


Figure 2.1: Graphical representation of non-negative matrix factorization.

(or intensity) measured for each m/z ratio can be seen as a random number of counts, allowing the assumption of MSI data being Poisson distributed, as previously claimed by Harn et al. and Piehowski et al. [90, 105, 45, 153]. Standard NMF minimizes the squared error assuming a Gaussian distribution of the noise, but Lee and Seung's 1999 paper [116] on NMF already proposed the use of Kullback-Leibler divergence (KL-NMF) instead of least squares as error measure to incorporate Poisson noise for the factorization of face images .

Another approach to include the Poisson prior to NMF can be found in the statistical framework: Probabilistic latent semantic analysis (PLSA) or indexing (PLSI) is a topic modelling technique often used for text mining. Here, a word count matrix of documents, a co-occurrence matrix, will be factorized in rank-K topics, each with a document-to-topic distribution and a word-to-topic distribution. It has been shown that maximizing the PLSA likelihood function yields the same objective as minimizing the KL divergence error. Under some conditions the KL-NMF and PLSA methods are equivalent, but they generally converge to different optima [219, 55]. In text mining, it is shown that latent Dirichlet allocation (LDA) outperforms PLSA in terms of disambiguation and precision due to the sparse Dirichlet prior [21]. LDA is generally not categorized as a factorization method but as a generative statistical model closely related to PLSA and thus indirectly connected to NMF [69].

Besides being non-negative and Poisson distributed, we also know that MSI data can be modelled as non-linear. Methods such as t-SNE and UMAP

perform a non-linear mapping of the data onto a lower dimensional structure and provide very intuitive visualizations because they take the entire feature space into account. One notable draw-back of these methods is the inability to simultaneously provide a spatial and spectral distribution of the data, making them harder to interpret than linear factorization methods.

In this paper, we will compare the results of the following methods: NMF, KL-NMF, PLSA, and LDA for the application on MSI data analysis. Although there are many more variants of NMF, we limit ourselves to these four because there is a clear resemblance in what they try to obtain, but their mathematics imply fairly different results. All but KL-NMF have already been investigated for the analysis of MSI data [123, 88, 197]. KL-NMF has been used in context of MSI in previous work, but has not yet been examined in detail or evaluated against standard NMF [25]. This is likely due to the theoretical resemblance with PLSA, which has also been stated specifically for MSI [201]. However, it has already been proven that both methods differ practically, and the theoretical resemblance only holds under certain circumstances [55].

In general, the selection of a certain algorithm for analysis is rarely substantiated, nor are the implications of a certain framework mentioned. Because MSI data is an exploratory technique, it is important to keep in mind the deviations resulting from the mathematical framework of the chosen algorithm. By comparing the results of different approaches, the correctness of results can be assessed with more certainty. Some frequently used numerical performance measures are the reconstruction error, computation time, and memory usage. Comparison of the non-linear mapping with the spatial distributions obtained with factorization methods could serve as an additional performance measure. As UMAP was found to be excellent for MSI data analysis [174, 10], we will examine whether the factorization method is able to capture the same regions as UMAP, for example. Furthermore, the mathematical complexity, interpretability, and scalability will also be used as evaluation criteria.

The next sections provide a mathematical description of the different algorithms, an experimental comparison and a discussion on the different methods. With this paper we hope to offer a comprehensive overview of the NMF-related methods mentioned and evaluate their performance on MSI data analysis by means of multiple measures, including the so far not yet compared KL-NMF method. We aim to provide a clear comparison and understanding of each method, listing their advantages and disadvantages. We will show that KL-NMF offers the best analysis for our data, when trying to match the obtained distributions of UMAP.

2.2 Methods

Define $\mathbf{X} \in \mathbb{R}_+^{M \times N}$, an MSI data set consisting of M rows, describing the measured m/z values, and N columns, describing all the (vectorized) pixels in the sample. We want to extract K non-negative components that represent the data as good as possible, according to a certain optimization function. The value K is called the rank of the decomposition. One straightforward way to determine the rank-K of a matrix is to look at the singular values. By taking K equal to the number of singular values of $\mathbf{X}$ that are distinctively higher than the other singular values, we get a good estimate for the rank of the non-noise part of the data. Alternatively, there exist several other, more complex methods for the rank estimation of a matrix as the determination of the rank is a research discipline on its own [193].

Extracting the non-negative components can be done using a variety of algorithms. Here, we will restrict ourselves to the examination of non-negative matrix factorization (NMF), Kullback-Leibler NMF (KL-NMF), probabilistic latent semantic analysis (PLSA), and latent Dirichlet allocation (LDA).

2.2.1 Non-negative Matrix Factorization (NMF)

The standard non-negative matrix factorization decomposes the matrix $\mathbf{X} \in \mathbb{R}_+^{M \times N}$ into two non-negative factor matrices: $\mathbf{W} \in \mathbb{R}_+^{M \times K}$, of which the columns represent the rank-K spectral basis elements, and $\mathbf{H} \in \mathbb{R}_+^{K \times N}$, describing the contribution of each basis element to the spectrum at each pixel p_i. This can be written as:

$$\mathbf{X} \approx \mathbf{WH}, \tag{2.1}$$

and yields the following optimization problem:

$$\min_{W,H \geq 0} ||\mathbf{X} - \mathbf{WH}||_F^2, \tag{2.2}$$

where the standard Frobenius norm is used, denoted as $|| \cdot ||_F^2$. The factor matrices are *component-wise* non-negative, which we denote as $W, H \geq 0$.

The main advantage of NMF is the easy interpretability of the factor matrices. Using the additive linear combination of the spectral basis elements in $\mathbf{W} \in \mathbb{R}_+^{M \times K}$, weighed by the contribution matrix $\mathbf{H} \in \mathbb{R}_+^{K \times N}$, the spectral signature of each pixel can be approximated. This way, NMF provides meaningful

information, both spectral and spatial, making it a well-suited method for MSI data analysis.

Unfortunately, the NMF problem is NP-hard (non-deterministic polynomial-time hard) and not unique. In most cases, the objective cannot be solved exactly and the optimum needs to be approximated numerically instead. This typically results in only a local optimum. For the numerical approximation, many algorithms exist such as multiplicative update (MU) rules and alternating least-squares.

NMF decomposition, from the scikit-learn Python package[152], was configured with a coordinate descent (CD) solver. This is the framework used most for NMF; it optimizes in alternation one of the factor matrices W or H, while keeping the other one fixed. This will result in a nonnegative least squares subproblem with only one unknown, which is convex and thus easier to solve. Because this optimization problem is symmetric in W and H, the same update can be used for both factor matrices. These updates guarantee a decrease of the objective function. The computational complexity of NMF on most real-world problems equals $\mathcal{O}(mnk)$ operations per iteration [80].

2.2.2 Kullback-Leibler Non-negative Matrix Factorization (KL-NMF)

As mentioned before, there exists a variation on traditional NMF that minimizes the KL divergence [70]. In this case, Poisson distribution is included when factorization is performed, instead of the Gaussian noise implied by standard NMF. For each m/z value i, and for each pixel j, the observed intensity x_{ij} follows a Poisson distribution $P(x_{ij})$. The matrix $\mathbf{X} \in \mathbb{R}_+^{M \times N}$ is factored in similar matrices $\mathbf{W} \in \mathbb{R}_+^{M \times K}$and $\mathbf{H} \in \mathbb{R}_+^{K \times N}$, but the optimization problem now includes the following maximum likelihood function:

$$L(\mathbf{X}|\mathbf{W},\mathbf{H}) = \prod_{ij} P(x_{ij}) = \prod_{ij} \exp(-\lambda_{ij}) \frac{\lambda_{ij}^{x_{ij}}}{x_{ij}!}, \tag{2.3}$$

with rate parameter $\lambda_{ij} = (\mathbf{WH})_{ij}$. Or, similarly, it minimizes the negative log-likelihood:

$$\min_{\mathbf{W},\mathbf{H}\geq 0} -\ln L(\mathbf{X}|\mathbf{W},\mathbf{H}) = \min_{\mathbf{W},\mathbf{H}\geq 0} \sum_{i=1}^{N}\sum_{j=1}^{M} \lambda_{ij} - x_{ij}\ln(\lambda_{ij}) + \ln(x_{ij}!) \tag{2.4}$$

$$= \min_{\mathbf{W},\mathbf{H}\geq 0} \sum_{i=1}^{N}\sum_{j=1}^{M} (\mathbf{WH})_{ij} - x_{ij}\ln(\mathbf{WH})_{ij} + c, \tag{2.5}$$

with c a constant depending only on $\mathbf{X}$, not on $\mathbf{W}$ and $\mathbf{H}$.

To solve KL-NMF, multiplicative update rules (MU rules) are most commonly used [92]. However, similar to standard NMF, the problem is NP-hard and not unique. The interpretation of the factor matrices is similar but using the KL-NMF now implies Poisson distribution on the result. This corresponds, better than standard NMF, to the prior information known about MSI data, but has, as known so far, not yet been examined and only rarely been used for MSI data analysis. In practice, many NMF estimators exist for different stochastic distributions of data.

The scikit-learn Kullback-Leibler extension of NMF uses multiplicative update rules (MU) with β-divergence for $\beta = 1$ as factorization algorithm, as this is equivalent to the KL divergence [152]. The MU are the most widely used approach for KL-NMF where matrices$\mathbf{W}$ and $\mathbf{H}$ are updated with component-wise multiplications, defined as follows:

$$\mathbf{H} \leftarrow \mathbf{H} \circ \frac{\mathbf{W}^T\mathbf{X}}{\mathbf{W}^T\mathbf{WH}}, \tag{2.6}$$

$$\mathbf{W} \leftarrow \mathbf{W} \circ \frac{\mathbf{XH}^T}{\mathbf{WHH}^T}. \tag{2.7}$$

Here, $\circ$ (resp. $\frac{\cdot}{\cdot}$) denotes the *component-wise* product (resp. division) of matrices. These rules are easy to implement and scale well. This approach also has a computational complexity of $\mathcal{O}(mnk)$.

2.2.3 Probabilistic Latent Semantic Analysis (PLSA)

The probabilistic equivalent of KL-NMF is PLSA. In essence, both methods have the same theoretical objective, but they use different approaches of the MU rules to achieve the optimum. Also, PLSA, as a statistical approach, will result in probability distributions, while KL-NMF will consist of values relative to the given data set. This makes them inherently dissimilar in practice and causes them to converge to different local optima.

PLSA originates from text mining, as it is an extension on latent semantic analysis (LSA). In text mining, PLSA observes the correspondence between words (w) in multiple documents and collects the word counts per document in a co-occurrence matrix [95]. Additionally, in the setting of PLSA, we define as hidden variables the topics (c) across the data set to which certain words and documents are related.

$$P(w_i, d_i) = \sum_{c=1}^{K} P(c)P(d_j|c)P(w_i|c). \tag{2.8}$$

Translating this to MSI, as has been done in previous work, we can see this as the correspondence between the m/z bins (words) and the spectra (documents), located at each pixel entry [88]. The hidden topics, in case of MSI, can be defined as the different cell types within the tissue. When the number of m/z bins ℓ in the data set is sufficiently large, PLSA results in the following factorization model:

$$\mathbf{X} \approx \ell\hat{\mathbf{A}}\text{diag}(\hat{\mathbf{s}})\hat{\mathbf{B}} = \mathbf{WH}. \tag{2.9}$$

Here, $\hat{\mathbf{s}} \in \mathbb{R}_+^K$, corresponding to the factor $P(c)$ of the previous equation, is a vector describing the probability of a random m/z value being related to the kth topic, matrix $\hat{\mathbf{A}} \in \mathbb{R}_+^{M\times K}$, or $P(w_i|c)$, describes the probability of a certain m/z value contributing to a certain spectrum, and matrix $\hat{\mathbf{B}} \in \mathbb{R}_+^{K\times N}$, or $P(d_j|c)$, the probability of each spectrum belonging to a certain topic or tissue type. As they entail probabilities, the vector and matrices are exposed to sum-to-one constraints with respect to the spectra. If we define $\mathbf{W} = \hat{\mathbf{A}}$ and $\mathbf{H} = \ell\text{diag}(\hat{\mathbf{s}})\hat{\mathbf{B}}$, the PLSA model can be seen as an NMF. More specifically, PLSA assumes $\mathbf{X}$ to be Poisson distributed in terms of $\hat{\mathbf{A}}\text{diag}(\hat{\mathbf{s}})\hat{\mathbf{B}}$. As the stationary points of KL-NMF preserve the row and column sum of $\mathbf{X}$, it can be proved that they correspond to a solution of PLSA [55]. This allows the correspondence of PLSA to KL-NMF in the local optima, despite the normalization difference between two formulations. Nevertheless, in practice both methods do differ. The MU rules of KL-NMF do not take the normalization of PLSA into account explicitly. Moreover, PLSA uses the Expectation-Maximization (EM) algorithm, where it applies a normalization in each step. For a more elaborate mathematical derivation on the correspondences and differences of these two methods, we refer the reader to a more elaborate work [81].

Using the EnsTop [1] Python package, the expectation step calculates the probability of each topic given a word-document pair:

$$P(c|w,d) = \frac{P(c|w)P(d|c)}{\sum_{c=1}^{k} P(c|w)P(d|c)}. \tag{2.10}$$

This is followed by the maximization step that uses the result of the expectation step to calculate the estimates $P(w|c)$ and $P(c|d)$:

$$P(w|c) = \frac{\sum_d \mathbf{X}_{w,d} P(c|w,d)}{\sum_{d,c} \mathbf{X}_{w,d} P(c|w,d)}, \tag{2.11}$$

$$P(c|d) = \frac{\sum_w \mathbf{X}_{w,d} P(c|w,d)}{\sum_{w,d} \mathbf{X}_{w,d} P(c|w,d)}. \tag{2.12}$$

The computational complexity of PLSA equals $\mathcal{O}(mnk)$ [95].

2.2.4 Latent Dirichlet Allocation (LDA)

Although LDA is not generally thought of as a factorization method, it is a generalization of PLSA and is often considered as the better of the two in the field of text mining, where they both originate from. The main difference between both methods is that PLSA aims to perform an exact modelling of the data set, while LDA tries to find unknown groupings within the data. LDA is a generative statistical model that uses a Dirichlet prior to impose sparsity on the solution, making it more robust to modelling the noise in the data set. For MALDI-TOF MSI, it has been used as a soft clustering technique and with tandem MS, one can use LDA to cluster the fragmentation of the data in substructures, as has been done in previous research [40, 197, 196, 160].

Let us define similarly as for PLSA: N as the number of m/z bins (or words) in the data, K as the number of components (or topics), which can be seen as the different cell-types within the tissue, and D as the number of spectra or pixels in the MSI data set (which corresponds to the number of documents in the setting of text mining). Also, let θ_d be the probability of a spectrum (d) belonging to a cell type (c) according to a Dirichlet prior α, and β_c be the probability of a certain m/z bin (w) belonging to a cell type (c) according to another Dirichlet prior η.

A pseudo algorithm [21] of the generative LDA method can now be defined as:

For each component or cell type $c \in K$, draw $\beta_c \sim \text{Dirichlet}(\eta)$.

For each spectrum d $\in D$, draw $\theta_d \sim \text{Dirichlet}(\alpha)$.

For each m/z bin i in document d : (2.13)

Draw the cell type assignment $z_{di} \sim \text{Multinomial}(\theta_d)$,

Draw the observed m/z bin $w_{di} \sim \text{Multinomial}(\beta_{z_{di}})$.

This process generates a latent structure that can be used for either prediction of new entries or for data exploration. Using the scikit-learn package for LDA, one can use an efficient batch variational Bayes algorithm to estimate the posterior distribution $p(z, \theta, \beta|w, \alpha, \eta)$ [152]. Using this algorithm, LDA has a computational complexity of $\mathcal{O}(mnk)$. For a more detailed description on this algorithm, we refer the reader to the work of Hoffman and Blei [94].

Note that if we sum over the different topics, LDA can be thought of as a "multinomial PCA" interpretation, i.e., a probabilistic factorization optimizing the squared error.

2.2.5 Evaluation criteria

To compare the methods described above, multiple evaluation criteria are investigated. First, we compare all methods numerically: we will calculate the reconstruction error, the computation time, and the memory usage for each experiment.

The reconstruction error will be investigated by means of three different measures: the relative $\boldsymbol{\ell}_1$-norm, the relative $\boldsymbol{\ell}_2$-norm and the KL divergence. We define the relative errors as the norm of the difference between the original and reconstructed data divided by the norm of the original data:

$$\frac{||\mathbf{X} - \mathbf{WH}||_p}{||\mathbf{X}||_p}, \tag{2.14}$$

with the $\boldsymbol{\ell}_p$-norm for a vector x and a real value $p \geq 1$ formulated as:

$$||x||_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{\frac{1}{p}}. \tag{2.15}$$

The relative KL divergence between $\mathbf{X}$ and $\mathbf{WH}$ is calculated as:

$$\sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right). \tag{2.16}$$

Here, P and Q correspond respectively to the vectorized original data $\mathbf{X}$ and the vectorized approximation $\mathbf{WH}$.

As has been shown in the previous section, NMF and LDA impose a Gaussian distribution of the noise and KL-NMF and PLSA a Poisson distribution. Because of these statistical priors, we can assume NMF to perform better on the squared error (ℓ_2-norm, Frobenius norm) and KL-NMF and PLSA better on the KL divergence error. To this end, we included a third, unrelated error measure: the ℓ_1-norm. This norm calculates the error by means of the absolute value, rather than the squared values used by the ℓ_2-norm, this makes the ℓ_1 reconstruction error more robust towards outliers. Note that for LDA, the comparison using reconstruction errors makes less sense as the model is generative and not aiming to reconstruct the data. We will thus only calculate the KL divergence in the case of LDA, to estimate to what extent the model follows a similar distribution as the given data set.

Additionally, we also measure the average computation time and memory usage for each experiment. For each method and for each data set, we perform the experiment ten times until convergence, or the maximum number of iterations is reached. Because of the size of real MSI data sets, which is expected to still increase as the technology advances, we monitor time and memory usage as they might reveal practical limitations of the algorithms.

Besides numerical performance measures, there are a number of additional criteria that should be investigated as well. First, we will check how well the spatial distributions resemble the solutions obtained by the non-linear mapping obtained with UMAP. We know these methods are better to capture the non-linearity of MSI data, and thus could serve as an additional quality check. Secondly, since MSI data is ever increasing in size, sparse representations could offer a solution in the future. This way, the large MSI data sets can be represented more efficiently, by using less memory and less computational power. We want to know how well the examined methods are able to perform with sparse representation, and thus whether they are able to scale up together with the data size. Thirdly, as mentioned before, we want to understand what the meaning is of the results. Depending on the mathematical framework and the used model, interpretation of the components will differ. We will compare each method in terms of their implications on the interpretation. Additionally, the computational complexity and mathematical simplicity of each algorithm will be reviewed. Finally, aside from the mathematical aspects, we will take a closer look at the biological implications for a human-lymph-node sample. Using all these numerical and non-numerical evaluation criteria, we aim to offer a complete comparison of NMF, KL-NMF, LDA, and PLSA for the analysis of MSI data.

2.3 Experimental setup

2.3.1 Data

In order to make proper conclusions on the performance of each method, we executed the experiments on multiple real MSI data sets: mouse pancreas data (M1, M2, and M3) and human-lymph-node data (P1, P2, P3, and P4). The data have been analyzed using a Bruker rapifleX MALDI-TOF mass spectrometer, with a mass resolution up to 50.000. For the mouse pancreatic tissue, 7 μm thick cryosections were prepared and mounted on ITO glass slides, with sinapinic acid (SA) as the matrix applied via Bruker ImagePrep. The pixel size is 50 μm, the recorded m/z range is 2-20 kDa in positive linear mode, and the acquisition speed is 9 pixels/sec with 1000 laser shots/pixel and a laser repetition rate of 10 kHz. For the human lymph node samples, 5 μm thick cryosections were mounted on ITO glass slides, with 2,5-Dihydroxybenzoic acid (2,5-DHB) as the matrix, applied via sublimation. The pixel size is 10 μm, the recorded m/z range is 620-1200 Da in positive reflector mode, and the acquisition is performed with 200 laser shots/pixel and a laser repetition rate of 10 kHz, resulting in an acquisition speed of 32 pixels/sec. Total-ion count (TIC) normalization was used for each data set. Table 1 gives an overview of the specifications of each data set.

2.3.2 Implementation

We examined the four methods (NMF, KL-NMF, PLSA, and LDA) on the seven tissue samples described above. The experiments were done in Python using the scikit-learn package [152] for NMF, KL-NMF, and LDA, and the EnsTop [1] package for PLSA.

For fair comparison, all experiments were executed using HPC provided by the Flemish Supercomputer Centre (VSC). Each setup used 36 computing nodes on a 2 Xeon Gold 6140 CPUs@2.3 GHz.

All four algorithms require a given number of components for the decomposition, or rank. The meaning of this rank is the same for each of the methods: it is the number of components or cell types we expect to contribute to the data. We determined the rank of the decomposition by means of the singular value decomposition, as explained in section 2.4.1.

Each experiment was run until the update between two iterations reached a threshold below 1e-3, or until a maximum number of iterations is reached, whichever comes first. We set the maximum number of iterations to 400 for

NMF, KL-NMF and PLSA, and to 20 for LDA. The number of iterations for LDA is much lower because LDA has a much longer computation time per iteration.

Finally, we tried to improve the optimization by using a sparsity favouring initialization by using Non-negative double singular value decomposition (NNDSVD) instead of random initialization [26]. This was done for NMF and PLSA. For KL-NMF, solved with multiplicative update rules, we used the NNDSVD variant that filles the zeros with small random values, such that they can still be updated during the iterations. LDA, being a statistical generative model, uses a prior for the topic word distribution (β) and a prior for the document word distribution (θ). Both were initialized at the default value $1/(n_{comp})$.

2.4 Results

Each algorithm ran multiple times on each of these samples, while measuring the numerical evaluation criteria. The additional non-numerical criteria are investigated separately.

2.4.1 Rank selection

MSI data contains a significant amount of noise, it is obvious we do not want to model all this noise exactly. We want to select a number of components of the factorization methods equal to the rank of the non-noise part of the data. Using a singular value decomposition, one can plot the singular values in a semi-log manner and select a rank equal to the knee in this graph [216, 208, 78]. This way, we estimate the number of components based on the variance explained by the singular values. As soon as the magnitude of the singular values stops decreasing significantly, one can expect the rest of the singular values are describing the noise present in the data. For a detailed investigation of a MSI data set, one can use the SVD as a first guess, but it is advised to inspect neighboring values as well or use an additional method such as the AIC criterion for rank selection. We refer to Ubaru and Saad for further information on rank estimation [193].

Figure 2.2 and Figure 2.3 display respectively for the M2 pancreas data and the P1 human-lymph-node data, the formation of the knee around the fifth largest singular value. For this, we will decompose the data in five components for both the pancreatic and lymph-node samples. This choice of rank, $k = 5$,

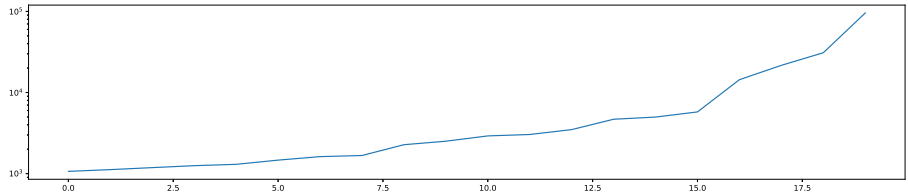


Figure 2.2: Rank selection of the pancreas sample M2 - 20 largest (right-to-left) singular values are plotted on a semilog scale. There is a kink in the curve around the 5th highest singular value. A good estimate for the rank of this matrix will thus be $K = 5$.

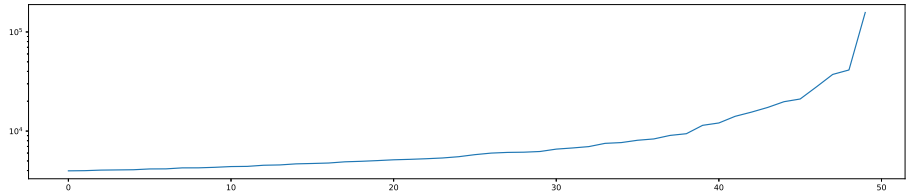


Figure 2.3: Rank selection of the lymph-node sample P1 - 50 largest (right-to-left) singular values are plotted on a semilog scale. There is a kink in the curve around the 5th highest singular value. A good estimate for the rank of the matrix will thus be $K = 5$.

can be justified by the visualization done with UMAP where one can find five distinct regions for M2 as for P1, see Figure 2.6.

2.4.2 Pancreas data

The first data set on which we will compare the factorization methods is the mouse pancreatic tissue. We have three samples: M1, M2, and M3 of relatively small size for MSI data, see Table 2.1. Each sample was decomposed in five components, using NMF, KL-NMF, PLSA, and LDA. Table 2.2 shows the average relative $\boldsymbol{\ell}_1$-norm, relative $\boldsymbol{\ell}_2$-norm, KL divergence, computation time, and memory usage after 5 runs each. In case of the pancreatic samples, each method reaches convergence before the maximum number of iterations. Note that each method converges to exactly the same value for the multiple runs, making the standard deviation zero for the reconstruction errors. Only the computation time and memory usages fluctuate over multiple runs. We see, as expected, NMF minimizing the relative $\boldsymbol{\ell}_2$-norm and KL-NMF the KL divergence

Table 2.1: Specifications of the mouse pancreas data (M1, M2, and M3) and human-lymph-node data (P1, P2, P3, and P4). We define the sparsity of the data sets as $1 - \frac{\#_{nonzero}(A)}{\#_{total}(A)}$, which shows the percentage of zero entries in the matrices.

sample	**pixels**	**m/z bins**	**sparsity**	**size (GB)**
M1	10 606	14 000	0.030	1.19
M2	14 791	14 000	0.090	1.66
M3	6 937	14 000	0.095	0.78
P1	572 173	8 000	0.005	36.62
P2	552 701	8 000	0.015	33.33
P3	354 444	12 000	0.041	34.03
P4	275 133	12 000	0.020	26.41

error. Additionally, KL-NMF scores also best on the relative ℓ_1-norm. Here, it is interesting to see that PLSA scores second best on the KL divergence error as we would expect. LDA only scores reasonably well on the KL divergence for the M3 sample.

The computation times differ significantly between the investigated methods. NMF clearly executes the fasted, and LDA is by far the slowest, taking into account we limited the number of iterations to twenty. PLSA is a bit faster than KL-NMF, but of the same magnitude. For the memory usage, NMF is clearly the most memory efficient, this can be explained as NMF reaches the update threshold within only one iteration for all samples. LDA, KL-NMF, and PLSA are again of the same magnitude, with LDA the most memory efficient of the three.

The resulting decompositions can be visualized as a spectral and spatial distribution for each component. Figure 2.4 shows the obtained results for sample M2. A first glance at these results shows that the PLSA decomposition misses a lot of information in its spatial distributions and that the LDA results show poor spectral distributions.

Remark that the spectra have different scales for the intensities across all methods and components. With the pancreas data, most differences between the decompositions are noticeable in the smaller peaks.

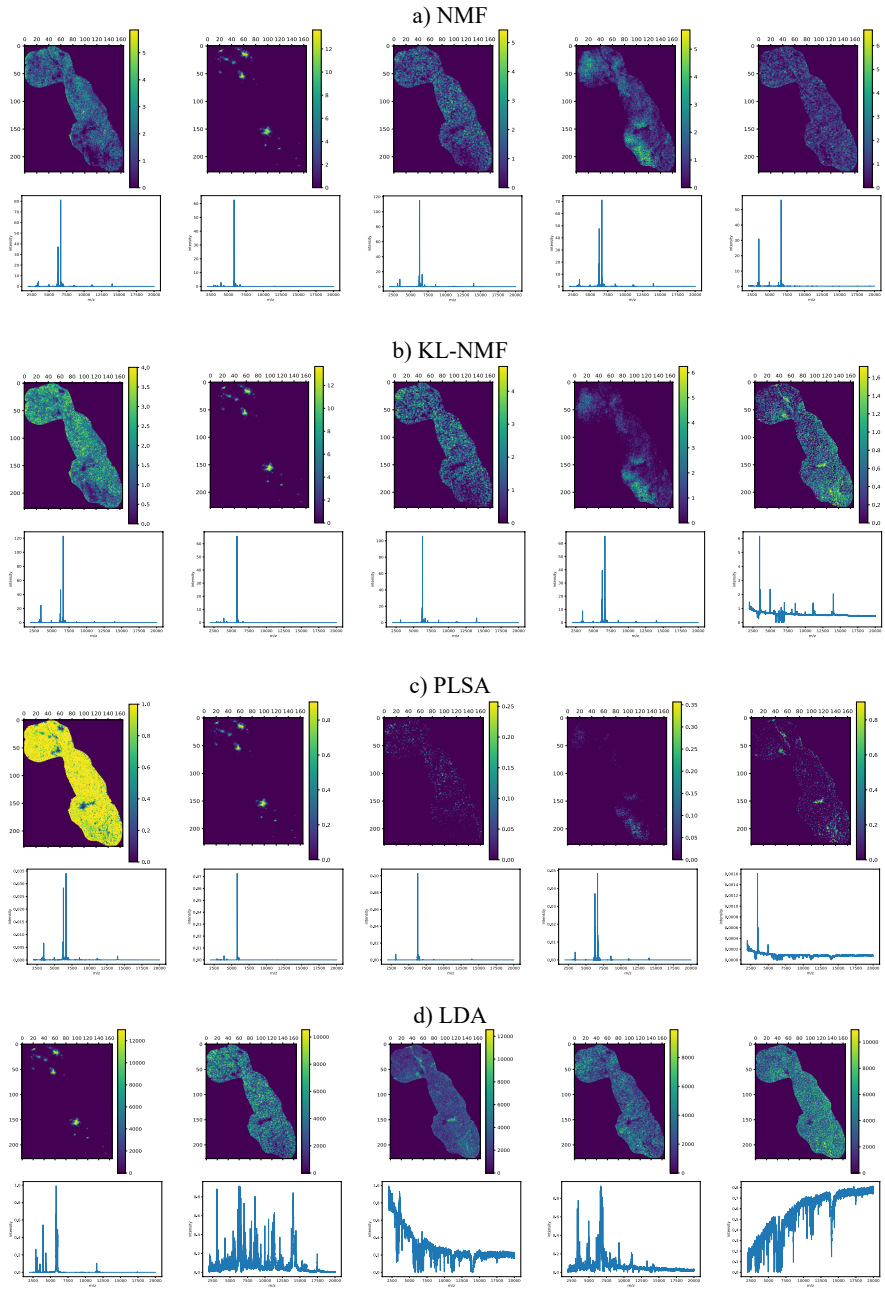


Figure 2.4: Visual results of the mouse pancreas sample M2. For each method: a) NMF, b) KL-NMF, c) PLSA, and d) LDA we have executed a decomposition in 5 components. The spatial distributions are given in the top rows and the spectral distributions in the bottom rows.

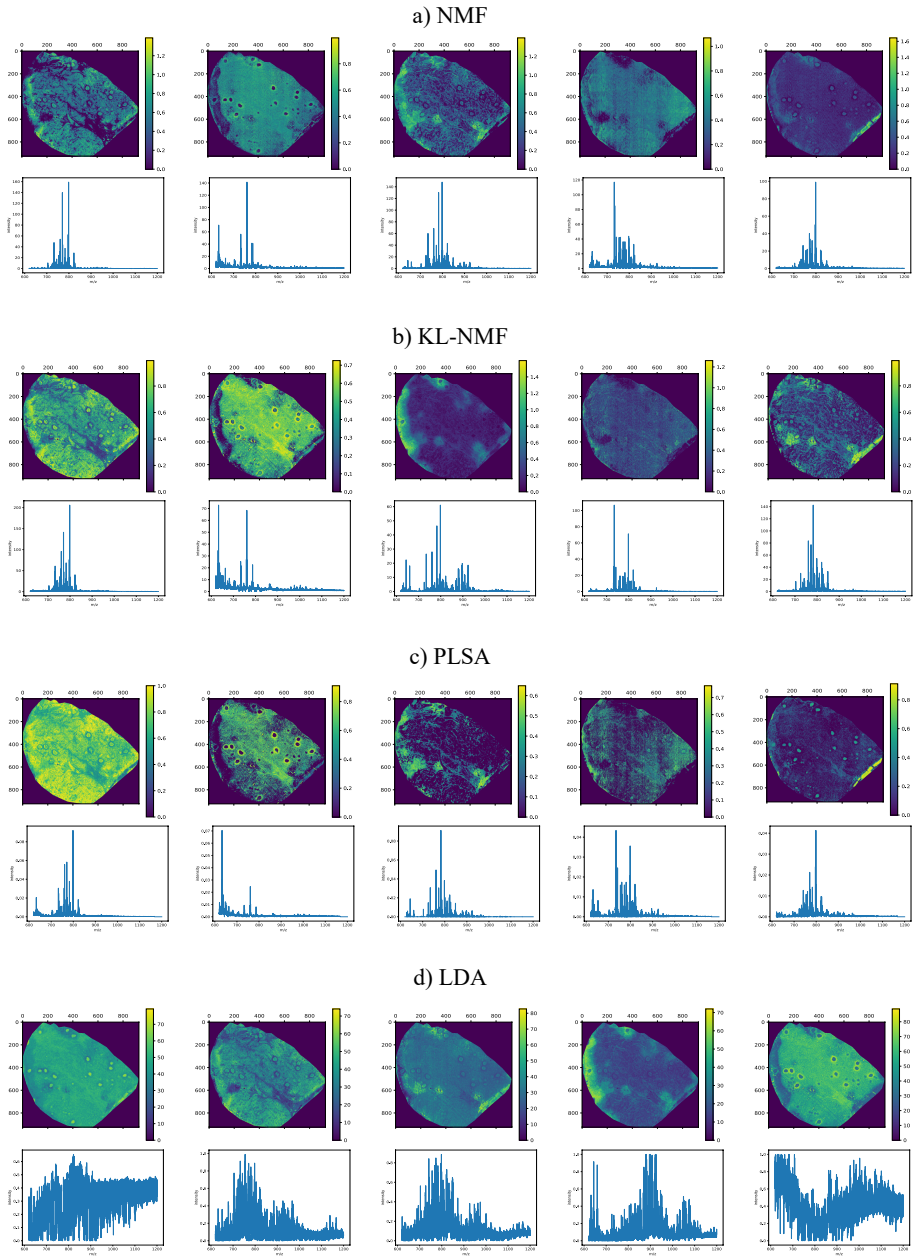


Figure 2.5: Visual results of the human-lymph-node sample P1. For each method: a) NMF, b) KL-NMF, c) PLSA, and d) LDA we have executed a decomposition in 5 components. The spatial distributions are given in the top rows and the spectral distributions in the bottom rows.

Table 2.2: Numerical analysis of NMF, KL-NMF, PLSA, and LDA on three mouse pancreas samples: M1, M2, and M3. For the reconstruction error, we use three measures: the relative ℓ_1 and ℓ_2-norm, and the Kullback-Leibler divergence. Each experiment ran until convergence in the update was reached, with a threshold of 1e-3.

	method	ℓ_1	ℓ_2	**KL**	**time (s)**	**mem. (GB)**
M1	NMF	0.4516	**2.486e-3**	0.2108	**2.6 ± 0.0**	**1.1 ± 1.0**
	KL-NMF	**0.3955**	3.233e-3	**0.1480**	60.9 ± 3.1	6.3 ± 1.4
	PLSA	0.8700	0.1350	0.1658	44.4 ± 0.9	6.6 ± 2.3
	LDA	-	-	11770	3970.4 ± 27.7	1.2 ± 0.0
M2	NMF	0.4770	**1.838e-3**	0.2578	**3.7 ± 0.1**	**0.8 ± 1.6**
	KL-NMF	**0.4366**	2.290e-3	**0.1922**	88.2 ± 7.2	7.4 ± 1.7
	PLSA	0.8283	0.0909	0.2096	63.9 ± 5.1	6.5 ± 2.4
	LDA	-	-	15376	4221.0 ± 29.1	1.7 ± 0.002
M3	NMF	0.5719	**1.266e-2**	0.2872	**1.6 ± 0.0**	**0.2 ± 0.4**
	KL-NMF	**0.5585**	1.442e-2	**0.2782**	35.8 ± 4.7	3.6 ± 1.0
	PLSA	0.7465	0.3067	0.2919	27.1 ± 0.3	2.1 ± 1.6
	LDA	-	-	0.5658	3761.2 ± 51.6	0.8 ± 0.002

2.4.3 Lymph-node data

The second data set we will investigate are the human-lymph-node samples. We perform the same experiments on the four lymph-node samples: P1, P2, P3 and P4, with more or less comparable data size, see Table 1. Their sizes are however multitudes greater than the pancreatic samples, affecting as a consequence the peak memory usage significantly. In fact, the lymph-node data resembles much more the current standard of MSI data with higher spatial resolutions. Again, each sample was decomposed in five components, using NMF, KL-NMF, PLSA, and LDA. From Table 2.3 we can deduce similar trends as for the pancreatic sample. Here, as was the case for the pancreas samples, NMF outperforms the other in minimizing the relative ℓ_2-norm and the same is true with KL-NMF for the relative ℓ_1-norm and the KL divergence error. Looking at the standard deviation for each method, we note that they all keep converging to the same solution. Only for LDA convergence of the update between two iterations was not reached, LDA was stopped after its maximum number of iterations (12) as the amount of computation time, roughly twelve hours, was already many times higher those of the other methods.

Table 2.3: Numerical analysis of NMF, KL-NMF, PLSA, and LDA on four human lymph samples: P1, P2, P3 and P4. For the reconstruction error, we use three measures: the relative ℓ_1 and ℓ_2-norm, and the Kullback-Leibler divergence. Each experiment for NMF, KL-NMF, and PLSA ran until convergence in the update was reached, with a threshold of 1e-3. For LDA convergence was not reached within the maximum number of 12 iterations.

	method	ℓ_1	ℓ_2	**KL**	**time (s)**	**mem. (GB)**
P1	NMF	0.8042	**2.740e-3**	0.5559	**120 ± 3**	104 ± 2
	KL-NMF	**0.7865**	2.869e-3	**0.5365**	2919 ± 133	218 ± 1
	PLSA	49300	0.1475	11545	2159 ± 46	422 ± 8
	LDA	-	-	10710	46 297 ± 811	**37 ± 1**
P2	NMF	0.6213	**1.08e-4**	0.4185	**112 ± 8**	92 ± 3
	KL-NMF	**0.6065**	1.147e-4	**0.3308**	2 687 ± 105	196 ± 2
	PLSA	30266	0.0778	0.4077	1754 ± 69	307 ± 15
	LDA	-	-	12304	42 981 ± 2415	**33 ± 1**
P3	NMF	0.6210	**3.233e-4**	0.3342	**119 ± 12**	93 ± 3
	KL-NMF	**0.5958**	3.406e-4	**0.3133**	3 761 ± 52	197 ± 1
	PLSA	42952	0.1926	0.3418	1856 ± 151	303 ± 15
	LDA	-	-	0.9790	42 455 ± 264	**35 ± 2**
P4	NMF	0.6823	**2.557e-4**	0.3899	**102 ± 11**	73 ± 2
	KL-NMF	**0.6402**	2.706e-4	**0.3531**	2 096 ± 60	155 ± 1
	PLSA	34761	0.1882	0.3899	1379 ± 120	242 ± 14
	LDA	-	-	0.9769	33 897 ± 261	**28 ± 2**

For the computation times, again NMF is by far the fastest, PLSA and KL-NMF are again of the same order of magnitude and still reasonably fast, with KL-NMF being to slower of the two, but reaching better reconstruction errors. Memory usage is now the lowest for LDA.

Visually, Figure 2.5 shows the spectral and spatial distributions per method and per component obtained for P1. A gamma-correction of $\gamma = 0.5$ was applied to the 2D visualisations for higher contrast. NMF, KL-NMF and PLSA all show clear decompositions, but differ in their details. Each method extracts different contributions of the relevant peaks. Again, as for the pancreatic sample, LDA results in poor spectral representations.

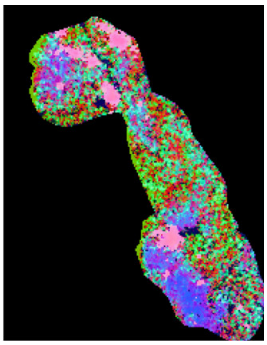
2.4.4 Additional evaluation criteria

Aside from the numerical experiments, measuring reconstruction errors, computation time and memory usage, we want to assess the performance of each method more generally as well. Since there no criterion is the most designated one for the evaluation of MSI data analysis, we will include several more evaluation criteria in order to obtain a complete review of every method.

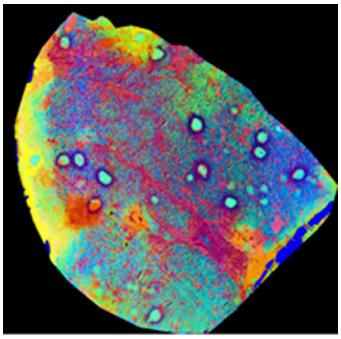
Visual inspection using UMAP

As already described in the works of Smets et al. [174] and some recent computational review papers [201, 10], UMAP-based data processing is very well suited for the analysis of MSI data sets. Figure 2.6a, obtained from Smets et al. [174], shows the visualization of a three-dimensional UMAP embedding with cosine distance of sample M2. From the UMAP embedding, we can distinguish five different regions based on the five distinct colours: pink, blue, black, red and green. Based on previous research [175], we know that the pink region corresponds to the Islets of Langerhans, the blue region is linked to exocrine tissue and the black region to blood vessels. The red and green region are not very clear to interpret and have not yet been related to an anatomical structure. To this end, we will perform a visual comparison based solely on the pink, blue and black regions of the UMAP embedding.

For all methods, the two most first components give a partition between the islets and the rest of the tissue. This can be explained by the high intensity of insulin (m/z = 5,800) in the data set. For the exocrine tissue, we see a similar component in NMF and KL-NMF, which is not detected by LDA. PLSA also uncovers the exocrine tissue, but as a less clear image. And lastly the blood vessels are not found in NMF, but are in KL-NMF, PLSA, and LDA. Note that for this anatomical component, we see more noise in the spectra than for the others. LDA shows visually rather good distributions but lacks precision in the spectra. For our MALDI data set, LDA is too general in the decomposition of the spectra. Similarly, we take a look at the correspondence of the decompositions to the UMAP embedding of P1, see Figure 2.6b, also obtained from Smets et al. [174]. The five coloured regions we can distinct here are: yellow, red, blue, green and orange. The spatial distributions of NMF for P1 shows green, blue/green, orange/yellow, and two unidentifiable regions compared to the UMAP decomposition. KL-NMF shows the green region in its first component, next blue/red, yellow, an unidentifiable region and lastly the orange part. For the P1 sample, we see that PLSA performs better than for the M2 sample, we now get a visualisation of similar good correspondence to UMAP as KL-NMF. Both methods however show clearly different decompositions of



a)



b)

Figure 2.6: UMAP visualization with cosine distance metric of samples M2 (a) and P1 (b). Obtained from the work of T. Smets et al. [174], see acknowledgements for permission granted.

the sample, but neither are able to extract the exact five regions visible in the UMAP representation. The spatial components of LDA show less detailed visualisations, they are of similar correspondence to UMAP as NMF.

Scalability

Anticipating the increase in resolution of future MSI data sets, we want to investigate how the factorization methods will be able to cope with even bigger data in the future. The use of sparse representations will be essential, but so far, not all algorithms are able to provide equally good results using sparse operations. With the algorithms we have used for the numerical experiments, we tested their sparse equivalents to gain some insight on their behaviour. NMF improves in speed with little loss in performance. Sparse KL-NMF also improves computation time, but there is a significant loss in accuracy when doing so. The sparse equivalent of MU is not yet satisfying for MSI data analysis. The EnsTop PLSA algorithm still shows some bugs when there is a sparse input. LDA, however, is faster and more accurate using a sparse input because of the sparse Dirichlet prior.

Note that research towards factorization methods is still ongoing and might lead to new extensions for big data analysis in the future.

Additionally, when looking at Table 2.1, the data sets do not exhibit large sparsity properties. Not many zeros are present, but there are many small values. By adding a cut-off, one might increase sparseness. How to establish a cut-off value and use sparse MSI data analysis, is still something to be explored in the future.

Interpretability

The interpretation of the decomposition depends on the method that was used. For the algebraic methods, the obtained factors can be seen as the spectral basis elements of the data matrix and a weight matrix describing their contribution to each pixel. The components retrieved using NMF and KL-NMF contain information, which is deduced directly from the data matrix.

For PLSA, we get similar components except that the weight matrix is now a probability matrix. The results show how likely a certain m/z value is present at a certain pixel. Although the spectra may look similar, they have to be approached differently regarding interpretation.

LDA adds a Dirichlet prior to PLSA, imposing more sparsity in the result and better properties for generation of new "documents". This generative property has little use in the field of MALDI-MSI and that is why the resulting spectral distributions are worse when decomposing with LDA; it generalizes what spectral information might be involved in that component. It could be more useful when an MSI data sets contains fragmentation effects we do not want to model, as is the case for tandem MS.

2.5 Discussion

For the selection of a MSI data analysis approach there will always be a trade-off; there is no holy grail (yet). This work focusses on NMF-related factorization methods as there are many it is not always clear how they are alike and how they differ. We examined NMF, KL-NMF, PLSA, and LDA both experimentally and theoretically. We can conclude that each method has its own strengths and depending on the goal of the analysis, one might out-perform the other. Interestingly, KL-NMF was not yet thoroughly evaluated in MSI context but proved to be an, at least, equally useful approach.

NMF is a well-known method for data analysis in many fields as well as for MSI. Its popularity comes from the well interpretable decomposition. From the numerical assessment, we find that NMF outperforms all other methods by far in terms of computation time and has often the smallest memory usage as well. Naturally, NMF also scores best on the relative ℓ_2 -norm, as its objective function is the minimization of the squared error. Additionally, the results are easy to interpret as basis spectral elements and their spatial weighing across the samples. For MSI data factorization, we advise to always start with NMF. Even towards future, possibly bigger MSI data sets, NMF is easy to extend to its sparse equivalent with only small loss in information.

A mathematically more suited extension of NMF for the factorization of MSI data is KL-NMF. By imposing Poisson distributed noise, rather than Gaussian, we see that the visual results are even better. Comparison with the UMAP embedding of Figure 2.6 shows that the KL-NMF components (Figure 2.4 and Figure 2.5) extract all regions visible for M2 and almost all regions for P1. This can be seen as a huge advantage, offering a spectral distribution for the spatial regions UMAP identifies. It is in fact the only method that consistently is able to capture the spatial distributions so well, making it even more surprising it has rarely been used for MSI so far. KL-NMF also was able to outperform the other methods in terms of both the relative ℓ_1 reconstruction error and the KL divergence error. Nevertheless, there are some disadvantages of KL-NMF. First, computationally it takes a bit longer than NMF, with a difference of one order of magnitude. Also, the sparse implementation of the algorithm, which could be essential in the future, performs much worse.

At the statistical end, we took a closer look at PLSA. It has already been known for decades that it theoretically resemblances KL-NMF [77]. More recently, it has been shown in many fields that there are in fact differences between both methods [222]. PLSA and KL-NMF optimize the same objective but are only identical methods if KL-NMF adds an additional column normalization and if they would use the same algorithm. MU and EM are different and will result in different solutions when tested on the same data with the same initialization [219]. This statement is supported by looking at the reconstruction errors; every run they each seem to go to their own optimum. The same applies to the visual results, where we can clearly see that they are converging to different local optima. When running until the update between two iterations reaches a threshold of 1e-3, KL-NMF shows a better result than PLSA, which is still missing information. PLSA is a popular method for MSI, where good results are obtained using it, while in text-mining it is becoming less and less popular. Looking at the mathematics, PLSA will work best when there are a lot of samples available [81]. This especially is the case for MSI data, where each spectrum can be seen as a sample. Additionally, PLSA is often critiqued in text-

mining for being too sensitive to noise in comparison to LDA; it approximates the given data set following the objective and does not try to find a generative model for the data. However, in the case of MSI, we specifically want every m/z bin to be equally important. Unfortunately, due to the normalization and the use of the EM algorithm, it does not perform consistently good visually. Although it is a bit faster in runtime, the statistical, normalized counterpart of KL-NMF, shows less accurate spatial distributions when we look at all samples tested, with an exception for P1 where the quality of its spatial distribution matched that of KL-NMF. The spectral basis elements resemble the ones of KL-NMF, albeit in a different scale due to its statistical nature. PLSA does not show absolute contributions of m/z values in its spectral components, but rather the probabilities of each bin contributing.

Finally, LDA is often seen as a generalization of PLSA, as it is able to capture more general trends, emitting the overfitting of noise. For MALDI-MSI, however, generalization is not preferred. In the details of the data, crucial information might still be present. To define what is noise and what is meaningful is very hard for MSI data. Here arises a difference with the field of text mining, where it might be easier to determine what is essential information and what are filler words. MSI is an exploratory technique where little known information is available, and it is harder to check whether something is essential or not. In our experimental set-up, LDA performed worst. Computationally, it took much longer to run until convergence. For the lymph-node data, it took approximately 12 hours to run the 12 iterations. This is the reason the iterations were limited to 12. Even with a sparse input, the computation time stays within the same order of magnitude. Looking at the UMAP resemblance, the spatial components often performed better than PLSA, as they are smoother and have better contrast, but not as good as KL-NMF. Spectrally, the distributions display little information on the chemical properties making it harder to interpret the spatial results. It has been used on MSI for the clustering of the data, improving memory usage with respect to hierarchical clustering and for tandem MS to cope with fragmentation effects [40, 196]. From Table 2.2 and 2.3, we see that the memory usage is lower than for PLSA, but NMF is still performing better.

Aside from the mathematical comparison, we also want to take a closer look at the biochemical interpretation of the resulting spectra. For the pancreatic tissue M2, we refer to the works of T. Smets et al. [174, 175]. We focus on the human-lymph-node sample.

It has become obvious that MALDI based Mass spectrometry imaging with a focus on the lipid mass range (600-1200 m/z,"Lipid Imaging") is also very well suited to acquire profiles directly from the tissue. These profiles can be applied for facilitating diagnosis, prognosis and follow-up of diseases and therapy. An example of this 'lipid profiling' of selected anatomical regions is shown in

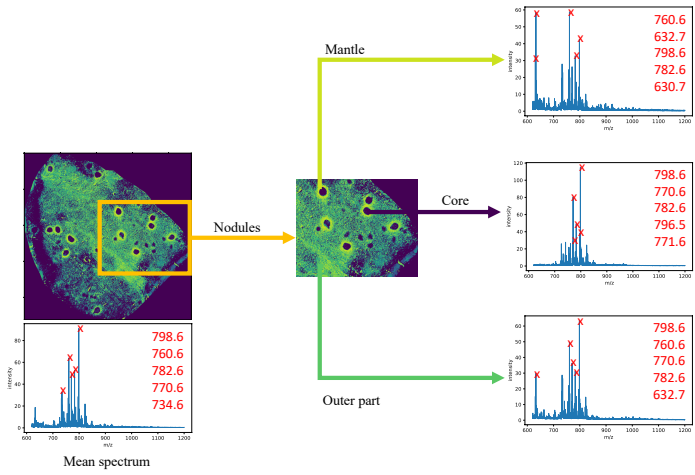


Figure 2.7: Detailed analysis of the human lymph sample P1. From the entire sample, with its mean spectrum given, we deduce three regions present in the orange box around the nodules: 1. the mantle zone, 2. the core zone, and 3. the outer part. In red, we show the m/z values belonging to the top five intensities present in each region.

Figure 2.7 and Figure 2.8. The differences in MS-profiles around the lymph nodules inside the tissue are explored. As shown, the majority of the molecular information (= the number and the height of mass peaks) in positive ion mode, is concentrated in a relative small mass range (700-850 m/z). This range indeed corresponds to masses of (membrane) lipids (especially phosphocholine derivates (PC's) in positive ion mode). The profiles extracted from various regions, demonstrate that an important number of lipid mass peaks are common between the various anatomical structures, and that their relative abundance is of crucial importance. On the other hand, some masses seem to have a higher discriminative value. The masses of 630.7 and 632.7 m/z are found particularly in the mantle zone (see Figure 2.7). Comparison of normal versus tumoral lymph node tissue in the P1/P3/P4-samples (P2 did not show tumoral tissue) suggest a slightly stronger appearance of m/z 741.6, 761.6, 820.6 and 849.6 in the tumoral transformed regions (see Figure 2.8 for P1). These observations all underscore the importance of the data processing methods such as the UMAP based methods described earlier or the KL-NMF based method described in this paper, which yield excellent decomposition and high quality visualization with high overall sensitivity.

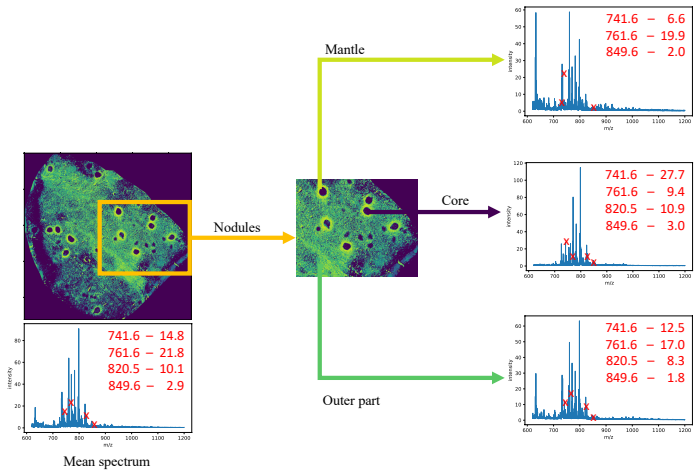


Figure 2.8: Detailed analysis of the human lymph sample P1. From the entire sample, with its mean spectrum given, we deduce three regions present in the orange box around the nodules: 1. the mantle zone, 2. the core zone, and 3. the outer part. In red, we show respectively the m/z values and intensities of the four m/z values of our interest: 741.6, 761.6, 820.6, and 849.6.

It should be noted, that a positive identification of the mass peaks is less straightforward as multiple lipids can share the same or nearly the same molecular weight. For instance, lipids with the same number of double bounds but with a difference in the position of these double bonds, all have identical masses. A suggestion of the possible identity of the mass peaks can be retrieved be consulting the Lipid Maps database (www.lipidmaps.org), but for final identification complementary analytical strategies are required, such as ESI-LC-MS with internal lipid standards to support the identity and provide (semi)-quantitative data on the lipid content.

2.6 Conclusion

Within the field of MSI, factorization methods are still a popular tool to interpret the chemical properties present in a tissue sample. We focussed on four NMF-related methods (NMF, KL NMF, PLSA, and LDA) and evaluated their performance on multiple criteria and on multiple MALDI-MSI data sets. Each method has its significant contribution to the analysis of MSI data, but

rarely are they compared to one another. We have shown that NMF is a fast, easy and accurate factorization method. However, it implies Gaussian noise whereas MSI data follows a Poisson distribution.

KL NMF and PLSA both take this characteristic into account, but unlike often assumed, do not result in the same decomposition. LDA can be seen as a generalization on PLSA, and is often considered the better of the two, but does not offer an advantage in the context of MALDI-MSI data factorization. In our experiments, KL NMF was able to represent the data most similar to the UMAP visualisation, while surprisingly, it has rarely been used for MSI data analysis so far.

Were UMAP exhibits state-of-the-art MSI visualisations, factorization methods such as KL-NMF improve data interpretation. They simultaneous decompose MSI data in both spatial and spectral distributions, offering more possibilities for further analysis of the biochemical structures present in tissues. In this work, we restricted ourselves to the four methods, but extrapolating this investigation to the implications and interpretations of other methods will offer better prospects for the interpretation of MSI data.

2.7 Acknowledgements

This work was supported by KU Leuven: Research Fund (projects C16/15/059, C3/19/053, 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 S005319N, Infrastructure project I013218N, TBM Project T001919N; PhD Grants (SB/1SA1319N, SB/1S93918, SB/151622)), This research received funding from the Flemish Government (AI Research Program). Bart De Moor and Melanie Nijs are affiliated to Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium. 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: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 885682); (EU H2020-SC1-2016-2017 Grant Agreement No.727721: MIDAS), KOTK fundation. 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. Figure 2.6 was obtained with permission

from https://pubs.acs.org/doi/10.1021/acs.analchem.8b05827 and further usage related to the material excerpted should be directed to the ACS.

Chapter 3

Spatiochemical characterization of mouse pancreas MSI data using topological data analysis

This chapter has been previously published as:

Derwae, H., Nijs, M., Geysels, A., Waelkens, E., and De Moor, B. Spatiochemical Characterization of the Pancreas Using Mass Spectrometry Imaging and Topological Data Analysis. *Analytical Chemistry 95*, 28 (July 2023), 10550–10556. Publisher: American Chemical Society

Contributions: Topic conceptualization, supervision of the experiments, proofreading of the manuscript, revision process

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.

3.1 Introduction

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 [29, 143]. 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) [195, 174, 198, 148, 9].

Topological Data Analysis

Topological Data Analysis (TDA) has proven efficient in analyzing high dimensional biological and medical data [173, 30, 156]. For the latest work on TDA for biomedical applications, we refer the reader to Singh et al. (2023) and Skaf and Laubenbacher (2022) [171, 173].

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 [13].

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 [142]. 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 [146]. The topology of a simplicial complex, and its corresponding data set, can be analyzed using its homology [142]. Zero-dimensional homology 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 [146, 129, 150]. 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 [129, 150, 200].

Two main techniques are used within the field of TDA. The first technique is called persistent homology [142, 200, 39] and was first introduced in 1990 [72, 73, 74, 202, 140]. Recent developments such as persistent landscapes [28] 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) [180, 107]. The second technique, called Mapper, was introduced by Singh et al. in 2007 [170]. Mapper has shown to be promising for clustering and classification applications [93, 60] 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. 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) [175]. UMAP too is based on TDA principles, as it builds fuzzy simplicial complexes [135]. A more extensive explanation about TDA can be found in Skaf and Laubenbacher (2022) [173].

3.2 Experimental section

3.2.1 Data acquisition

The used data comes from MSI experiments performed on healthy mouse pancreatic tissue, using a Bruker rapifleX MALDI-TOF mass spectrometer,

with a mass resolution up to 50.000. For the mouse pancreatic tissue, 7 μm thick cryosections are prepared and mounted on ITO glass slides, with sinapinic acid (SA) as the matrix applied using Bruker ImagePrep. The pixel size is set to 50 μm, the recorded m/z range is 2-20 kDa in positive linear mode, and the acquisition speed is 9 pixels/sec, with 1000 laser shots per pixel and a laser repetition rate of 10 kHz. The data sets correspond to Sample 2 and Sample 1 in Smets et al. (2020) and (2019), [174, 175] 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.

3.2.2 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 3.1.

In order to make the Mapper graph, the Python toolbox giotto-tda [186] was used. To cluster the resulting Mapper graph, the scikit-learn implementation [152] of the spectral clustering method was used [206]. The Calinski-Harabasz (CH) index [12] was used as a quantitative metric for the clusters, again, using the scikit-learn implementation [152, 32]. 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:

$$BCD = \frac{\sum_{k=1}^{K} n_k |c_k - c|^2}{K - 1} \tag{3.1}$$

$$WCD = \frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} |d_i - c_k|^2}{N - K} \tag{3.2}$$

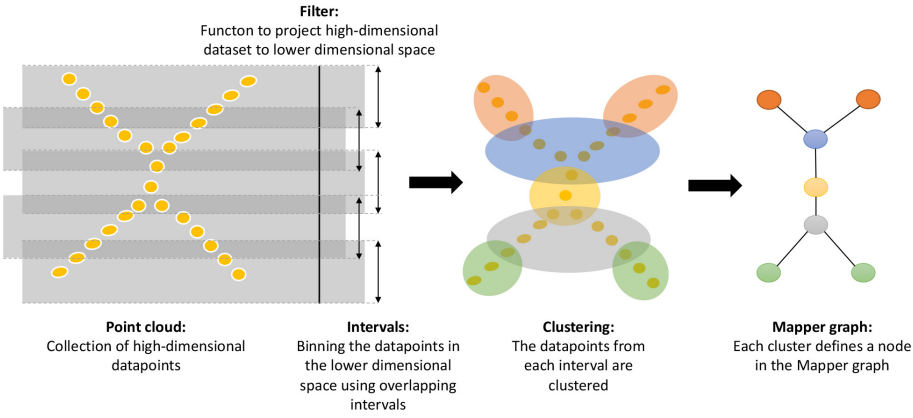


Figure 3.1: Example of the Mapper algorithm for a two-dimensional data set. Figure modified from Chazal and Michel [39].

$$CH = \frac{BCD}{WCD} \tag{3.3}$$

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.

3.2.3 Mapper algorithm

In order to have a benchmark clustering score, we reproduced the spatial clustering part of Smets et al. (2020) [175] 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

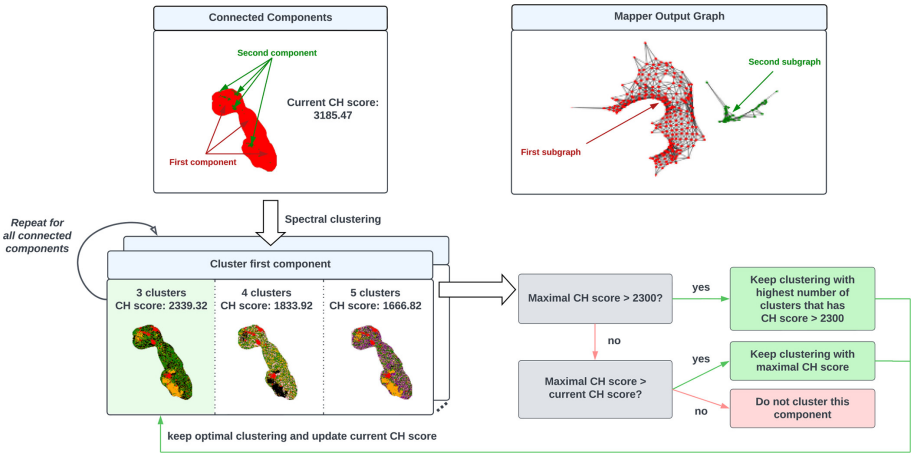


Figure 3.2: Diagram of the clustering workflow together with the original constructed Mapper graph. The parameters to construct the Mapper graph were UMAP, 25 intervals, 25% overlap. The connected components correspond to the subgraphs of the Mapper graph. The original "current score" is the CH score obtained from the connected components. When a clustering is chosen for a component, the current score gets updated.

can be seen in Figure 3.2. Only components containing more than 10 nodes were considered for clustering.

3.2.4 Parameter selection

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) [208] and UMAP [135]. 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 [208].

In order to cluster the data points from each interval, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method [68] 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 [152, 125], 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. 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) [174, 175].

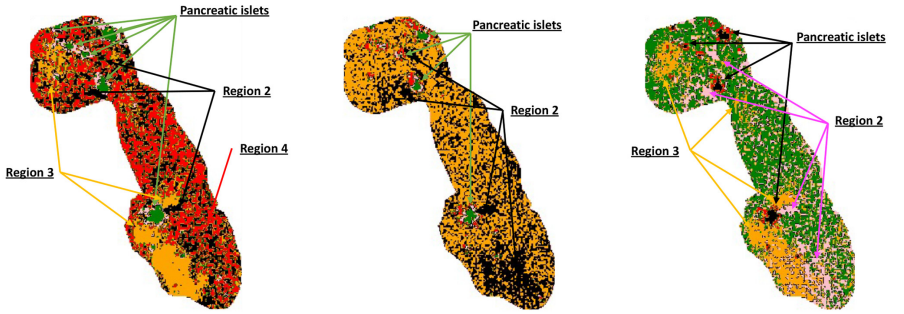
The number of intervals and the overlap percentage were determined using an extensive parameter search. A test interval for both parameters was determined based on an initial random search. Afterward, the algorithm was executed for all parameter combinations within these intervals. The final CH scores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

3.3 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 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 3.2.

3.3.1 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 [44, 27, 31]. Further analysis is required to uncover whether the ring structure is based on a biological structure



(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 3.3: Visualisation of 5 clusters obtained from Mapper graphs using different settings. (a) UMAP filter function, 2 dimensions, 15 intervals, and 25% overlap. (b) SVD filter function, 2 dimensions, 25 intervals and 85% overlap. (c) SVD filter function, 4 dimensions, 5 intervals, and 55% overlap.

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 3.3a, 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 3.3b. 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 3.3c.

3.3.2 Quantitative comparison

In order to compare the obtained results with the spatial clustering part of Smets et al. (2020) [175], 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 3.3a 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

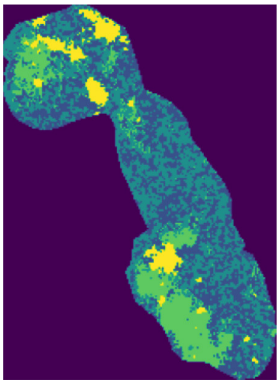


Figure 3.4: Benchmark clustering made using UMAP and k-means clustering, based on the spatial clustering part of Smets et al. (2020) [175].

data set can be analyzed in less than 30 seconds 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.

3.3.3 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 3.4. 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 3.3a.

When using svd as a filter function, on the other hand, region 2 was clustered differently, see Figures 3.3b and 3.3c. 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 [33]. 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 3.3c. In Figure S4, the evolution of clusters can be seen for this experiment. The CH

score gradually decreases as the number of clusters increases.

3.3.4 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 3.5. 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. $DICE = \frac{2|A||B|}{|A|+|B|}$. The resulting Dice score was 0.89, which is close to 1, meaning a high overlap [37].

3.3.5 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

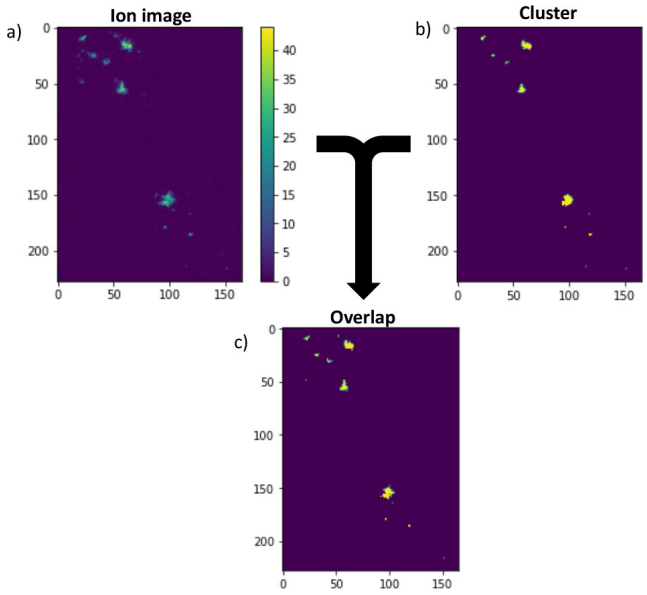


Figure 3.5: (a) Ion image for m/z value 6006.67. (b) Mask for the cluster representing the cores of the pancreatic islets. (c) Overlap between intensities higher than 15 from the ion image and the cluster.

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 3.6. 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.

3.4 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

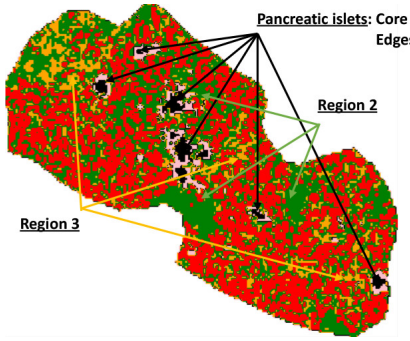


Figure 3.6: Clusters obtained from the Mapper graph using svd as a filter function for the second data set with the following parameters: 4 dimensions, 15 intervals, 70% overlap, 5 clusters.

exploratory 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.

3.5 Acknowledgements

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.

Chapter 4

Hierarchical bi-clustering of mouse pancreas MSI data using recursive rank-2 NMF

This chapter has been submitted for publication as:

Nijs, M., Waelkens, E., and De Moor, B. Hierarchical bi-clustering of mouse pancreas mass spectrometry imaging data using recursive rank-2 non-negative matrix factorization. Submitted to Journal of the American Society for Mass Spectrometry

The work in this chapter has been awarded *Best poster abstract* at the IMSIS conference, 2023, Montreal, Canada.

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.

4.1 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 [36]. With the advantage of measuring all these molecules simultaneously, MSI allows for the exploratory analysis of tissue sections and aims to uncover novel biomarkers [10].

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 such as ion suppression and noise artifacts.

There are indeed multiple factors increasing the complexity of the MSI signal, e.g.: sample preparation, matrix effects, biological noise, fragmentation of molecules, adducts, isotopes, and so on [15, 29]. To minimize artefacts resulting from sample preparation, the MSI field is working towards more standardized procedures [194]. Unfortunately, the other sources of noise are inherent to the data and its technology. This underscores the importance of data scientists to develop software that 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 [98, 201].

Not only noise artifacts but also high abundance of particular molecules may cause 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), can 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 applied on just a sole dimension rather than both spectral and spatial [5, 175].

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 [137] method. Although bi-clustering has been demonstrated successfully on MSI data in earlier works [126, 151], 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 data sets. 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 [155, 84, 64]. 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 [175, 50, 148] 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.

4.2 Methods

4.2.1 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, with a mass resolution up to 50.000. The tissue is prepared by mounting 7 μm thick cryosections on ITO glass slides, where sinapinic acid (SA) serves as the matrix applied via Bruker ImagePrep. The pixel size is 50 μm, with a recorded m/z range of 2-20 kDa in positive linear mode, and an acquisition speed of 9 pixels/sec, utilizing 1000 laser shots per pixel at a laser repetition rate of 10 kHz.

All three data sets 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 [9], we apply the contrast enhancement method CLAHE (Contrast Limited Adaptive Histogram Equalization), as provided by https://github.com/anntzer/clahe.

Throughout this manuscript we use additional abbreviations for the + and 2+ charge states of both insulins. We write the single (+) charged state [Ins + H] as Ins^{+}, and similarly, the double (2+) charged stated $[\text{Ins} + \text{H}]^{+}$ as Ins^{2+}.

4.2.2 An 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 data set [132]. 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 $\mathbf{X} \in \mathbb{R}_{+}^{M \times N}$, with M rows, the m/z values, and N columns, the (vectorized) pixels, in two non-negative factor matrices: the spectral basis elements $\mathbf{W} \in \mathbb{R}_{+}^{M \times K}$ and the weight matrix $\mathbf{H} \in \mathbb{R}_{+}^{K \times N}$. In short, we write:

$$\mathbf{X} \approx \mathbf{WH}. \tag{4.1}$$

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 [148, 81].

The usefulness of NMF and its related methods for the analysis of MSI data has been investigated in earlier work [148], where NMF is used for dimensionality reduction and visualization of the main trends in the samples. In continuation

of the 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 developed a recursive rank-2 NMF algorithm.

Shown in pseudocode below, rr2-NMF starts with a rank-2 NMF on the full dataset. This is a matrix with dimensions $M \times N$, as defined above. The resulting factor matrices $\mathbf{W}$ and $\mathbf{H}$ will respectively have dimensions $M \times 2$ and $2 \times N$ after every iteration of the rr2-NMF algorithm. Based on the obtained spectral and spatial information of the two NMF components and utilizing the linearly additive property of NMF, we sub-select from the original data two new datasets of co-localized peaks. For each subset, as defined in the pseudocode, we take the pixels and m/z-values most contributing to each of the two components by putting the other values to zero.

The rank-2 NMF decomposition is repeated on each subset, until either: (Algorithm 1) a predefined depth α is reached, or (Algorithm 2) the subcluster 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: the spectral distribution, the spatial distribution and 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.

4.2.3 Implementation details and code availability

All implementation has been done using Python and the scikit-learn [152] 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 $+1 = \alpha$ **then**
    $W, H \leftarrow$ NMF(data, rank=2)
    PLOT_COMPONENTS(W, H)
  else
    $W, H \leftarrow$ NMF(data, rank=2)
    $subset1 \leftarrow$ data$[W_1 > W_2, H_1 > H_2]$
    $subset2 \leftarrow$ data$[W_1 \leq W_2, H_1 \leq H_2]$
    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)
  $W, H \leftarrow$ NMF(data, rank=2)
  $subset1 \leftarrow$ data$[W_1 > W_2, H_1 > H_2]$
  $subset2 \leftarrow$ data$[W_1 \leq W_2, H_1 \leq H_2]$
  if SIMILARITY(data) $-$ SIMILARITY($subset1$) $< \beta$ **then**
    PLOT_RESULT($subset1$)
  else
    RR2-NMF($subset1$)
  end if
  if SIMILARITY(data) $-$ SIMILARITY($subset2$) $< \beta$ **then**
    PLOT_RESULT($subset2$)
  else
    RR2-NMF($subset2$)
  end if
end function

4.3 Results and discussion

4.3.1 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 4.1, a graphical representation of the algorithm is shown where rr2-NMF is applied on the M2 data set with as a stopping criterion depth of two layers. Here, we see step-by-step how rr2-NMF iteratively splits the data set 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 4.2. 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 $m/z = 3485.7$ Da (glucagon) becoming more significant in these subclusters as opposed to non-hierarchical approaches, see also Figure 4.3 A. and C.10. The same is the case for samples M1 and M3 (see Supporting Information S3 and S5).

The method confirms itself in a way that expected signals originating from one biomolecule are clustered together, e.g.: multiple charge states, and adducts such as sodium and potassium. In Table 4.1, we list the m/z peaks of cluster "M2_convergence_2_1". Here we see the value corresponding to Ins2 (m/z = 5800.8 Da), with the 2+ charge state (m/z = 2901.5 Da), as well as the added potassium (m/z = 5838.1 Da), and both potassium and sodium (m/z = 5857.4 Da).

4.3.2 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 by production of insulin and releasing 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.

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) [167]. As known for rodents (and not the case for humans), we

Table 4.1: List of the *m/z* values corresponding to the peaks of cluster "M2_convergence_2_1".

m/z value [Da]	**intensity**	**(possible) annotation**
5800.8	391.0	$Ins2^{+}$
5814.9	43.0	$Ins2^{+}$ + 14.15
5838.1	20.0	$Ins2^{+}$ + K
5782.8	11.0	$Ins2^{+}$ – H_2O
6005.4	11.0	
5857.4	10.0	
6023.4	8.0	
2901.5	5.0	$Ins2^{2+}$
3920.7	4.0	
5755.7	3.0	
5896.0	3.0	
5910.2	1.0	
5977.1	1.0	

Table 4.2: 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	$Ins1^{+}$
5843.2	23.0	
3924.5	22.0	
5822.7	17.0	$Ins1^{+}$ + 14.15
5863.8	7.0	
5898.6	3.0	
3122.8	2.0	
5916.6	2.0	
2906.6	1.0	$Ins1^{2+}$
3485.7	1.0	glucagon

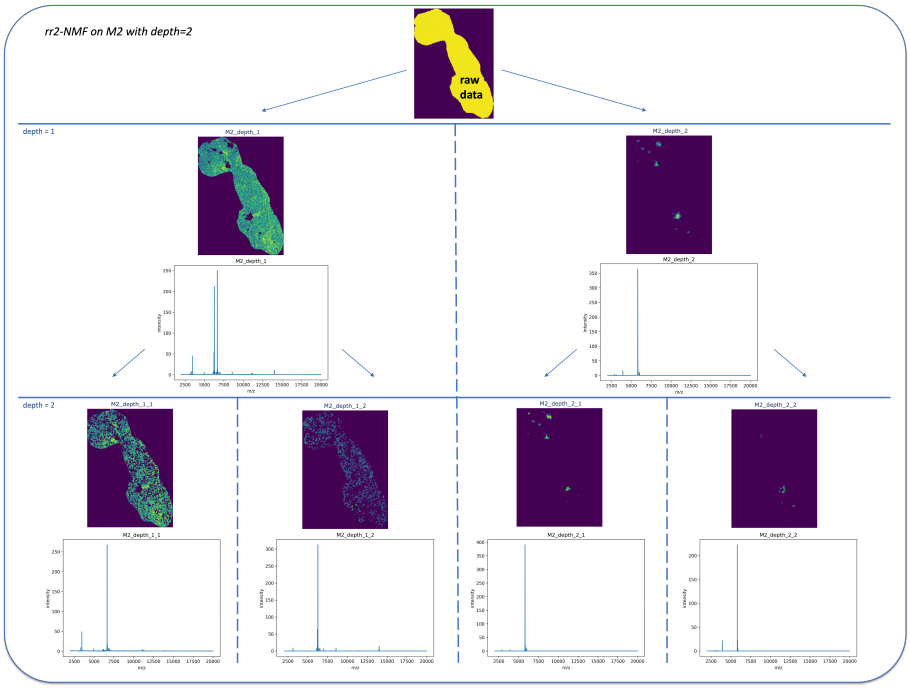


Figure 4.1: Graphical overview of the rr2-NMF algorithm of the M2 pancreatic tissue sample, simultaneous clustering of the spectral and spatial dimensions. With stopping criterion: a fixed depth α of 2 levels. High resolution images are available in the Supporting Information (S3a).

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 [33].

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 4.3, 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 [187], Table 4.2 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 4.3. 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.

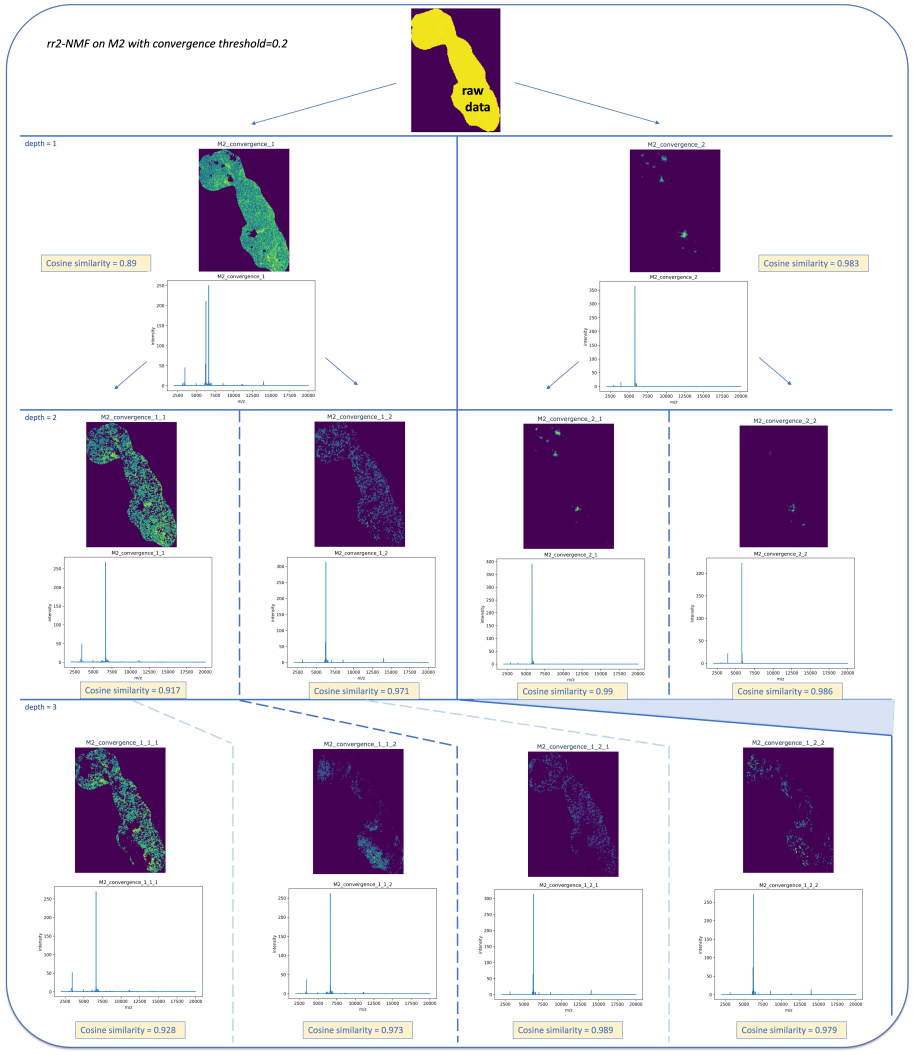
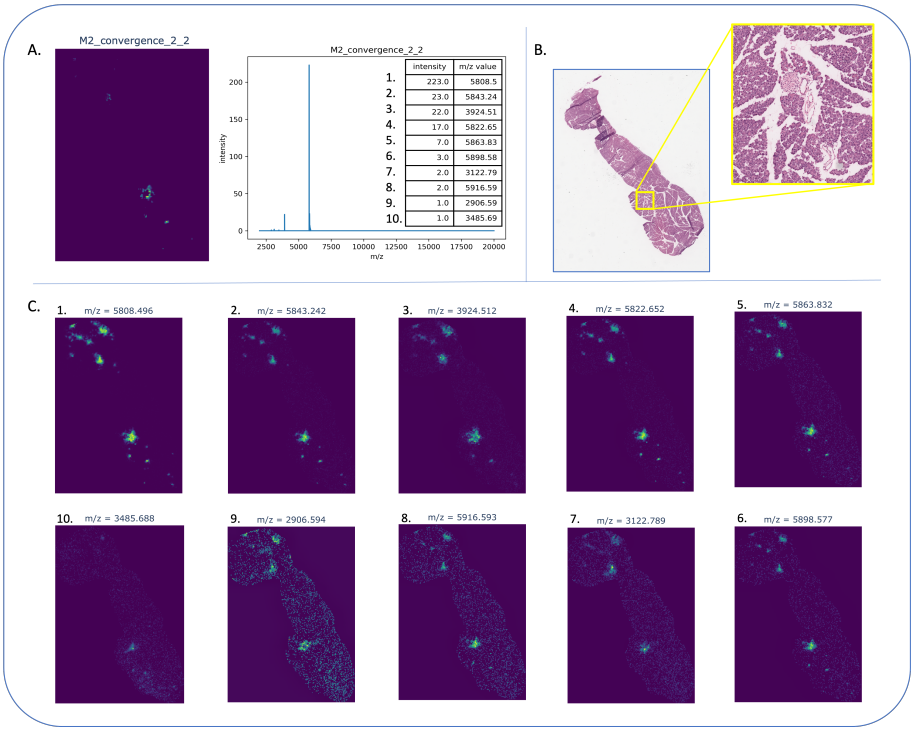


Figure 4.2: Graphical overview of the rr2-NMF algorithm of the M2 pancreatic tissue sample, simultaneous clustering of the spectral and spatial dimensions. With stopping criterion: the convergence (increment $\beta < 0.02$) of the cosine similarity of the spectra within each cluster. High resolution images are available in the Supporting Information (S3b).



	intensity	m/z value
1.	223.0	5808.5
2.	23.0	5843.24
3.	22.0	3924.51
4.	17.0	5822.65
5.	7.0	5863.83
6.	3.0	5898.58
7.	2.0	3122.79
8.	2.0	5916.59
9.	1.0	2906.59
10.	1.0	3485.69

Figure 4.3: 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 = 2906.6 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.

Although lowly abundant, it is not a surprise that m/z = 3485.7 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), Figure 4.3C.1., it suggests an increased intensity along the outer border of the islets. The same is true for the 2+ charge state of Ins1 (m/z = 2906.6 Da), Figure 4.3C.9. 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).

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 4.5 shows similar clustering for the M1 sample, albeit with a small mass shift, 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.

Differences in expression levels between Ins1 and Ins2

The study of the spatial distrubution and localization of Ins1 and Ins2 is not straightforward as the masses of 5800.8 Da (Ins2) and 5808.5 Da (Ins1) have partially overlapping isotope distributions, see also the Supporting Information S2. This explains mainly why earlier approaches sometimes could not distinct the two. In order to compare the spatial distributions of both molecules, we have analyzed 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 4.4, 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 $m/z = 5795.6$ Da) is clearly higher than the respective isotope peak of Ins1 (59 for $m/z = 5813.6$ Da).

This is suggestive for a difference in expression level of Ins2 and Ins1 in the center and in the border of the islets. Eventhough a distinct regulation of the expression levels of Ins1 and Ins2 have been reported earlier [49, 161], a different spatial-related expression in normal mouse pancreatic islets has not been studied yet and opens new horizons for further biomedical research.

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 [217].

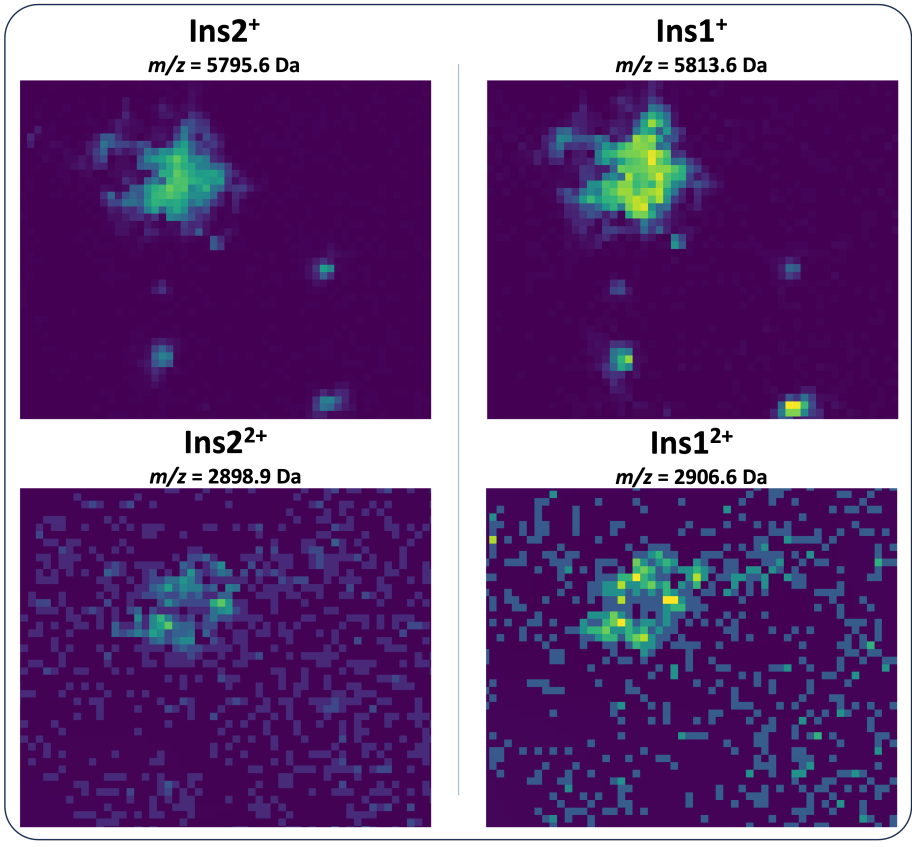


Figure 4.4: Zoomed-in islets of Langerhans of the M2 sample for Ins1 and Ins2 of the isotopes at – 5 Da (Ins2) and + 5 Da (Ins1) of their peak values in + charge state, and – 2.5 Da (Ins2) and + 2.5 Da (Ins1) for the 2+ charge states. We see higher correlation of Ins1 with the outer border of the islets than Ins2.

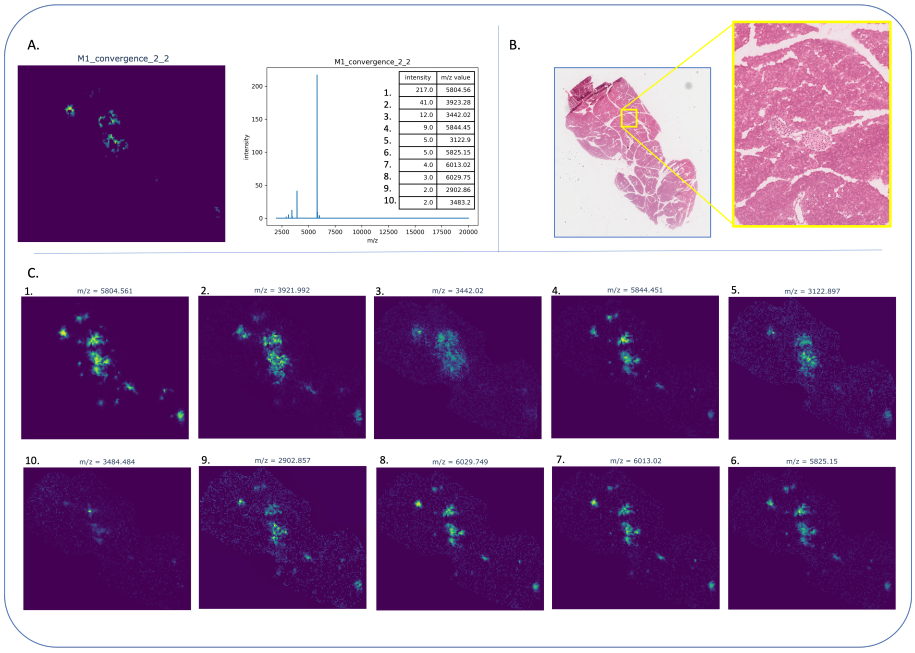


Figure 4.5: Spatial correspondence of cluster "M1_convergence_2_2" and the ion images of the corresponding m/z peaks. The cluster shows the potentially higher presence of Ins1 in the outer periphery of the islets of Langerhans in this sample as well, albeit with a shifted mass range. (A) Result cluster obtained with rr2-NMF. (B) H&E staining of the M1 sample. (C) Ion images of the peak values corresponding to this cluster.

4.3.3 Computational efficiency and limitations

The gain in information compared to some traditional data-analytical tools comes at the small cost of more computational resources. This, however, is negligible. 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.

4.4 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 4.1, Table 4.2 and the intensities tables in Figure 4.3 and 4.5).

Furthermore, the application of the method to the mouse pancreatic MSI data sets 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.

4.5 Acknowledgements

This work was supported by KU Leuven Research Fund (pro-jectsi BOF/23/064, C16/15/059, C3/19/053, C24/18/022, C3/20/117, C3I-21-00316), Industrial Research Fund (Fellow-ships 13-0260, IOFm/16/004) and several Leuven Research and Development bilateral industrial projects; Flemish Gov-ernment Agencies: FWO: EOS Project no G0F6718N (SeL-MA), 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).

Part II

Applications to other omics modalities

Chapter 5

An NMF-MLP pipeline for fast exploration and annotation of single cell MILAN data

This research was conducted as part of a KU Leuven C3 project: "MILAN: next-generation pathology by MILAN multiplex analysis and advanced trainable image analysis" in collaboration with the department of Imaging and Pathology at UZ Leuven.

This chapter demonstrates the use of non-negative matrix factorization in the context of targeted **spatial proteomics** data analysis obtained with MILAN immunohistochemistry technology. Within the project, a pipeline has been developed for unsupervised clustering of the pixels and supervised classification of the components, i.e. the different cell-types.

5.1 Introduction

At the end of 2016, the human cell atlas initiative [115] was launched, aiming to describe all different types and characteristics of human cells, both healthy and sick. In this chapter, we look how multiplexed immunohistochemistry (IHC) data allows for the expression profiling of single cells in their natural micro-environment.

Using immunofluorescence, IHC detects specific antigen-antibody reactions, as shown in Figure 5.1. This makes it a valuable tool in many applications from research to clinical diagnostics, e.g.: biomarker discovery, therapy response, drug development, etc. [61].

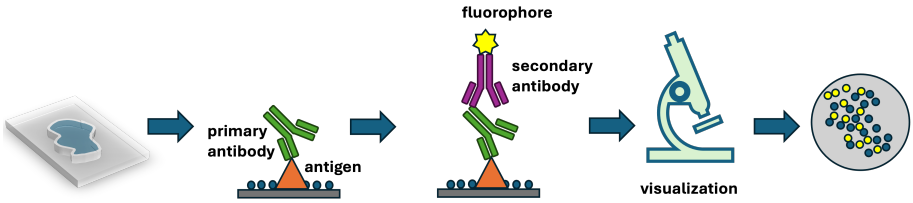


Figure 5.1: Principle of indirect immunohistochemistry using a primary and secondary antibody to detect specific antigens with immunofluorescence.

Although antibody and morphology based methods for spatial expression profiling have been around for decades, older approaches of IHC only allowed for the simultaneous measurement of a few proteins in a single tissue slide [159]. MILAN technology, which stands for Multiple Iterative Labeling by Antibody Neodeposition, can measure multiple (> 80) markers on the same slide. This offers huge advantages towards interpretation of the status and interactions of cells in a tissue.

In brief, MILAN is a multiplexed IHC (mIHC) technology that consists of a cyclic repetition of (i) antibody staining of FFPE (formalin-fixed paraffin-embedded) tissues, followed by (ii) high resolution fluorescence image extraction, and (iii) antibody removal while maximizing antigen preservation. Every repetition, three markers are measured, after which the slides are rinsed and new markers are added to the same tissue slides. After all rounds, the output is stored for computational analysis. The high resolution images can easily mount up to Terabytes in storage memory and thus requires efficient algorithms for further analysis. Figure 5.2 shows the overview of the MILAN process.

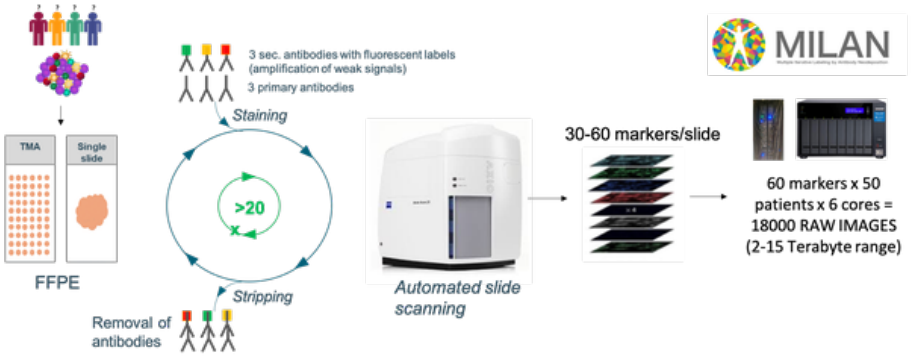


Figure 5.2: **MILAN process for spatial proteomics.** Left: tissues are obtained from patients and embedded in FFPE for stability, either on a single slide or as a tissue micro array (TMA). Middle: the cyclic repetition of antibody staining with fluorescent tags, slide scanning and stripping of three markers. Right: after 20 rounds of repetition, 60+ markers are obtained for each slide and the high resolution datasets are exported. Source: MILAN website [2].

Because of the cyclic routine, where inter-cyclic distortions are inevitable, preprocessing the data correctly is of the utmost importance. After the data acquisition, a preprocessing pipeline is executed to correct all images before the downstream analysis takes place. Figure 5.3 shows an overview of commonly applied preprocessing steps and data analysis [23]. We discuss the most important steps in this pipeline:

- **Image registration (REG)**

  Because the images are acquired in multiple scanning rounds, it is important that the different marker expressions are spatially aligned with minimal error. To this end, one could perform either rigid or elastic registration of the images [141]. Given the possible deformations of the sample between subsequent rinsing, elastic registration will perform better.

- **Autofluorescence (AF) removal**

  Aside from the antibody fluorescence tags, a tissue has also inherent, autofluorescence properties emitting light under certain wavelengths. This disturbs the true signals of the markers and should be corrected for. To do so, an untagged slide is measured to find the AF signal and subtract it from the other measurements. For different wavelengths, different autofluorescence can be excited. This means that AF should be subtracted for each round in measurement.

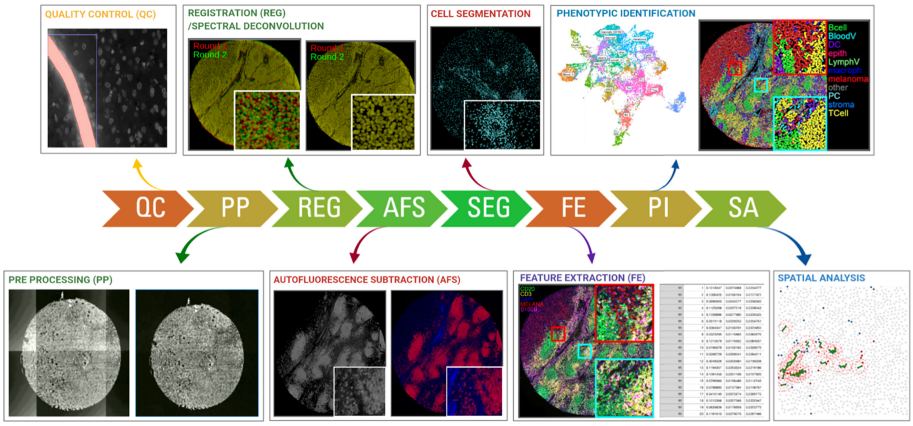


Figure 5.3: **Most commonly used steps in the data processing of fluorescent, cyclic methods for multiplexed IHC.** Images are collected across multiple cycles but still need to be cleaned (QC), corrected (PP), registered/aligned (REG), autofluorescence removed (AF), segmented (SEG), feature extracted (FE), phenotypically annotated (PI), and spatially resolved (SA). Figure reprinted with permission form [23]. Copyright 2022 Bosisio, Van Herck, Messiaen, Bolognesi, Marcelis, Van Haele, Cattoretti, Antoranz and De Smet, under the terms of the Creative Commons Attribution License (CC BY).

- **Cell segmentation (SEG)**

  For the correct identification of each cell, a good segmentation method is required to detect the cell edges, cell features or tissue regions. Deep learning techniques, such as convolutional neural networks (CNNs), perform well for this task and are the current state-of-the-art.

- **Downstream analysis (FE, PI, and SA)**

  After all the preprocessing steps are done, further downstream analysis can be performed. In case of multiplexed IHC for spatial proteomics, this entails feature extraction (FE), phenotype identification (PI) and spatial analysis (SA).

The pipeline presented in this chapter, aims to tackle all three data downstream analysis tasks at once. Using non-negative matrix factorization (NMF), we get both spatial (SA) and marker (FE) distributions of the main trends in the data. In combination with a subsequent multi-layer perceptron (MLP) classifier trained on the provided annotations, we can classify each NMF marker distribution to the most corresponding cell-type and give a first estimate for the identification of the phenotypes (PI).

MILAN vs MALDI spatial proteomics

Both mass spectrometry imaging (MSI) and IHC allow for spatial proteomics: to study proteins in their micro-environment. While they are different assays, the information they provide can be seen as complementary.

While MSI allows for the simultaneous and untargeted measurement of thousands of molecules present in a slice of tissue, identification of these molecules remains challenging. In contrast, IHC using MILAN can only measure in a targeted manner by applying preselected markers, and only around 80 at once. However, it has the ability to then identify the molecules easily based on the antigen properties. MILAN delivers microscopic resolutions, while MALDI-MSI is often measured at a higher spatial resolution [23].

Additionally, MSI measurements depend on the chemical characterisations of the molecules such as their weight and ability to ionize. Using IHC, proteins with high molecular weight can still be detected with the antibody markers, and will provide immediate identification. In MALDI-MSI, subsequent analyses are required for the specific identification of the (larger) molecules. On the other hand, with MSI the spatial distribution of lipids can be measured as well, which cannot be done using IHC.

Normalization and preprocessing of the data are very challenging in MSI and IHC. In both cases, technical and biological noise is challenging to correct for. In MSI, there can be technical artifacts from sample preparation, matrix effects, background noise, etc. With multiplexed IHC, as mentioned above, there are in addition spatial shifts between the images and autofluorescence effects causing the technical artifacts. Normalization of both IHC and MSI data is still an open problem and a ground truth is often not available. Furthermore, in terms of biological noise, MSI has isotopes and double charged states causing positive correlation of certain m/z values, and similarly, in IHC, markers could be related to each and tag similar cells.

A robust downstream analysis for both IHC and MSI should be able to cope with the potential high presence of noise in the data to discover the underlying spatial clusters and their corresponding fingerprints. This demonstrates that similar approaches, while tailored to each data format, could benefit both MSI and IHC data analysis for spatial proteomics.

To this end, we extrapolate our knowledge from the first part of this dissertation and apply NMF on IHC data. We present a downstream analysis pipeline for the automatic clustering and annotation of melanoma IHC data obtained with MILAN.

5.2 MILAN data of human melanoma samples

The data used for the development of the NMF-MLP pipeline is a human melanoma dataset. This data set has been annotated by pathologist based on previous clustering results, see [22], which allows us to train a classifier for the NMF results. To reduce the data size, the average expression for each cell is stored with the (x, y) coordinates of its nucleus. This averaged data set requires 1.4 Gb in disk space and is loaded from a `.csv` file into Python. We given an overview on the main properties of the data set in Table 5.1.

Table 5.1: Properties melanoma data

Property	**Details**
# cells	$\approx$ 1.5 million cell nuclei for 70 different samples
# markers	84 markers of which 37 phenotypic (first tier) markers
# cell-types	18 different cell-type predictions

- **Markers used**

  ['AID', 'AXL', 'bcl2', 'BCL6', 'CD137', 'CD138', 'CD14', 'CD141', 'CD16', 'CD163', 'CD1A', 'CD1c', 'CD2', 'CD20', 'CD207lang', 'CD21', 'CD23', 'CD248', 'CD25', 'CD27', 'CD3', 'CD303', 'CD31', 'CD32abc', 'CD34', 'CD4', 'CD4543', 'CD5', 'CD56', 'CD64', 'CD68', 'CD69', 'CD7', 'CD74', 'CD79a', 'CD8', 'CD83', 'CD86', 'CK', 'CLEC10A', 'CLEC9A', 'cMAF', 'CXCR5', 'DAPI', 'FCepsilonR', 'FoxP3', 'GATA3', 'GRB7', 'HistH3', 'HLADR', 'ID1', 'ID2', 'IDO', 'IgD', 'IRF4', 'IRF8', 'kappa', 'Ki67', 'LAG3', 'lambda', 'LEF1', 'LYZ', 'Melan-A', 'MX1', 'MYC', 'OX40', 'p27', 'Pax5', 'PD1', 'PDL1', 'PNAd', 'Podoplanin', 'PRDM1', 'pSTAT1', 'PU1', 'S100B', 'SOX9', 'TBET', 'TCF7', 'TCRd', 'TIM3', 'TP53', 'VISTA', 'ZEB1']

- **Phenotypic markers**

  ['CD34', 'CD31', 'CD141', 'PNAd', 'CD25', 'CD14', 'CD1c', 'CK', 'CD21', 'FoxP3', 'CD23', 'GRB7', 'CD1A', 'Podoplanin', 'CD138', 'CD248', 'CD64', 'CD163', 'Pax5', 'IRF8', 'CD20', 'CD8', 'CD303', 'LYZ', 'CD16', 'CD2', 'HLADR', 'IRF4', 'CD5', 'CD79a', 'CD68', 'CD3', 'CD4', 'CD27', 'PRDM1', 'Melan-A', 'S100B']

- **Different cell-type predictions**

  ['B Cells', 'Blood Vessels', 'Classical Dendritic Cells Type 1', 'Classical Dendritic Cells Type 2', 'Cytotoxic T Cells', 'Epithelial Cells', 'Follicular

Dendritic Cells', 'High Endothelial Venules', 'Lymphatic Vessels', 'M1-like Macrophages', 'M2-like Macrophages', 'Melanoma', 'Plasma Cells', 'Plasmacytoid Dendritic Cells', 'Regulatory T Cells', 'Stroma', 'T Helpers', 'other']

5.3 The NMF-MLP pipeline

The proposed downstream analysis pipeline combines both unsupervised ML using NMF for the marker/spatial clustering of the data and supervised ML using an MLP classifier for the cell-type annotation of the obtained marker fingerprints (the obtained NMF components). Figure 5.4 shows a schematic representation of the proposed approach. Implementation has been done in Python using the scikit-learn [152] library.

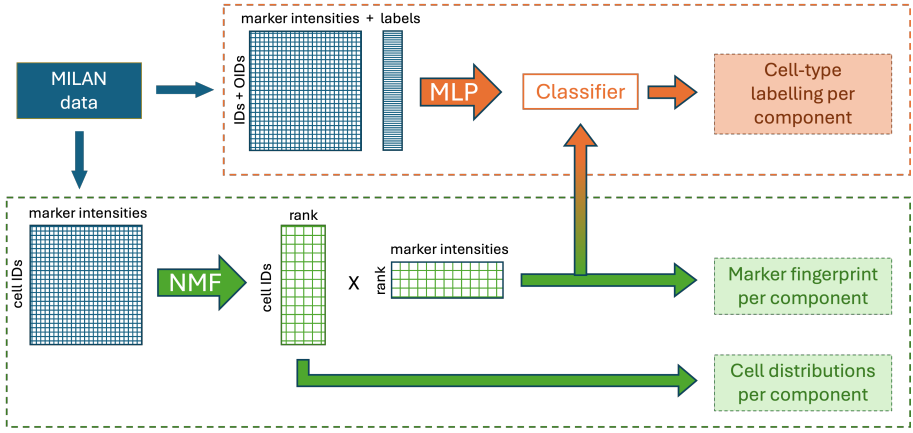


Figure 5.4: **Schematic overview of the proposed NMF-MLP pipeline. Top**: An MLP is trained using the fingerprints and annotations for each cell ID, the resulting model is used to classify new fingerprints. **Bottom**: NMF decomposes the unannotated marker intensities into a rank r number of components: for each component we obtain both cell and marker distributions. The cell distributions allow for the spatial visualization of the component in the tissue and the marker distributions allow for the phenotypic identification either by expert analysis or using the trained MLP classifier.

Here, a first part consists of training an MLP to learn the mapping from each cell's marker fingerprint to its corresponding label. In the second part, we apply NMF to cluster the data into its main components without using the labels. The resulting factor matrices can be used for the spatial visualization of the cells

and the phenotypic identification of each component. Finally, we can combine the two parts for the automatic annotation of the obtained marker fingerprints with the MLP classifier.

The remainder of this section provides more details on the theoretical background and experimental setups for each part of the pipeline.

NMF for fast clustering of single and multiple samples

Similar as for MSI, we define a matrix $\mathbf{X} \in \mathbb{R}_+^{m \times n}$ with m rows describing the measured markers and n columns for every cell ID. Depending on the selection of the rank r, we get a number of components representing the main trends. These components consist of the r row-column combinations of the factor matrices $\mathbf{W} \in \mathbb{R}_+^{m \times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r \times n}$. The columns of $\mathbf{W}$ contain the distinct marker basis distributions and the rows of $\mathbf{H}$ represent the corresponding weight of each cell to that basis.

For the selection of the rank one could plot the singular values in a semi-log scale and use the elbow approach. In the setting for MILAN, we apply different ranks depending on the application. Within the MILAN pipeline, NMF can offer a first insights in the data using a low rank decomposition for the main trends. Additionally, a more in-depth analysis can be obtained using a higher number of components, and this can be done both for a single sample as for a multi-sample analysis.

For the spatial visualisation of the clustering results, we opt for a more compressed approach where all components are plotted on top of each other. In this case, we color each cell according to the component with the highest weight. Due to the high resolution of the spatial dimension and the data only containing the (x, y) coordinates of each cell's nucleus, we need the corresponding area to visualize each cell clearly. We therefore use the segmentation map from the preprocessing pipeline as a mask to color the correct area based on the cell ID.

Classification of NMF clusters using an MLP framework

For the training of the classifier, we use an MLP framework implemented in Python's scikit-learn. To this end, we split the data in a test and train set, with respective proportions of 25 and 75 % of the total data. To have comparable values across the different markers we perform min-max normalization. We additionally apply some data balancing, using the scikit-learn stratification option, to accommodate for the uneven frequencies of the different cell-types.

For the MLP, we use 2 hidden layers each with 64 neurons. Additionally, we use a ReLU (rectified linear unit function) activation function, an L2 regularization $\alpha = 0.001$, and the ADAM solver for weights optimization. After 98 epochs we obtain a validation score of 0.86 and a final loss of 0.39.

With the model, we predict for a new fingerprint the corresponding cell-type by looking at the probability of it corresponding to all possible classes. This gives an indication in the certainty of the predicted label.

5.4 Results & Discussion

5.4.1 Multi-sample analysis

For the melanoma data, an useful software feature is the ability to perform multi-sample analysis and find trends specific to the pathology. We run an NMF with rank 8 on the 70 melanoma samples to detect the main features across all, indicative for melanoma samples.

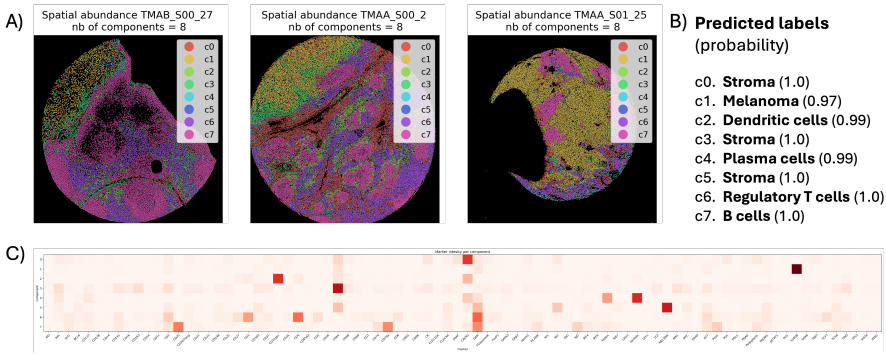


Figure 5.5: **Results for the melanoma multi-sample analysis with rank 8. (A)** The spatial visualizations for 3 of the 70 samples displaying the 8 components in the different colors. Corresponding cells can be identified across the different samples. **(B)** The predicted cell-types for each component based on the marker fingerprint. **(C)** A heatmap visualizing the contributions of all 84 markers to each component.

Figure 5.5(a) shows the combined spatial visualisation of the 8 components for 3 of the 70 samples. See Appendix E for results on the other samples. The output of the MLP classifier for each of these components can be seen in Figure 5.5(b). We give the predicted cell label as well as the probability indicating the

certainty. At the bottom, Figure 5.5(c), a heatmap representation is giving with the contributions of all 84 markers (x-axis) to each of the 8 components (y-axis). Using these fingerprints, we identify the visual clusters with our trained model. Appendix D shows the resulting visualisations of the other samples.

5.4.2 In-depth analysis of a single melanoma sample

For the in-depth analysis of the melanoma sample *TMAA_S00_2* we compare our results to the number of identified cell-types in [22] and take rank $r = 11$ equal to their obtained result. Figure 5.6 shows a comparison between the original (left), NMF (center), and UMAP (right) visualization of the data. The original demonstrates the result from earlier work [22] with 11 different cell-types manually annotated. Table 5.2 shows the predicted labels obtained by the trained MLP classifier as well as the corresponding result from [22] based on the spatial distribution. While not fully identical, our analysis instantly identifies most of the components and localises them in roughly the same areas. Further - independent - analysis is required to deduce the true labels for each cell.

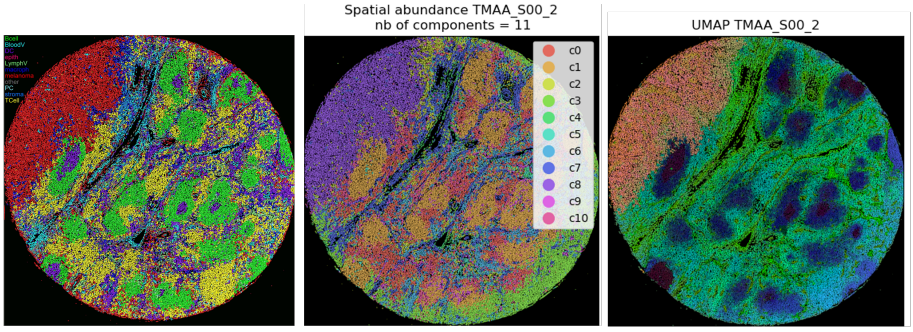


Figure 5.6: In-depth analysis of melanoma sample *TMAA_S00_2* using, from left to right: the original approach (see [22]), NMF with rank=11 and UMAP to 3 dimensions.

When decomposing using NMF in 11 components, we see a rather similar structure to the distributions of the original analysis. An important component is the melanoma tissue, identified by the 'Melan-A' marker. In Figure 5.7, the corresponding fingerprints are plotted for each of the NMF components. In Figure 5.6, c8 visually corresponding to the melanoma tissue, is the only component with this marker present, which confirms the correspondence.

Table 5.2: Predicted labels of NMF decomposition of melanoma sample *TMAA_S00_2* in 11 components and the most corresponding cell-type obtained by earlier work in [22].

component	**classification**	**≈ corr. annotation [22]**
c0	Regulatory T Cell	T Cell
c1	B Cell	B Cell / Lymph Vessel
c2	Stroma	Macrophage
c3	High Endothelial Venule	Melanoma / Epithelial Cell
c4	Stroma	Blood Vessel / Plasma Cell
c5	Plasma Cell	?
c6	Dendritic Cell Type 1	Dendritic Cell
c7	Blood Vessel	Blood Vessel / Plasma Cell
c8	Melanoma	Melanoma / Epithelial Cell
c9	Lymphatic Vessel	B Cell / Lymph Vessel
c10	Cytotoxic T Cell	T Cell

On the other hand, we also notice some differences in the spatial NMF result, for example, a more dominant presence of c3 on the lower right edge not present in the original analysis. This corresponds to DAPI as shown in the corresponding markers of c3 in Figure 5.7. DAPI binds specifically to DNA or cell nuclei. The higher presence on the edge might be an indicator of noise or improper AF removal in those pixel or artefacts from the normalization of the markers. Aside from this, the general structure of the cell distributions match. Here, we note that the original method used only a subset of markers and cells to perform the downstream analysis, which could explain local differences.

Lastly, we have the hyperspectral visualization of the data using UMAP to the 3 RGB dimensions. It is clear UMAP is not able to capture all the nuances of the dataset in the same level of detail as the other two approaches. Additionally, UMAP does not allow for the immediate inspection of the markers, making it harder to identify the obtained components in the data.

Figures 5.6 and 5.7 demonstrate the abilities and usefulness of NMF regarding feature extraction, phenotypic identification, and spatial analysis all in one. The obtained fingerprints can be further inspected for the identification of the components or fed to a trained classifier for this type of tissue.

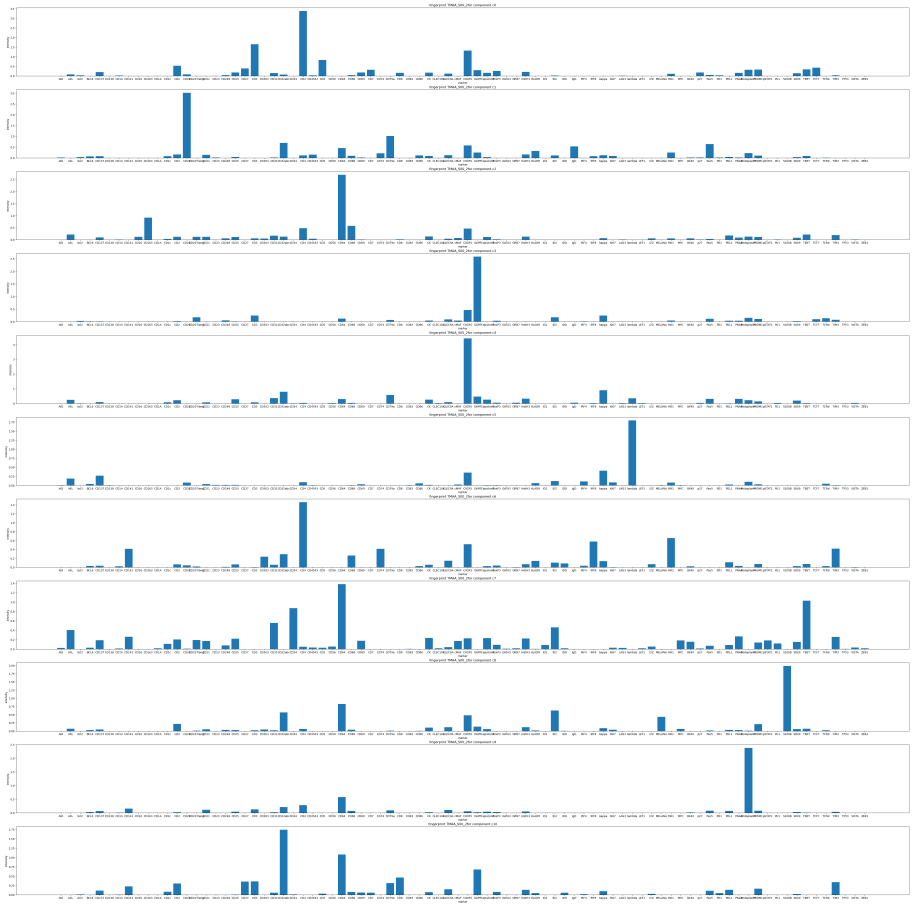


Figure 5.7: Fingerprints corresponding to the 11 components of melanoma sample *TMAA_S00_2*, see also Figure 5.6.

5.4.3 Computational benefits of the pipeline

To cope with the high dimensionality of the dataset, the original approach in [22] started from a subset of the full data. Both a subset of markers (the first tier markers) and cells (20K from the 1.5M) was used to perform an ensemble of three clustering methods: PhenoGraph, FlowSom, and KMeans. The subsampling of the cells was either random sampling, which could miss out on rare cell types, or stratified sampling, which is more likely to include noise. In their approach, the cells are clustered and for each cluster the average

fingerprint is taken for further expert annotation. After the annotations, clusters are merged. Only if more than two methods agree on the final annotation of a merged cluster, the most common annotation is kept. Otherwise, the final cluster is labelled as noise. Finally, to extrapolate labels to the full data set, the Euclidean, Pearson and Manhattan distances were calculated between the remaining cells and the representative fingerprints for each phenotype. The shortest distance was used to label the unused cells.

While effective in the analysis and insightful for the pathological hypotheses in the paper. The original method has some drawbacks, which the NMF-MLP pipeline can overcome:

- The NMF method is scalable to large datasets and does not require any subsampling of the data. As long as the hardware allows for the data to be easily loaded into memory, it can cope very well.
- Also its runtimes are impressive, for example, the in-depth analysis of the *TMAA_S00_2* melanoma tissue took less than 10 seconds to run on a personal computer (2019 MacBook Pro with a 2,6 GHz 6-Core Intel Core i7 processor). The most computational intensive scripts are the visualisation of the results due to the high spatial resolution, for that same sample this script took 20 minutes on the same computer.
- NMF works linearly additive, meaning it tries to construct components with distinct features. This omits the re-clustering step to merge the clusters used for annotation, as done previously.
- As shown in Figure 5.3, feature extraction, phenotype identification and spatial analysis are three steps, which the presented NMF-MLP pipeline tackles all in once. This makes it a valuable tool in real life applications.

5.4.4 Future work

While results are optimistic for MILAN data analysis, more advanced analyses are planned to further improve the results, both for visualisation and identification of the different phenotypes. The general goal of this part of the project is to include the final analysis pipeline in the web tool developed by the MILAN project. This allows other scientists to perform similar analyses on their own data.

Next step: subcellular analysis with NMF

In a first extension, we aim to fully exploit NMF's scalability and the information MILAN data provides: we want to analyze every pixel measured, not averaged per cell. This make it into a subcellular analysis where we hope to distinguish the distinct features within one cell and those similar within multiple cells. This will require external annotations of the clusters by experts as so far, no annotated dataset is available on this subcellular level. The gain in information using all the information MILAN data provides is expected to be substantial. The computations in this case will however require more advanced HPC infrastructure, which at KU Leuven is available through the Flemish Super Computer or VSC.

Next step: improve data quality

As for many omics techniques, data is often noisy and impurities can misdirect all downstream analysis. To this end, a second tier of improvements will be made through tailored normalization of the markers based on the expected decrease in signal between the different rounds of measurement. Additionally, as NMF only approximates the true distribution, we will perform a check on the discarded residual between the true data and the reconstructed data after NMF decomposition. This residual might still contain crucial information amongst the smaller signals and might aid biomarker discovery in complex cancers.

Next step: hierarchical analysis

The rr2-NMF method allows for hierarchical analysis based on the similarities in both factor matrices. Similar as to MSI, some markers can be correlated and some regions of cells can be further decomposed into a higher level of detail. This makes the application of rr2-NMF an interesting and feasible next step to research.

5.5 Conclusion

With multiplexed IHC (mIHC), such as MILAN, we work in a more targeted manner than MSI. We have only a limited number of markers to perform analysis on, but spatially have a much higher resolution to embed. This makes that we can extrapolate knowledge between the two fields of research to learn and improve analysis results. With the knowledge on NMF acquired in the MSI

research, we have now demonstrated how well it performs on MILAN data as well.

In this chapter, we have introduced an NMF-MLP pipeline for the fast decomposition and annotation of single cell MILAN data. We applied the analysis on human melanoma samples, and performed both multi-sample and in-depth analyses. We see promising results and easy and efficient pipeline that improves earlier work computationally.

While the collaboration is still ongoing, we have indicated several next steps to conduct and improve results even more. Here, a big focus lies on the more detailed analysis using subcellular information and hierarchical analysis. We want to dive into the details of the data. Furthermore, we underline that also in the case of mIHC data, denoising and normalization are extremely important steps in the preprocessing as they influence the data used for any subsequent downstream task.

5.6 Acknowledgements

This work has been funded as a C3 - KU Leuven Category 3 project: "Next-generation pathology by MILAN multiplex analysis and advanced trainable image analysis". The resources and services used to perform the multisample analysis were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.

Chapter 6

Matrix and tensor factorization for cell type clustering of single cell Hi-C data

The research presented in this chapter has been done in collaboration with Columbia University and the New York Genome Center under the supervision of Professor Gamze Gürsoy.

This chapter demonstrates the use of matrix and tensor factorization in the context of **genomics** data analysis. More specifically, we focus on distinguishing cell types based on the spatial organization of the chromatin in single cells.

6.1 Introduction

The field of genomics studies the structure, function, evolution and mapping of genomes. Knowing that humans have over 3 billion nucleotides in the 23 pairs of chromosomes, one can imagine the storage of this genetic material inside the small, spherical, nucleus of a cell has to be extremely compact and efficient, as shown in Figure 6.1.

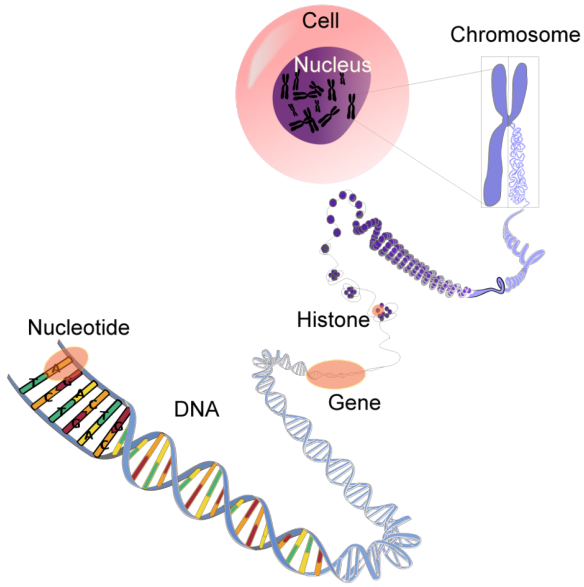


Figure 6.1: **Overview of the organization of DNA.** In each chromosome pair in a cell's nucleus, the DNA consists of nucleotides that form the base pairs. Defined regions of the genome make up a gene. The unused, inactive DNA for that cell is coiled efficiently around histones to save space in the nucleus. Copyright WaSu-Bio, CC0, via Wikimedia Commons

Despite each cell having the same information, how cells, tissues, and organisms switch on and off specific genes for specialization of function depends on how the genome is folded in the nucleus. It is in fact a dynamic, hierarchical, 3D organization of the chromatin, which ensures that the nucleotides that need to find each other, will do so [112]. The single cell high-throughput chromatin conformation capture (scHi-C) method [145] allows us to measure interactions taking place between far away loci of the genome within a cell's nucleus. With this data, we gain important insights in the biological processes of heterogeneous

cell populations such as: cell cycle, cell differentiation and disease progression. Where contacts occur, interactions between regulatory elements and genes can modulate transcription and thus regulate gene expressions. This is crucial information for the study of epigenetic modifications in diseases such as cancers or neurological disorders [184, 164].

The long-range interactions of the genome consist of two main mechanisms: structural looping and enhancer-promoter interactions. A brief overview is given below, but more information on the architecture of the 3D genome can be found in the work of Rowley and Corces (2018) [162].

- **Enhancer-promoter (E-P) interactions**
  Promoters are DNA sequences located near the start of a gene, it allows for RNA to bind to this site and begin transcription. When contact takes place, the enhancer - possibly far away from the gene it regulates - boosts the transcription of that gene. E-P interactions regulate gene expression directly.
- **Structural loops**
  Structural loops fold the genome into domains and can either prevent or enable E-P interactions. They thus indirectly regulate gene expression. With proteins such as CTCF and cohesin, distant genomic regions can be brought together. Interactions between loci occur more frequently within these regions than outside them, which is why they are called topological associating domains or TADs.

The scHi-C assay

Compared to earlier assays, such as: Fish [19], 3C [48], 4C [223, 169], and 5C [59], Hi-C [120] offers an *untargeted* analysis of the interactions. Contacts between all regions of the genome are measured. For scHi-C, this is done within each cell, rather than for a multiple of cells in bulk Hi-C.

In summary, to perform scHi-C according to the protocol [145], the following steps need to be executed:

1. Crosslinking: isolation and fixation of the cells using formaldehyde, preserving the spatial organization of the chromatin.
2. Lysis: extraction of the crosslinked chromatin from the cell structure.
3. Restriction digestion: using a restriction enzyme, the DNA is fragmented (digested) at specific nucleotide sequences.

4. Biotin fill-in (labelling): using biotin, the ends of each DNA fragment are labelled.
5. Ligation: connecting the fragments that were located in proximity (= the interactions).
6. Reverse crosslinking: releasing the linked fragments and pull down the biotinylated ligation junctions (= the interactions).
7. Second restriction digestion: fragmentation of the DNA.
8. Ligation for identification: unique 3-bp (base pair) identification tags are added to each fragment of DNA
9. PCR (Polymerase Chain Reaction) sequencing: using PCR, the fragments are amplified and sequenced.
10. Alignment: the resulting sequences are aligned to a reference genome to find which regions were interacting.

Figure 6.2(a) visualizes some essential steps of the protocol. The method returns a heatmap (or contact map) containing the number of interactions between each region and across all chromosome, this is shown in Figure 6.2(b). Using these heatmaps, we aim to differentiate the different cell types and find the contacts of interest. To do so, customised algorithms are required to identify the discriminating features in the data: we are less interested in the expected interactions on the diagonal, i.e. self contacts and contacts with neighbouring regions are redundant information, but very interested in the long-range ones, i.e. looping and E-P interactions. Understanding the organization of the chromatin is an active area of research with still many unanswered questions.

Obstacles analysing scHi-C

It is important to note, however, that there is a trade-off in using scHi-C compared to bulk Hi-C. Because of the single cell measurements, the number of contacts between regions is expected to be lower and decreases the signal-to-noise ratio (SNR). Furthermore, because of the high number of nucleotides in our DNA, even with bin sizes of 10Mb the resulting single cell contact map is very sparse. Additionally, a genomic distance bias - we have many more contacts close by than far away - should be taken into account when performing data analysis [224]. Lastly, batch effects from preparation and sequencing could influence comparability across the cells. At the cost of increased cell specificity, scHi-C requires more advanced algorithms and preprocessing steps for correct interpretation of the data.

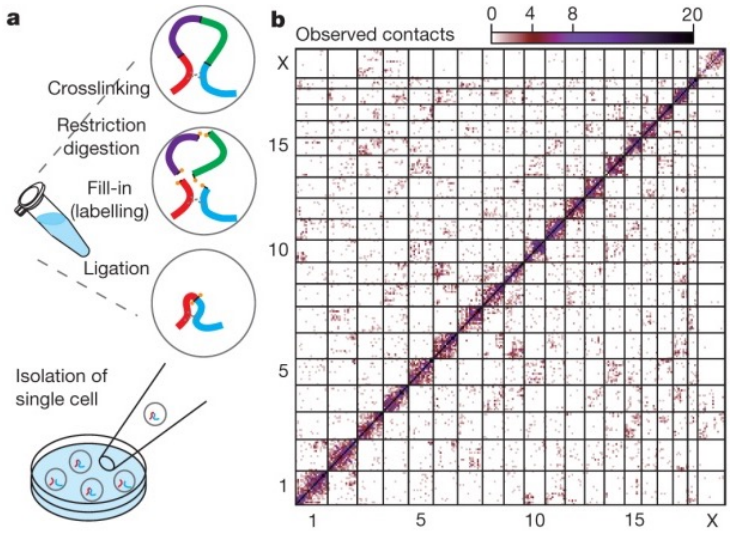


Figure 6.2: **Protocol and data format scHi-C.** (a) Protocol of scHi-C: a single cell is isolated, next the genetic material undergoes: crosslinking, restriction digestion, labelling, and ligation. (b) scHi-C heatmap showing interactions between the different 10Mb regions of a mouse genome, with single-copy X chromosome (chromosome Y and mitochondrial genome were ignored). The figure is reprinted with permission from [145]. Copyright 2013, Springer Nature Limited.

Towards multimodal multi-omic analysis and the relation with gene expression data

A natural extension of this technology is multimodal single cell omics assays that can characterize multiple omic modalities of a single cell. This enables deeper profiling of single cells and allows for the investigation of links between chromatin organization, epigenome and transcription. Indeed, these links lie at the heart of one of the most essential questions in molecular biology, the transcriptional regulations of genes. These multi-omic approaches suffer from low coverage, low SNR and are limited to a few combinations of modalities [199].

An alternative approach to generate multi-omics data is using computational methods to integrate two unimodal sc-omics data into a pseudo-multi-omic data set. Recently, there has been ample algorithm development to combine scRNA-seq (gene expression) and scATAC-seq (chromatin accessibility) modalities into a pseudo-multi-omic scRNA+ATAC modality [199]. However, integration of

single cell Hi-C (chromatin architecture) data with gene expression is still an open problem. The limitation is that we still do not have a robust computational approach that can assign cell types to single-cells using single cell Hi-C data. This is because the relationship between chromatin architecture and gene expression is cryptic, requiring sophisticated data analysis methods.

Gene expression profiles of single cells (captured with scRNA-seq) reveal which genes are actively transcribed or silent. The spatial proximity captured by scHi-C provides insights into how enhancers spatially organize with respect to their target promoters, influencing gene transcriptional regulation. Integrating scHi-C with scRNA-seq datasets aims to elucidate how the spatial proximity of regulatory elements to target genes influences gene activity. This integration requires aligning and correlating spatial chromatin interactions with gene expression profiles across thousands of individual cells.

When aligned, we can pinpoint which genes are influenced by specific chromatin architectures within individual cells. By correlating spatial chromatin interactions captured by scHi-C with gene expression patterns from scRNA-seq, comprehensive insights into how genome organization impacts transcriptional dynamics at a single-cell resolution can be achieved. This approach not only enhances our understanding of gene regulatory networks but also sheds light on the molecular mechanisms underlying cellular diversity, development, and disease.

As methodologies advance and computational tools improve, the synergy between scHi-C and scRNA-seq datasets holds significant potential for unraveling the intricate relationships between chromatin architecture and gene regulation at unprecedented resolution.

6.2 Single cell Hi-C genomics data of developing mouse brain

The data used is an open-source transcriptome and 3D genome atlas of developing mouse brain created by Tan et al. [182] (GEO: GSE162511). The data set consists of scHi-C measurements of diploid cells, obtained with their Dip-C assay [184], as well as MALBAC-DT transcriptomics measurements. The full data consists of 3,517 3D genome and 1,954 transcriptome cells of developing mouse cortex and hippocampus. The major cell types identified are: neurons, astrocytes, oligodendrocytes, and other cells, such as microglia and vascular cells. Figure 6.3 shows a schematic overview of the atlas development, as well

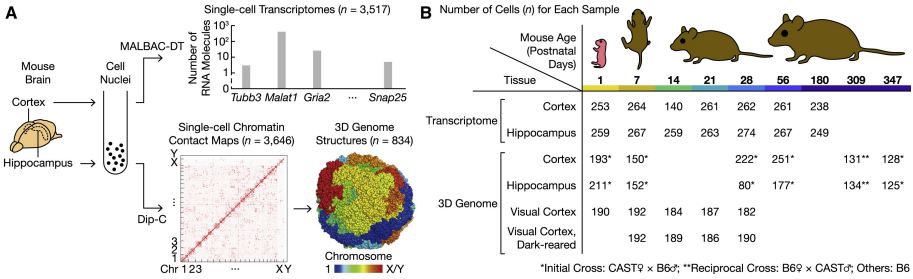


Figure 6.3: **Dip-C data set of developing mouse brain.** (a) Schematic overview of the atlas development. From developing mouse brain cortex and hippocampus cells, single cell transcriptomes and contact maps have been created, using respectively MALBAC-DT and Dip-C assays. (b) Overview of the data set in numbers. Figure reprinted with permission from [182]. Copyright 2020 Elsevier Inc.

as a tabular overview of the number of cells measured at nine distinct ages of the developing mouse brain.

One key advantage of using a Dip-C data set is the true cell type annotation they provide. This is done by isolating the cells into individual wells and adding barcoded tags (a unique synthetic DNA sequence), which are then analyzed using scRNA-seq [183]. Using these labels is a big advantage when evaluating the results obtained using solely the contact maps. When using a primary analysis to infer the cell-types from the contact maps, noise or discrepancies in the measurements could affect the annotations and be self-confirmed by secondary analysis of the contact maps. Another advantage of using this data is comparable research. The data set is public and consists of annotations for multiple modalities, cell-types, and cell ages, and has been re-analyzed several times [218, 224, 124]. This makes it an excellent tool for comparison of computational approaches for scHi-C data analysis and multimodal integration with scRNA-seq.

6.3 scHi-C data analysis using matrix and tensor factorization

Can we cluster scHi-C contact maps according to their cell types?

Next, can we find the driving forces, i.e. contacts, that differentiate these clusters?

Matrix and tensor factorization methods have been used extensively in many research domains and have proven effective for the dimensionality reduction and clustering of all sorts of data. They are scalable for high dimensional data sets, can incorporate constraints such as sparsity, and rely on explainable mathematical principles [41, 80]. This makes them an ideal starting point for our problem at hand. One benefit of using tensor over matrix factorization lies in its uniqueness properties, but the choice between the two algebras is problem dependent [168]. Additionally, we know scHi-C data is non-negative and expected to have Poisson noise, due to the counting nature of the data acquisition. These priors are also taken into account for analysis. In this section, we demonstrate the performance of a few well-suited factorization algorithms as well as the importance of the initial preprocessing.

The raw data

As a starting point, a subset of the data is used to build each pipeline and increase experimental set-up, runtimes and evaluation. Once results are satisfactory, more data is incorporated in the analysis. We limit ourselves therefore to the gene-rich chromosome 19 contact map for the cohort of 150 cortex cells with 7 days of age. We rely on the `hic-straw` GitHub repository to load the `.hic` files into Python [62].

Chromosome 19 consists of 61,431,566 base pairs, this will amount to a $(60, 60)$ matrix for each cell when binned at 1Mb resolution, see Figure 6.4(a) for an example contact map. To perform our analysis across the different cells, we can either create a matrix representation, see Figure 6.4(b), by vectorization of the bin-bin plane and stacking these vectors. Or, we could stack the original 2D contact maps into a third order tensor, as show in Figure 6.4(c).

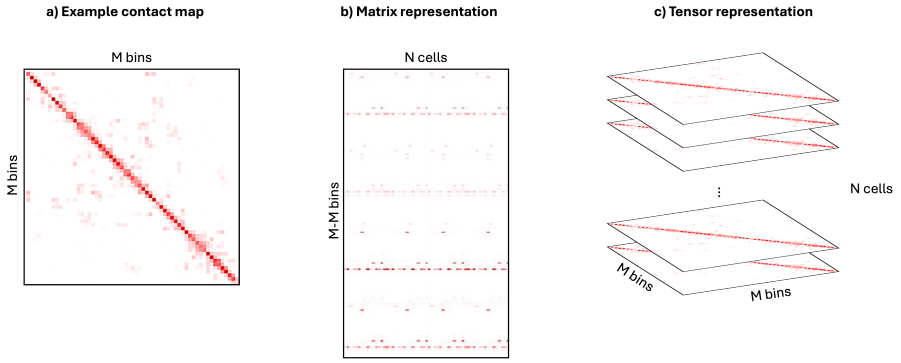


Figure 6.4: **Matrix and tensor representation for scHi-C data.** (a) Example of a scHi-C contact map for one chromosome-chromosome interaction pair. (b) Vectorization of the bin-bin space allows for stacking across multiple cells in a matrix format. (c) Tensor creation by stacking the contact maps for each cell on top of each other in a third, new dimension

6.3.1 Preprocessing

As with any omics technology, noise is highly present in the measurements and can mislead analysis. To conduct our experiments rigorously, a first step is the correction of the data regarding noise and artefacts. Using scHi-C data, as mentioned before, we notice high sparsity and low SNR ratios, which result in difficulties extracting the true signal. Correcting scHi-C towards noise and sparsity is a complex and extensive area of research. Here, we base ourselves on recent literature and in-house developed tools to assure the best representation of the data. Due to the lack of a golden standard, the performed preprocessing is subject to trade-offs and scientific preferences. We present the following preprocessing steps, each with an on/off functionality to test their influence on the final matrix and tensor factorization methods.

BandNorm normalization

Many approaches for scHi-C normalization exist, see [130]. It is important to normalize the data within a contact map as well as across all the contact maps used in the analysis. To this end, we apply BandNorm [224] normalization.

BandNorm first removes the genomic distance bias within each contact map by division of each interaction by the mean of that band, next, it normalizes for the sequencing depth between different cells (contact maps), and finally, a

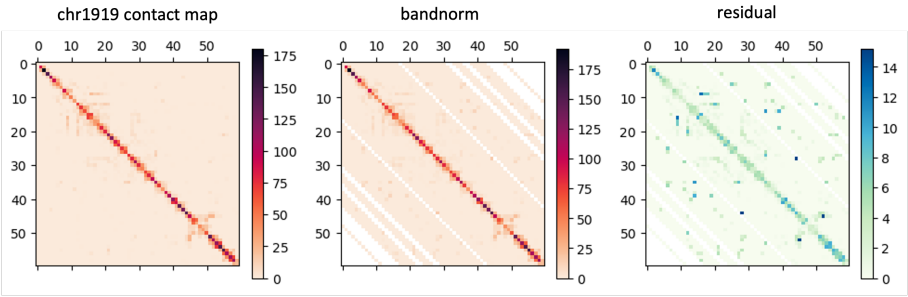


Figure 6.5: **BandNorm result on one of the chromosome 19-19 contact maps used for our analysis.** Shown left to right: the original contact map, the contact map after BandNorm normalization, and the residual of the two.

common band-dependent contact decay estimate across the different contact maps is added back. We write for contact value C for a bin pair r of cell c at band v:

$$C_r^{cv} = \frac{Y_r^{cv}}{L^{cv}/m_v} \frac{\alpha(v)}{m_v} = \frac{Y_r^{cv}}{L^{cv}} \alpha(v), \quad \forall r \in \mathcal{A}(v), \tag{6.1}$$

where m_v represents the number of bin pairs in band v, $\mathcal{A}(v)$ denotes the bin pairs at band v, Y^{cv} describes the vector of interaction frequencies of bin pairs at band v for cell c, $L^{cv} = \sum_r^{m_v} Y_r^{cv}$ gives the total interaction frequency of cell c at band v. Finally, $\alpha(v)$ is the average interaction frequency of band v across all cells, it is calculated as $\sum_{c=1}^{N} L^{cv}/N$.

Figure 6.5 shows an example of a contact map before and after BandNorm normalization. The empty bands are the result of division by a mean of zero contacts in that band, which are therefore put to zero.

Band correction towards the off-diagonals

Another important concept in scHi-C data analysis is the genomic distance bias. The interaction frequency decreases as the genomic distance increases. It is expected that loci in close proximity have more interactions than those far away. This theory is confirmed by polymer physics models, which state that the decay of interactions follows a power law. The following decay function is observed [120] for bulk Hi-C data:

$$f_{contacts_bulk} = dist(x, y)^{-1.08}. \tag{6.2}$$

However, looking at the single cell data an ever stronger decay is noticed, probably caused by the increased sparsity. Fitting the observed function, see

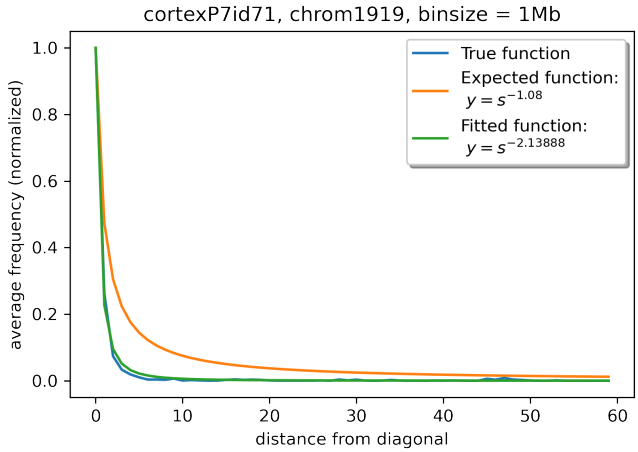


Figure 6.6: Power law fitting of the interaction frequency compared to the genomic distance of a scHi-C data set compared to bulk Hi-C decay. X-axis shows the distance of the bins from the diagonal (i.e. a self contact). Y-axis shows the average frequency of the contacts occurring (1 for a self contact). Blue shows the true data, orange the expected decay for bulk Hi-C and green shows a fitted power law for the single cell data of the chromosome 19 contact map, exhibiting a faster decay than for bulk.

Figure 6.6, we get the following power law for scHi-C:

$$f_{contacts_sc} = dist(x, y)^{-2.14}. \tag{6.3}$$

Data imputation using polymer chain models

Addressing the sparsity issue in the contact maps, a physical polymer chain model is used to fill in missing, but expected contacts. We apply an in-house model based on constrained optimization of a polymer chain generation process using importance sampling [86].

In the model, first, a set of spatial distance constraints is deduced from the raw scHi-C data. Then, a set of conformations (or chains) is generated that satisfy the constraints while minimizing an error function ϵ measuring the deviations from the distance constraints. The set of conformations is expected to follow a target distribution: $\exp(-\epsilon)$. As this target distribution of all distances is complex to sample from, a geometrical importance sampling approach is applied

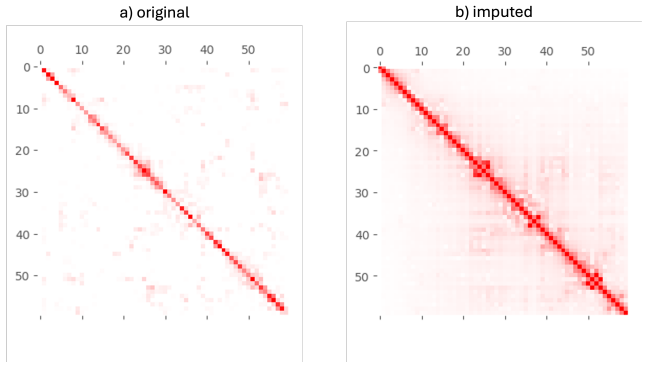


Figure 6.7: Polymer chain model imputation of a chromosome 19 scHi-C contact map.

to select a subset of the constraints. To accommodate for biases using this approach, we weigh each final chain accordingly.

To create a proposed chain $\boldsymbol{x}$, a growth (or trial) function is followed. This function determines the 'best' next node t in chain l based on all nodes before and the physical properties, i.e. collision and self-avoidance, of chromatin chains. Mathematically, the objective is to minimize the following priority score $\beta_t^{(l)}$:

$$\beta_t^{(l)} = exp\left[-\frac{\rho_1 f_1(x_t^{(l)}) + \rho_2 f_2(x_t^{(l)})}{\tau_t}\right], \tag{6.4}$$

with f_1 a function denoting the self-avoiding property of chromatin fiber, f_2 a function incorporating the derived distance constraints for this node, ρ_1 and ρ_2 the coefficients for each function, and τ a temperature parameter.

The trial distribution consists of the probability of all possible spatial positions for the next node t, given the constraints. With the final distribution, distances in the original contact map are updated and the process is repeated for the new target distribution. Iteration will stop once all constraints are met. Based on the resulting chains for each cell, contacts are defined depending on the resolution of the data. An example of an imputed contact map is given in Figure 6.7.

Using a polymer chain model, the physical rules of chromatin conformation are used to enhance the scHi-C contact map for missing contacts. This includes more physical knowledge in the imputation than statistical models for smoothing such as Gaussian smoothing [228] or the work by Xie et al. [212].

6.3.2 Non-negative matrix factorization

To understand scHi-C data better, and investigate the effects of the preprocessing steps mentioned above, we inspect different NMF decompositions to uncover the different cell-types in the data base on the chromatin contacts. Figure 6.8 shows a basic schematic of NMF.

For the scHi-C data mentioned in the previous section, we define the matrix $\mathbf{X} \in \mathbb{R}_+^{m \times n}$ where each contact map is represented as a column by vectorizing the bin-bin dimension. Here, the m rows describe each bin-bin combination and the n columns the different cells. The resulting factor matrices are $\mathbf{W} \in \mathbb{R}_+^{m \times r}$, of which each column represents a basis element of the contact patterns, and $\mathbf{H} \in \mathbb{R}_+^{r \times n}$, describing the contribution of each basis element to each cell. For a good representation of the data, we aim to get a different contact pattern for each cell-type, thus taking the rank r equal to the number of different cells. For the selection of the data used, i.e. day P7 for the cortex cells, we have $r = 7$ corresponding to the number of different cell-types. We confirm this by looking at the SVD decomposition of the matrix and plotting the singular values in a semi-log graph, shown in Figure 6.9. We see that around the 7th singular value the *elbow* in the curve appears, making it a good estimate for the number of main trends.

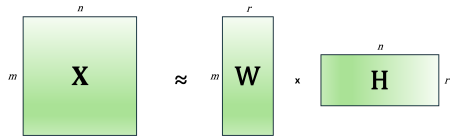


Figure 6.8: Schematic representation of non-negative matrix factorization.

The resulting factor matrices can be read-out to understand for each component how the contact basis pattern contributes to the different cells. An example for one component is shown in Figure 6.10.

This section investigates the performances for standard NMF, KL-NMF, and LDA. For a fast inspection of each decomposition, we visualize the cells based on their weights in the $\mathbf{H}$ matrix. We apply UMAP to 2 dimensions and color-code the embeddings using the true label, i.e. cell-type, for each cell. Once this embedding shows a good clustering, we can inspect all components as done in Figure 6.10 to detect the basis contact patterns and see which contacts are specific for certain clusters. We start from a 1Mb binned representation of chromosome 19 and compare to the 500kb and 100kb resolutions, we do this for each of the different preprocessing steps.

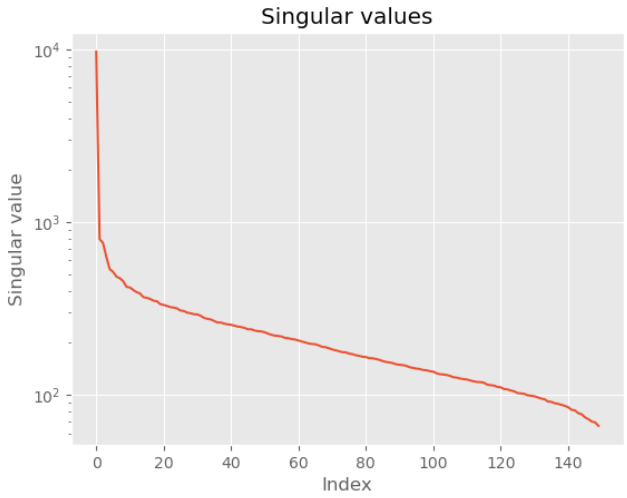
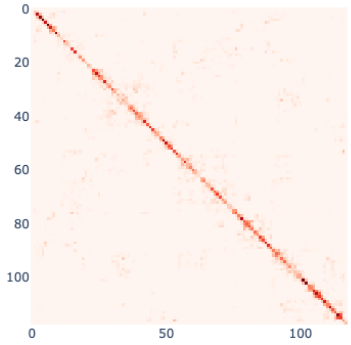


Figure 6.9: Semi-log plot of the singular values of the chromosome 19 matrix representation (day=P7, tissue=cortex, 150 cells) with 1Mb resolution.



(a)

cell id	cell type	contribution to component
179	Oligodendrocyte Progenitor	33.26
002	Neonatal Neuron 2	8.0
101	Neonatal Astrocyte	5.77
051	Oligodendrocyte Progenitor	4.17
065	Neonatal Neuron 2	3.7
068	Neonatal Astrocyte	3.49
045	Neonatal Astrocyte	3.23
143	Neonatal Astrocyte	3.21
149	Oligodendrocyte Progenitor	2.13
055	Neonatal Astrocyte	1.54
067	Unknown	1.41
135	Pyramidal Cell	0.48
044	Neonatal Neuron 2	0.0
158	Neonatal Astrocyte	0.0
028	Neonatal Neuron 2	0.0
043	Neonatal Astrocyte	0.0
115	Neonatal Neuron 2	0.0
188	Neonatal Astrocyte	0.0
026	Neonatal Neuron 1	0.0
122	Neonatal Neuron 2	0.0

(b)

Figure 6.10: **An example of a single NMF component.** (a) Basis contact pattern for a single NMF component at 500kb resolution. (b) The top 12 most contributing cells to this basis: cell id, cell-type and contribution weight.

Appendix E contains the UMAP embeddings for each experiment. For easy comparison of the embeddings, we have calculated the adjusted Rand index (ARI) between each low rank embedding of the cells according to the decomposition and the true cell labels. The ARI is defined as:

$$\text{ARI} = \frac{\text{RI} - \text{Expected_RI}}{\max(\text{RI} - \text{Expected_RI})}, \tag{6.5}$$

with the Rand index describing the percentage of *correctly* clustered elements: $\frac{TP+TN}{TP+FP+TN+FN}$, with TP the true positives, TN true negatives, FP false positives, and FN the false negative cell clusterings [100]. To compare with literature [226], we used spectral clustering to determine the predicted labels. The ARI score ranges from -1.0 to 1.0, with 1.0 a perfect match and 0.0 random labels, using Python's scikitlearn environment [152] this is bounded at -0.5 for very discordant clusterings.

Here, we give an overview of the obtained ARI scores for the various setups: the different resolutions and preprocessing techniques. The next section discusses the implications of these observations in more detail.

Standard non-negative matrix factorization

Standard NMF, as discussed in more details in Chapter 2, implies a Gaussian distribution of the data. Although scHi-C data is expected to follow a Poisson/binomial distribution [224], NMF should still be able to detect the main trends. Figures 6.11 and 6.12 show the performance a rank-7 NMF decomposition for the different bin resolutions and preprocessing approaches.

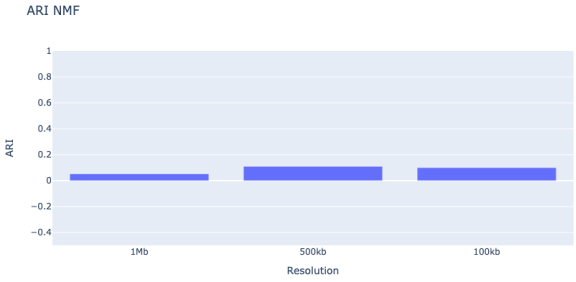


Figure 6.11: The ARI scores of the NMF result for different resolutions: 1Mb, 500kb and 100kb.

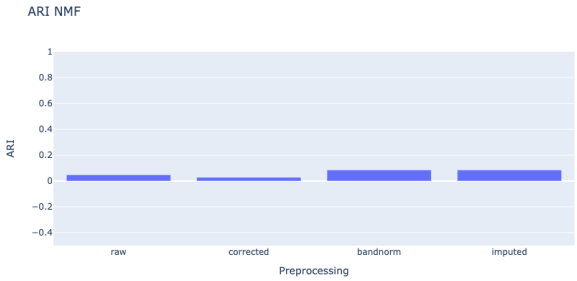


Figure 6.12: The ARI scores of the NMF result for different preprocessing techniques: no preprocessing, corrected for the decay function, BandNorm normalization and imputed using a polymer chain model.

Kullback-Leibler NMF

Here, we try to improve results using Kullback-Leibler NMF (KL-NMF). As shown in earlier work on MSI data analysis [148], KL-NMF is able to capture the Poisson distributed nature of the data. Figures 6.13 and 6.14 show the performance a rank-7 KL-NMF decomposition for the different bin resolutions and preprocessing approaches.

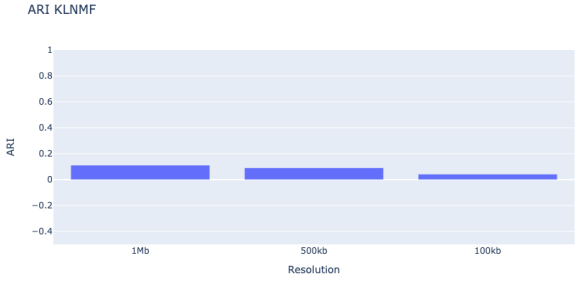


Figure 6.13: The ARI scores of the KL-NMF result for different resolutions: 1Mb, 500kb and 100kb.

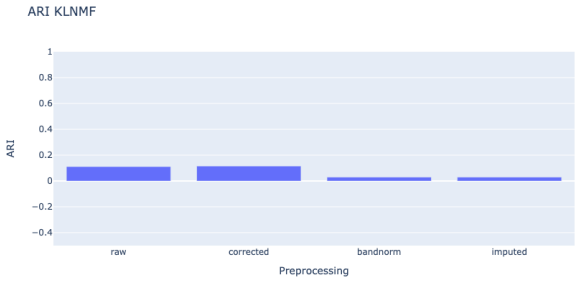


Figure 6.14: The ARI scores of the KL-NMF result for different preprocessing techniques: no preprocessing, corrected for the decay function, BandNorm normalization and imputed using a polymer chain model.

Latent Dirichlet Allocation

Finally, we applied Latent Dirichlet Allocation (LDA) for the clustering of the cells in matrix representation. As shown in the work by [106], LDA aims to uncover the hidden trends or topics by means of a generative probabilistic model, incorporating all *possible* interactions, making it less overfitting than NMF. We note that compared to NMF and KL-NMF, the runtime for LDA is one magnitude bigger. Figures 6.15 and 6.16 show the performance a rank-7 LDA decomposition for the different bin resolutions and preprocessing approaches.

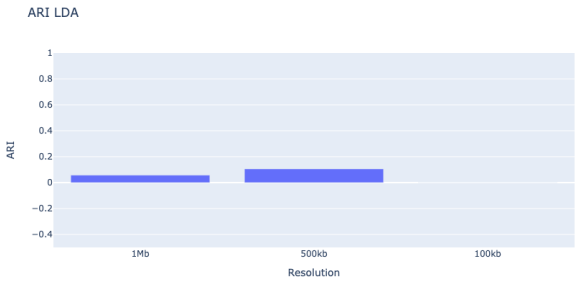


Figure 6.15: The ARI scores of the LDA result for different resolutions: 1Mb, 500kb and 100kb.

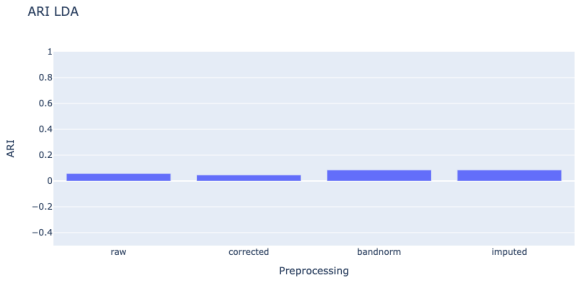


Figure 6.16: The ARI scores of the LDA result for different preprocessing techniques: no preprocessing, corrected for the decay function, BandNorm normalization and imputed using a polymer chain model.

6.3.3 Non-negative tensor factorization

To exploit the symmetrical structure of the contact maps, a higher order factorization might capture more information. As shown in Figure 6.4(c), the tensor representation can be thought of as a stacking of contact maps along a new dimension. The resulting non-negative tensor factorization is inspected, both for Tucker and CPD decompositions.

Non-negative Tucker factorization

Based on the notation of Kolda and Bader in [108], we define the tensor $\boldsymbol{\mathcal{T}} \in \mathbb{R}_{+}^{M \times M \times N}$ representing the stacked contact maps, in our case with $M = 60$ and $N = 150$. A non-negative Tucker factorization aims to decompose this tensor in a compressed core tensor $\boldsymbol{\mathcal{G}} \in \mathbb{R}_{+}^{P \times Q \times R}$ and connected factor matrices $\boldsymbol{A}, \boldsymbol{B},$ and $\boldsymbol{C}$. This can be seen as a higher order singular value decomposition (HOSVD) [114]: the factor matrices describe the basis vectors along each dimension and the core tensor shows how they are in mutual interaction. Mathematically, we write:

$$\boldsymbol{\mathcal{T}} \approx \boldsymbol{\mathcal{G}} \times_1 \boldsymbol{A} \times_2 \boldsymbol{B} \times_3 \boldsymbol{C} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} \boldsymbol{a}_p \circ \boldsymbol{b}_q \circ \boldsymbol{c}_r, \tag{6.6}$$

with $\circ$ the vector outer product. We note P, Q, and R as the number of components or columns in each of the factor matrices $\boldsymbol{A}, \boldsymbol{B}$, and $\boldsymbol{C}$, respectively, as shown in Figure 6.17.

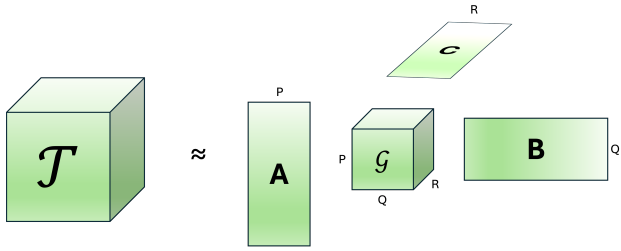


Figure 6.17: Schematic representation of a Tucker decomposition for a three dimensional tensor.

Using Tucker we aim to find a good clustering of the cells and use the core tensor to find the most important contacts driving these clusters. Figures 6.18 and 6.19 show the performance a rank-[7,7,7] Tucker decomposition for the different bin resolutions and preprocessing approaches.

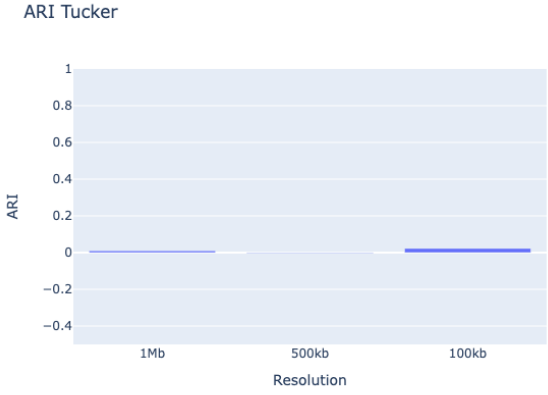


Figure 6.18: The ARI scores of the Tucker result for different resolutions: 1Mb, 500kb and 100kb.

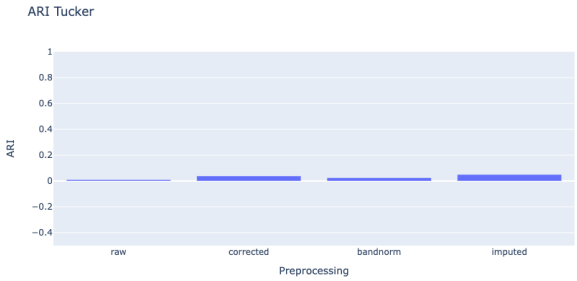


Figure 6.19: The ARI scores of the Tucker result for different preprocessing techniques: no preprocessing, corrected for the decay function, BandNorm normalization and imputed using a polymer chain model.

Canonical polyadic decomposition

Where a Tucker factorization models the interactions between all components of the different factor matrices, a CPD or canonical polyadic decomposition aims to find one-to-one relationships for each cluster. This means every cluster is described by a rank-1 tensor, i.e. one vector from each factor matrix. We now have only one value for the rank equal to the number of clusters. Figure 6.20 shows the schematic representation of CPD for a third order tensor, described by the following equation:

$$\mathcal{T} \approx \sum_{r=1}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r, \tag{6.7}$$

with $\circ$ the outer product of the vectors [108]. Figures 6.21 and 6.22 show the performance a rank-7 CPD decomposition for the different bin resolutions and preprocessing approaches.

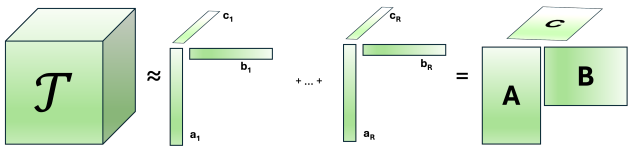


Figure 6.20: Schematic representation of a CPD decomposition for a three dimensional tensor.

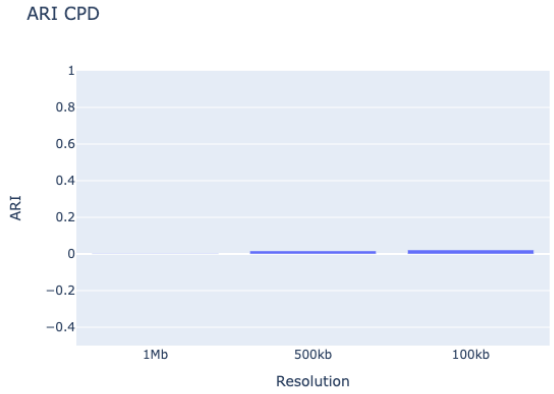


Figure 6.21: The ARI scores of the CPD result for different resolutions: 1Mb, 500kb and 100kb.

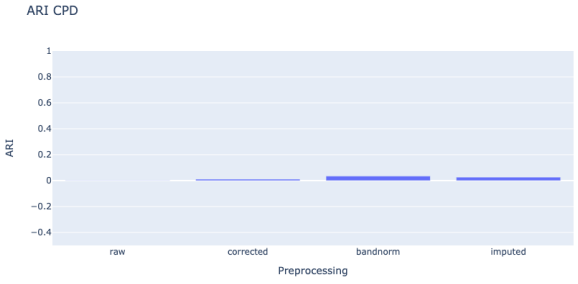


Figure 6.22: The ARI scores of the CPD result for different preprocessing techniques: no preprocessing, corrected for the decay function, BandNorm normalization and imputed using a polymer chain model.

6.4 Discussion

This section discusses the obtained results from the previous section and gives several directions for future work to improve the current clusterings.

6.4.1 Performance of the decompositions

The ARI scores obtained in the previous analyses do not get higher than a 0.2 score. Unfortunately, this means none of the techniques, resolutions or preprocessing techniques gives us enough information towards cell-type clustering. This is also demonstrated by the UMAP embeddings in Appendix E.

The ARI scores give indication that a 500kb resolution obtains better results than 1Mb. For 100kb, the highest resolution possible, the clusters are more distinct, but do not represent the cell-types, giving a lower ARI score than 500kb. In terms of preprocessing, even less improvements are observed w.r.t. the raw data. Except for KL-NMF, we see that BandNorm normalization and imputation of the data lead to marginally better ARI scores. Finally, we see that for the small data set used, the matrix factorizations outperform the tensor approaches. This might be caused by the shared dimension in the tensor representation, but could also be dependent of the chosen algorithms for decomposition.

We can conclude that ARI scores are not good enough for further analysis of the contacts. The most informative features of the data are still hidden in its current representation. Here, we underline that negative results are still results and could lead to more advanced approaches for the analysis of scHi-C data. To this end, we propose several next steps to improve the current clustering results with factorization methods.

6.4.2 Future work

Next step: more data

Even though chromosome 19 encodes for many genes, it seems that it might not be sufficient to distinguish the different cell-types solely by looking at the contacts of chromosome 19. A first next step would extend the experiments above for the combined analysis of all chromosome contact maps. Here, the main focus would be the intra-chromosomes contact maps, as little to none information is expected to be stored in the inter-chromosome contacts. This would imply essential contacts taking place between different contact maps, so we assume it is mostly noise in there. Secondly, we would incorporate the cells for all ages of the mice and those from the hippocampus as well as the cortex.

Next step: structured data fusion

To further improve the results other tensor factorization approaches can be researched. To aid the selection of which one to use, a Structured Data Fusion (SDF) approach, as presented in [177], could be applied to the data. SDF is an integrated tool in the TensorLab MATLAB toolbox [203] that tries out multiple solvers in selection of the best one, and it has the option to impose algebraic structures on the factor matrices if beneficial for the solution. A sparse format can be used to load the data for memory efficiency. Additionally, SDF has the option of analysing multiple tensors simultaneously and with different solvers - opening the door to multimodal analysis: scHi-C together with scRNA-seq.

Next step: uncovering the hidden trends

Throughout the experiments, and currently available literature, it seems clear that the most informative signals are hidden amongst the noise, preventing standard data analysis techniques (both linear and non-linear) from capturing them. As for other omics modalities, modelling noise and normalization of the data remains extremely challenging. A proposition is to go back to square one and find out where the signals are hidden to further develop appropriate data normalization and analysis. Here, a first approach will be to analyze the contact maps band per band. Maybe there are optimal distances for chromatin to loop within. Alternatively, a supervised, deep learning approach could be utilized to correctly cluster the different data sets, after which a feature importance analysis, such as SHAP-values, could be applied to derive where the discriminating contacts are hidden.

Next step: multimodal analysis

Once the clusterings are able to capture the different cell-types present, we can move on to the next step in the analysis. Using factorization techniques, we obtain not only the clusters of cells, but also the accompanied contacts describing each cluster. This information could be indicative of gene locations in our DNA. Using the accompanied transcripts (scRNA-seq data), we can connect each cell-type with its defining contacts and gene expression information. As mentioned above, coupled matrix and tensor factorization techniques can be extended to multimodal data analysis through shared factors, making them excellent candidates for further research.

6.5 Conclusion

Tensor and matrix factorizations show promising results towards scHi-C data analysis. They require little parameter tuning and are very scalable towards big data. Where we have not yet been able to crack the code for scHi-C data analysis by using solely the measured data, we performed extensive experimentation to understand the data and how factorization methods can cope with it. Different preprocessing approaches have been employed and first conclusions confirm indeed how the noise is suppressing the true signal, especially on the far diagonal.

With the groundwork done, experimentation will continue to the next level. Can structured data fusion leverage the current approaches by identification of possible algebraic structures on the factor matrices? Can a more advanced approach towards the preprocessing aid the representation of the data in such a way matrix factorizations can provide better ARI scores? Can we provide a tailored algorithm, designed to the complexity of scHi-C data, where we aim to automatically detect the most informative signals in the data?

At this point, the answer to these questions is unknown. However, the research in this chapter is still ongoing. As it remains our goal to discover chromatin regions that are essential for specific cell types and continue even further towards multi-omics analysis incorporating the scRNA-seq data. If we are able to solve this complex puzzle computationally, implications in the field will be profound. First of all, the cost of performing genomics experiments is of a different magnitude than computational ones. Money will be saved. Secondly, if the model works for scHi-C, it could be expanded to bulk Hi-C, pore-C, or other variants. And finally, we could draw new biological conclusions regarding the dynamical organization of chromatin towards the expression of cell-type specific proteins.

6.6 Acknowledgements

This work was supported by Fonds Wetenschappelijk Onderzoek Travel Grant V417023N. 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). The resources and services used in this work were provided by the New York Genome Center.

Carlos Angel is thanked for his help regarding the preprocessing of the data and additional insights in the underlying biology, also Anoushka Joglekar for the enriching collaboration on the topic and her feedback on this chapter.

Conclusion

Chapter 7

Conclusion and future perspectives

The devil is in the details.

In this thesis, we have investigated the use of non-negative matrix factorization (NMF) for the analysis of multiple omics data sets. We have demonstrated how NMF can aid in the unsupervised, but explainable analysis of three different omics modalities: mass spectrometry imaging (MSI), immunohistochemistry (IHC) and chromatin conformation (Hi-C). Here, we motivate the use of NMF for fast dimensionality reduction, visualization and clustering of omics data. And finally, we developed an extension to NMF for the tailored analysis of MSI data, able to cope with smaller and overlapping signals. This chapter discusses the contributions in more detail and provides an outlook on the future of omics data analysis.

7.1 Conclusion

To conclude, we highlight the main insights and implications of each chapter, both for the mass spectrometry imaging algorithms as for the extensions to other omics modalities.

7.1.1 Algorithms for the analysis of MSI data

Mass spectrometry imaging (MSI) allows for the untargeted measurement of thousands of molecules in a slice of tissue simultaneously. Over the past few decades, it has become a well-established method within the spatial omics community, offering an unparalleled richness in information for the *in situ* analysis of biomolecules. Nevertheless, as there is still room for improvement, MSI and MSI data analysis are still active areas of research.

The work presented in the initial section of this dissertation has made significant contributions to the MSI community in three key areas. Firstly, we offered fellow researchers a comprehensive review on the performance of NMF and related methods for MSI data analysis. Secondly, we introduced the application of TDA to enhance MSI data visualization. Lastly, we proposed an extension to NMF, tailored specifically for the detection and deconvolution of lowly abundant and overlapping MSI signals, respectively.

An in-depth comparison of NMF and related methods

Chapter 2 of this thesis starts with a detailed overview of NMF, KL-NMF, PLSA and LDA for the analysis of MSI data. This chapter discusses the theoretical framework for each method, and the most commonly used algorithms for their analysis. The four methods are evaluated on seven MSI samples, three pancreas and four lymphoma. By evaluating the results numerically and visually, we conclude that KL-NMF captures the Poisson noise of MSI data better than regular NMF, and that PLSA and KL-NMF differ in the practical setting, even though their objective functions are the same. LDA, while outperforming PLSA in the setting of text-mining, is less effective in MSI data analysis. With this in-depth comparison of the four methods, we offer fellow researchers in the field a concise summary of each method. While we cannot define a golden standard, one should be aware of the mathematical implications of each method when drawing conclusions from the results.

Visualization of MSI data using TDA

The next chapter, Chapter 3, describes a non-linear approach for the visualization of MSI data. Using Mapper, we color-code MSI data of a mouse pancreas based on the topological similarities between the different data points. In essence, a high dimensional point cloud is mapped onto a lower dimensional graph of connected components. The nodes in the graph represent clusters of data points, while the edges define the data points that are shared between the clusters. Mapper requires the selection of multiple parameters: the filter function, the n-D low dimensional space, the number of intervals, the percentage of overlap and the cluster method. The chapter describes an extensive comparison of different parameter combinations for optimal visualization results. Due to the lack of a ground truth, the Mapper results where compared to a standard UMAP visualisation by means of the between-cluster / within-cluster dispersion ratio. Additionally, for one of the parameter combinations, we noticed a different representation of the islets of Langerhans. In this case, the islets consisted of two components, describing an outer ring structure and an inner core, instead of just one. As Mapper does not provide molecular information on the obtained components, the cause of this detected ring structure remains unknown and can only be explained with subsequent analyses. This could then indicate if the different structures in the islets are the effect of structural differences or the result of artifacts describing the transition between the endocrine and exocrine region. This missing molecular information is resolved in the next chapter and allows for further speculation on the different clusters in the islets.

Detailed analysis of MSI data using a novel rr2-NMF approach

The work in Chapter 4 presents an extension to NMF specifically designed for the analysis of MSI data. Looking at MSI data in general, we noticed how difficult it is for most data analytical tools to cope with the smaller abundances and nuances between peaks in the m/z spectra. The highly abundant values, containing important information about the tissue under investigation, are easily detected by many algorithms. Amongst the lower abundances, however, peaks are hidden that contain information just as important, but, due to their intrinsic low signal or to the suppression by the high abundances, are easily disregarded as noise. To resolve this problem, we have extended a standard NMF method to a hierarchical variant: recursive rank-2 NMF. By iteratively performing rank-2 NMF on subsets of the data, a more nuanced and detailed analysis can be done. The pancreas samples studied in the previous chapters provide an excellent model to study the effect of suppression, as the pancreatic islets are characterized by their massive production of insulin. When applying

rr2-NMF, we detect again a ring structure in the islets of Langerhans. However, unlike for UMAP or mapper with rr2-NMF, we obtain the spectral distribution for each spatial pattern. Inspecting the m/z spectra for each cluster, we noticed a distinct expression profile between Ins1 and Ins2, each containing their corresponding two charged states as well as the presence of additives on the molecules, and glucagon correlating to the outer periphery cluster. These findings were confirmed in each of the three samples and could imply a difference in function between Ins1 and Ins2, as mentioned in earlier literature, yet not fully understood by experts. The method is able to deconvolute the two overlapping insulin distributions, with their peaks lying only 8 Da apart in an m/z range of 2-20 kDa. Additionally, it shows their relative abundance in the sample, where Ins2 is a third more abundant than Ins1, which was only confirmed recently by expensive mRNA measurements.

7.1.2 Application to other modalities

Demonstrating the versatility of NMF in omics data analysis, two use cases describe how well NMF performs on data obtained with other omics modalities, i.e. multiplexed IHC and chromatin conformation. Experimenting with different modalities allows to test the robustness and limitations of NMF. Additionally, it aids in a better understanding of other omics data sets, how they differ and how they are connected. This is a first step towards the development of multi-omics approaches.

Fast clustering and annotation of multiplexed IHC data

In Chapter 5, we shift gears from MSI towards multiplexed immunohistochemistry data obtained with MILAN. The data represents the measurement of 84 protein markers throughout different samples of melanoma biopsies. While this is much lower than the number of m/z values measured with an MSI experiment, the spatial resolution is much higher. MILAN data measures up to a sub-cellular level. We show that NMF can easily handle this spatial resolution, even for a multi-sample analysis of more than 1 million cells, and thus omitting the need for subsampling. By performing the NMF analysis simultaneously for multiple samples, we provide a fast spatial representation of the corresponding marker fingerprints across the samples. Alternatively, as for MSI, with NMF we can perform an in-depth representation of the main trends in a single sample. Using annotated data sets of similar pathologies, the obtained bases or fingerprints are further classified to the corresponding cell-types. Further analyses, such as subcellular resolution, between-cluster / within-cluster dispersion, and inspection of the NMF residual, allow for more

precision in the model and can aid experts/pathologists in the time-consuming task of manual image inspection. Furthermore, an interactive implementation of hierarchical NMF could resolve the difficulty of rank estimation and can be included in the webtool currently being developed by the MILAN group.

Understanding scHi-C data using matrix and tensor factorization

Finally, in Chapter 6, we move away from spatial proteomics and apply matrix and tensor factorization for the analysis of single cell Hi-C data. Even more sparse and noisy, single cell Hi-C measures the pairwise contacts of the chromatin within the nucleus of a single cell. Using the annotated data of developing mouse brain, we evaluated and compared different approaches for preprocessing and factorization analysis. This way, we aim to get a better understanding of which specific chromatin interactions are manifested across different cell types. A complex and still open problem in the field of genomics. Through experimentation, we noticed how the low SNR is complex to adjust for, and misdirects standard approaches in preprocessing and clustering. Using an explainable technique, such as factorization, we aim to detect which preprocessing techniques improve the SNR and where the crucial information for clustering is stored in the data. Once a proper ARI clustering score is reached, the resulting factor matrices allow for the immediate detection of the most contributing contacts for each cluster. In a next phase, this information can be linked to the accompanied transcriptome data for the detection or confirmation of (un)known genes.

7.2 Remaining challenges

While it is clear how MSI and other omics technologies are highlighting the future in disease diagnostics and precision medicine, there are still challenges to overcome regarding their data analysis. We discuss some important ones.

Handling the noise

Data acquired using omics technologies are almost by definition extremely noisy. This includes both technical noise between experiments caused by the equipment, and biochemical variations throughout the samples caused by the complexity that is our biology. Since we still don't fully understand all the (overlapping) sources contributing to the noise signal, the modelling and removal of the noise is extremely difficult.

In this dissertation, we provided several approaches to cope with noise in omics data: comparison of Gaussian vs. Poisson distribution of the noise in NMF, application of noise robust topological data analysis, and a tailored approach able to deal with overlapping peaks. However, we noticed how difficult it is to fully capture all the effects of the noise sources and modelling the data accordingly. For example, a recent MSI study [83] showed how the noise distribution throughout the m/z spectrum varies from dominance of the technical noise for the lower m/z values to dominance of the chemical noise at the higher end. Additionally, proteins and metabolites have different chemical properties, which means that their chemical noise has different characteristics. Does this mean separate models need to be developed for each type of molecule? It remains challenging and requires bioinformaticians to have an excellent understanding of all aspects, both biological and computational, for the development of tailored denoising algorithms.

Normalization

Another challenging consequence of the complex relationship between signal abundance and importance, is the normalization of omics data. Normalization allows for the correct comparison between different experiments as well as between different measurements within the same experiment. In MSI, for example, the uneven application of a matrix might cause one part of the tissue to ionize more than another part, which causes higher intensities in the measurements. In multiplexed IHC data, such as MILAN, a set of markers is measured and cleaned off, before new ones are applied. This causes variation between the signals throughout the different rounds of measurements. In scHi-C data, which is very sparse,

Except for the work in Chapter 6, we have refrained ourselves from experimentation with normalization approaches. However, it remains a first step prior to further downstream analysis and could influence the performance of all subsequent methods. Proper understanding of the experimental and biological bias between measurements is necessary for the development of good normalization algorithm. A gold standard for data normalization is still an open problem in many areas of omics data analysis.

Data quality, availability, benchmarking and reproducibility

The availability of annotated and open source data is still a major hurdle in omics data analysis. Reasons for this are the cost of these experiments, i.e. the machinery, and the cost of proper annotations, i.e. the experts and their time.

This makes high quality data sets scarce and hinders many labs of publishing their data as well. This is why the focus in omics is more on the unsupervised rather than supervised analysis.

In extension of the data quality and availability, we highlight the need for benchmarking data sets to objectively compare analytical tools. Computational analyses are handling the complexity of the data better than ever before. With a lack of ground truths and new approaches presented continuously, there is an increasing demand towards benchmarking. This could even be extended to benchmarking data sets for each pathology, as data may differ significantly and require different tools. The selection of high quality benchmarking data sets requires detailed review and consensus within each omics community. This way omics data analysis can move faster towards the best performing approaches through more objective comparisons. Where both positive and negative results should be embraced.

Finally, differences in sample preparation and preferences in data preprocessing across labs can introduce (minor) differences to the data sets and make comparison more difficult [10]. Not only for MSI, but for many omics modalities this remains a challenge. To perform more robust comparisons across both wet and dry labs, the omics community should focus on reproducibility: sharing as much as possible regarding sample preparation and downstream analysis.

Rank estimation for NMF

When using NMF, the only hurdle remains proper estimation of the rank, i.e. the number of components. When a data set contains noise, we want to model the true signal without the noise. This means there exists a rank r, smaller than the size of the data, able to capture the r main trends in the data. In an ideal setting, the residual between the original data and the NMF approximation would only contain the noise. Unfortunately, without any prior information, estimating the number of main trends is difficult. In this work, we used an SVD decomposition and looked at the size of the singular values to estimate the rank. Other, more advanced techniques exist [193], but a detailed review on rank selection specifically for MSI and other omics is still missing.

7.3 Future perspectives

While there are still many challenges to overcome, the future of omics data analysis is bright. With omics, and more advanced computational tools, we can expect major medical breakthroughs in the coming years.

Mass spectrometry imaging

Within MSI data analysis, the opportunities for improvement tackle the challenges mentioned above.

- **Standardization**

  To allow for more reproducibility across labs, standardized protocols and quality control should be in place for data acquisition, preparation, measurement, and analysis [211]. When good standards are in place, we can undertake the hurdles of reproducibility and high quality data sets. This will improve the availability of annotated data sets and allow merging them for better accuracy. With better standardization protocols in place, computational pipelines can provide support to the pathologist and allow for faster, and maybe better diagnostics [23].

- **Open-source data sets and software**

  For MSI, the *.imzml* format and the METASPACE platform [207] have contributed to a shift in uniformity of software tools and data availability, respectively. With METASPACE, labs all around the world are now uploading their data for metabolite annotations and in exchange making the data publicly available. By being more open in terms of data and software, benchmarking can take place. This allows for faster and better understanding of omics data. To this end, the work in Chapter 4 is accompanied by a GitHub tutorial demonstrating its use. Providing peers in the field the possibility to apply the rr2-NMF code on their own data.

- **NMF for MSI**

  As mentioned before, the rank selection should be further investigated when using NMF for MSI or other omics data. To accommodate for a possible bad estimate of the rank, the residual could still be inspected for signals suppressed as noise. Performing a subsequent NMF or UMAP analysis on the residual, could identify smaller signals and improve the denoising of the data, by reconstruction of the data without the true noise component. Alternatively, more intelligent preprocessing and normalization of the data, might allow subsequent NMF analysis to perform even better.

Towards multi-omics data analysis

Once an omic modality is fully understood and presented in a comparable manner, we can perform a joint analysis and derive confirmations about the

tissue under investigation Here, factorization techniques can be easily extended to a multidimensional analysis. This could be either tensor representations, as demonstrated in Chapter 6, or coupled analysis such as SDF [177] or joint-NMF [35, 67]. Because NMF is not a black-box approach, it can be extended by imposing shared or related factor matrices. This way, trends can be discovered and explained across multiple modalities. When these multimodal clusters are discovered, they could reveal interactions between the different modalities or increase the certainty for possible biomarkers. While every omics tool in itself already offers great gains in the understanding of our functioning, using multiple omics technologies simultaneously only widens the possibilities of what can be discovered.

Recent literature confirms how NMF can be of use in multi-omics studies [221]. NMF is explainable, produces easy-to-interpret results, and scales well to bigger data sets. This an important aspect in the future of omics data analysis, where fast and explainable data processing allows for applications in surgical settings and precision medicine.

An important preprocessing step of multi-omics analysis is the correct alignment of the different modalities. Where MSI has a big feature space, it has a lower spatial resolution than microscopic or scRNA-seq. Additionally, most of the time data for the different modalities is obtained from subsequent slices, imposing minor shifts between the regions. Robust registration methods are essential for the performance of combined analyses.

Towards precision medicine

Where artifical intelligence will not take over medicine as we know, we believe it to improve drastically in terms of throughput, both in time and in numbers. Already starting now and even more so in the coming decades, better computational analysis for omics data will pave the way for precision medicine. After an in-depth analysis, mostly done in an unsupervised manner, we can use the obtained knowledge on biomarkers and disease models to build (supervised) tools in the assistance of surgery and diagnostics. This will aid medical experts in their decision making processes and will allow for more time with patients, resulting in better patient care in general.

The devil is in the details

To end this chapter how it began: *the devil is in the details*. This might be the key take-away message when conducting research in omics data analysis methods.

A variety of tools exist to analyze omics data, and they all have relatively good performances. The most abundant signals will always be detected, but since signal magnitude and signal importance is not always positively correlated in omics data, we need to start with more tailored analyses towards the smaller signals and nuances in the data. Especially in complex heterogeneous tissues such as tumours. With each omics modality correctly processed for all the essential information it provides, multi-omics analysis will be next and the remaining mysteries in molecular biology will be further unravelled.

Appendices

Appendix A

Supporting Information for Non-Negative Matrix Factorization and related methods for MSI Data Analysis

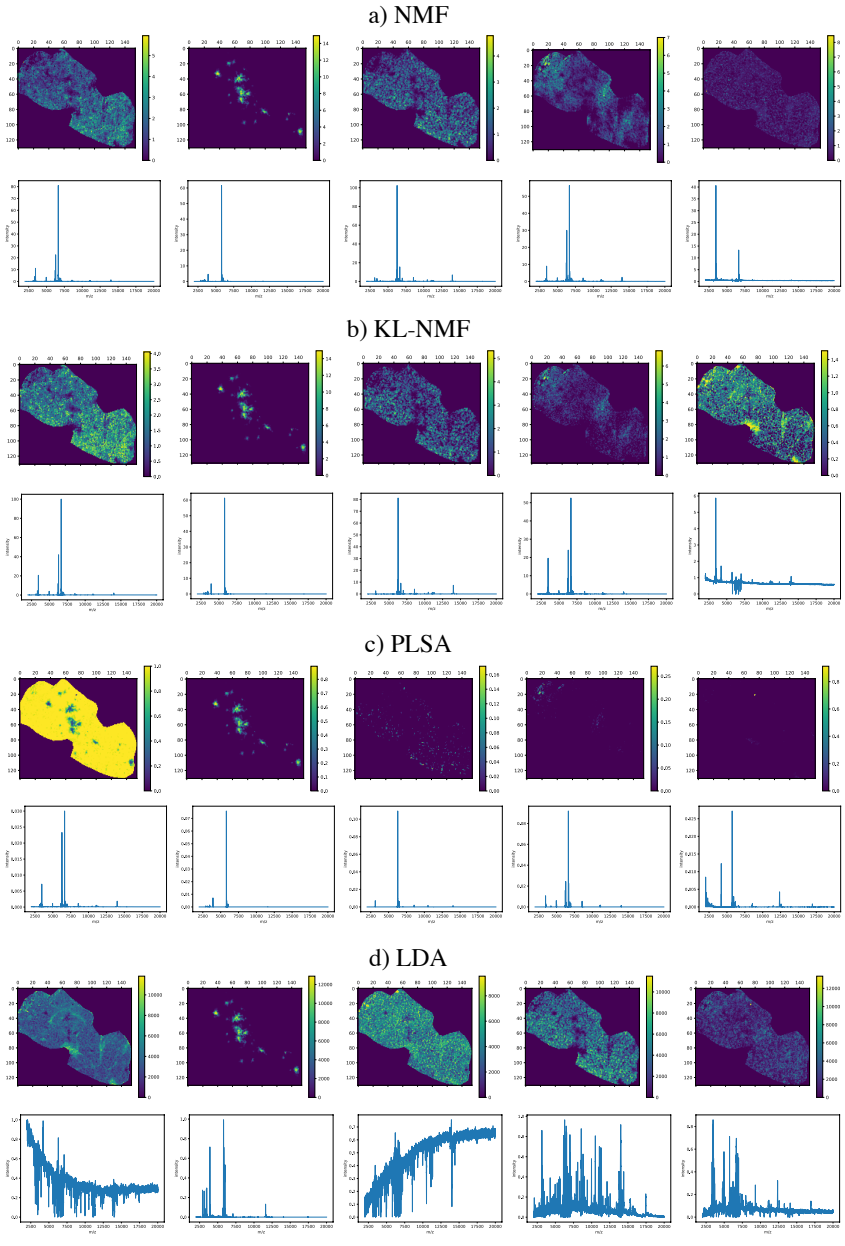


Figure 1: Visual results of the mouse pancreas sample M1. For each method: a) NMF, b) KL-NMF, c) PLSA, and d) LDA we have executed a decomposition in 5 components. The spatial distributions are given in the top rows and the spectral distributions in the bottom rows.

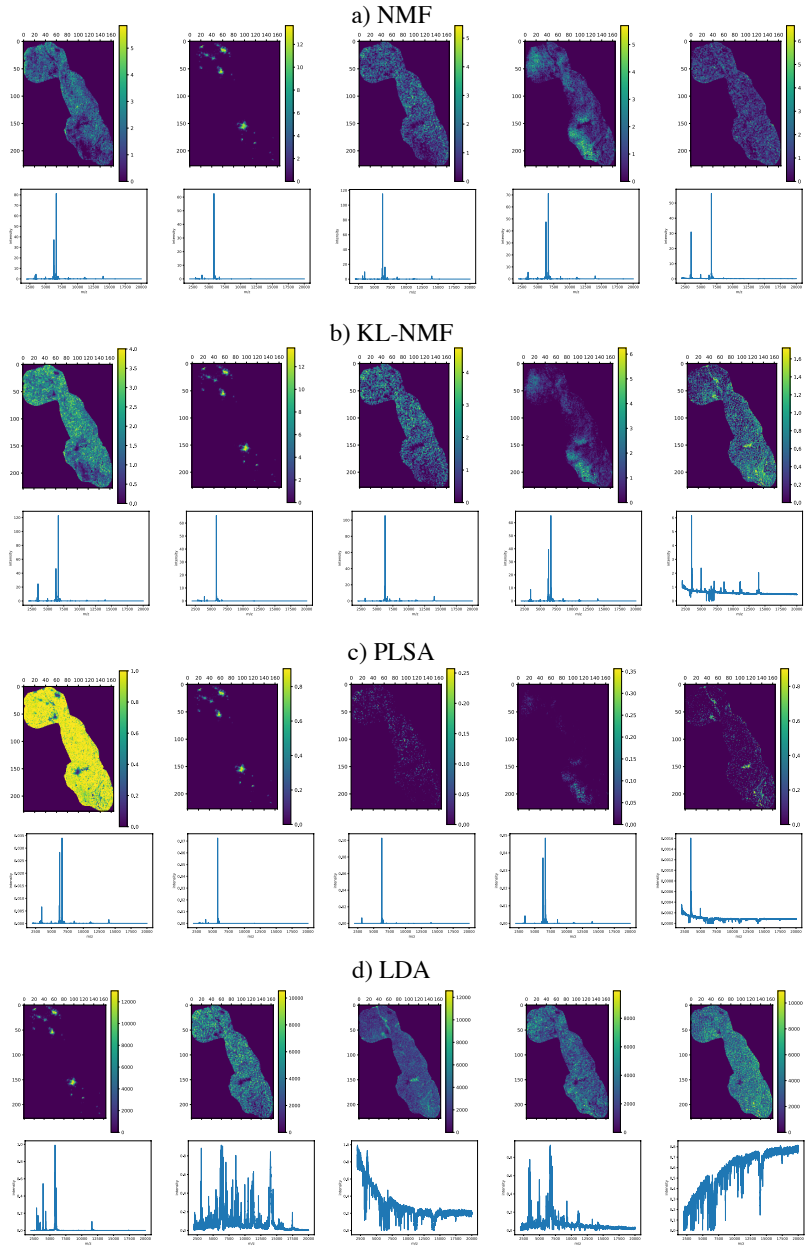


Figure 2: Visual results of the mouse pancreas sample M2. For each method: a) NMF, b) KL-NMF, c) PLSA, and d) LDA we have executed a decomposition in 5 components. The spatial distributions are given in the top rows and the spectral distributions in the bottom rows.

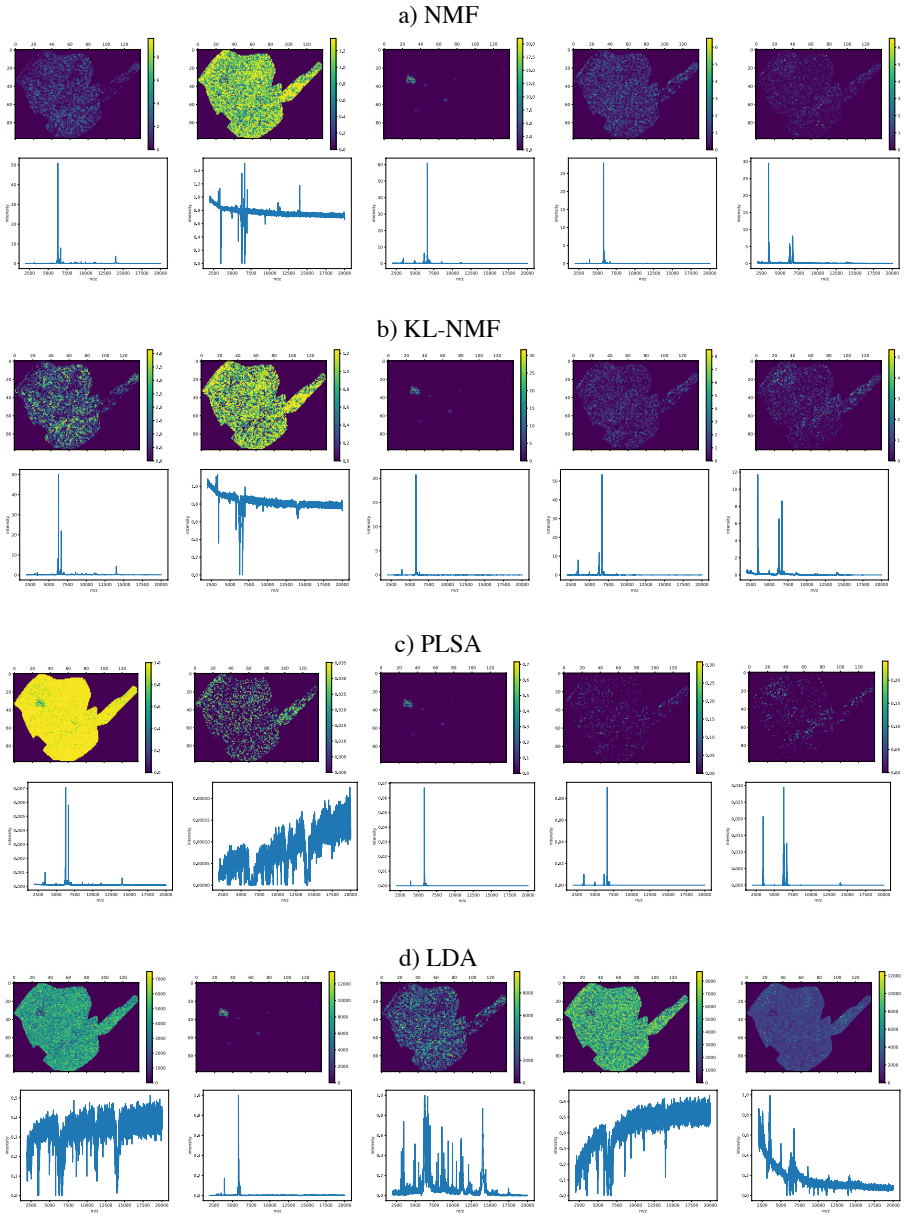


Figure 3: Visual results of the mouse pancreas sample M3. For each method: a) NMF, b) KL-NMF, c) PLSA, and d) LDA we have executed a decomposition in 5 components. The spatial distributions are given in the top rows and the spectral distributions in the bottom rows.

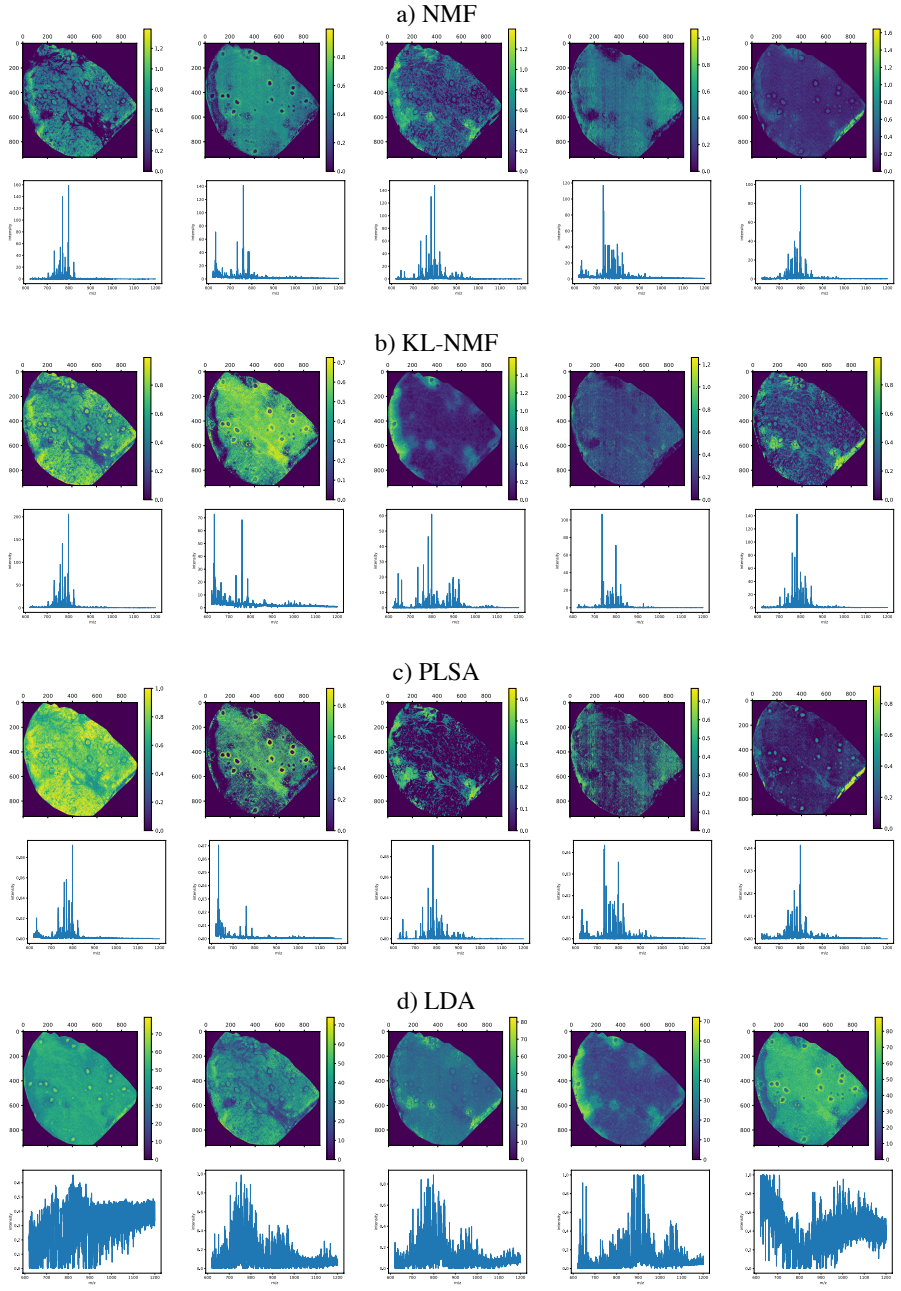


Figure 4: Visual results of the human-lymph-node sample P1. For each method: a) NMF, b) KL-NMF, c) PLSA, and d) LDA we have executed a decomposition in 5 components. The spatial distributions are given in the top rows and the spectral distributions in the bottom rows.

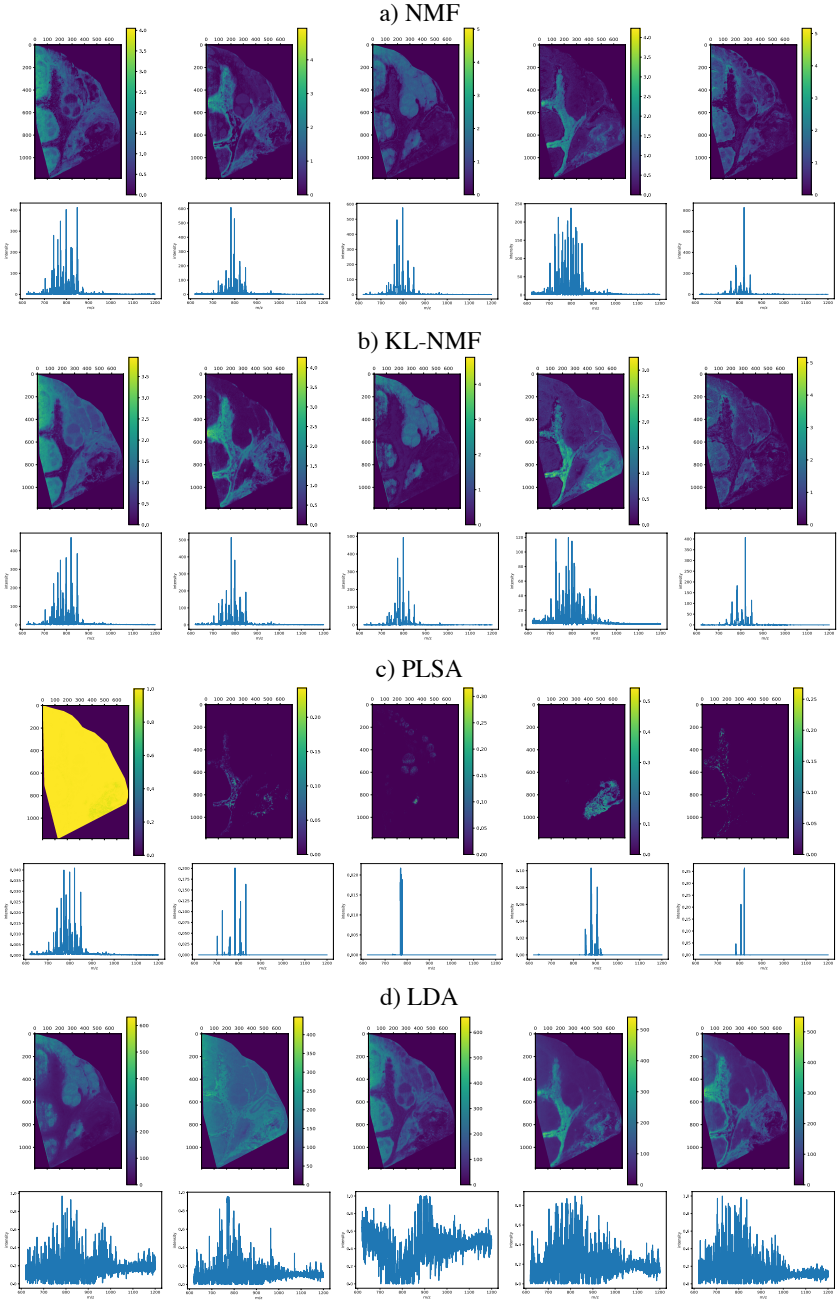


Figure 5: Visual results of the human-lymph-node sample P2. For each method: a) NMF, b) KL-NMF, c) PLSA, and d) LDA we have executed a decomposition in 5 components. The spatial distributions are given in the top rows and the spectral distributions in the bottom rows.

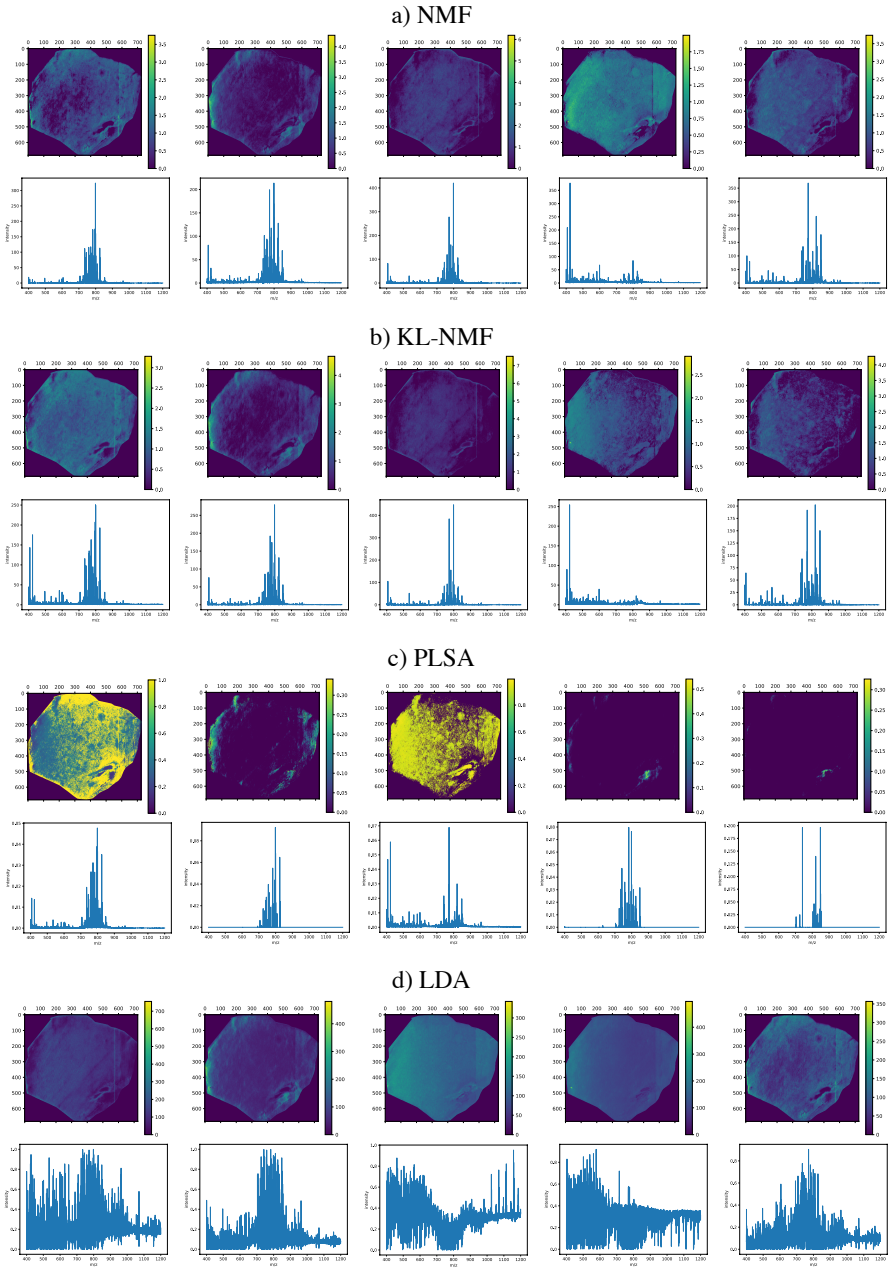


Figure 6: Visual results of the human-lymph-node sample P3. For each method: a) NMF, b) KL-NMF, c) PLSA, and d) LDA we have executed a decomposition in 5 components. The spatial distributions are given in the top rows and the spectral distributions in the bottom rows.

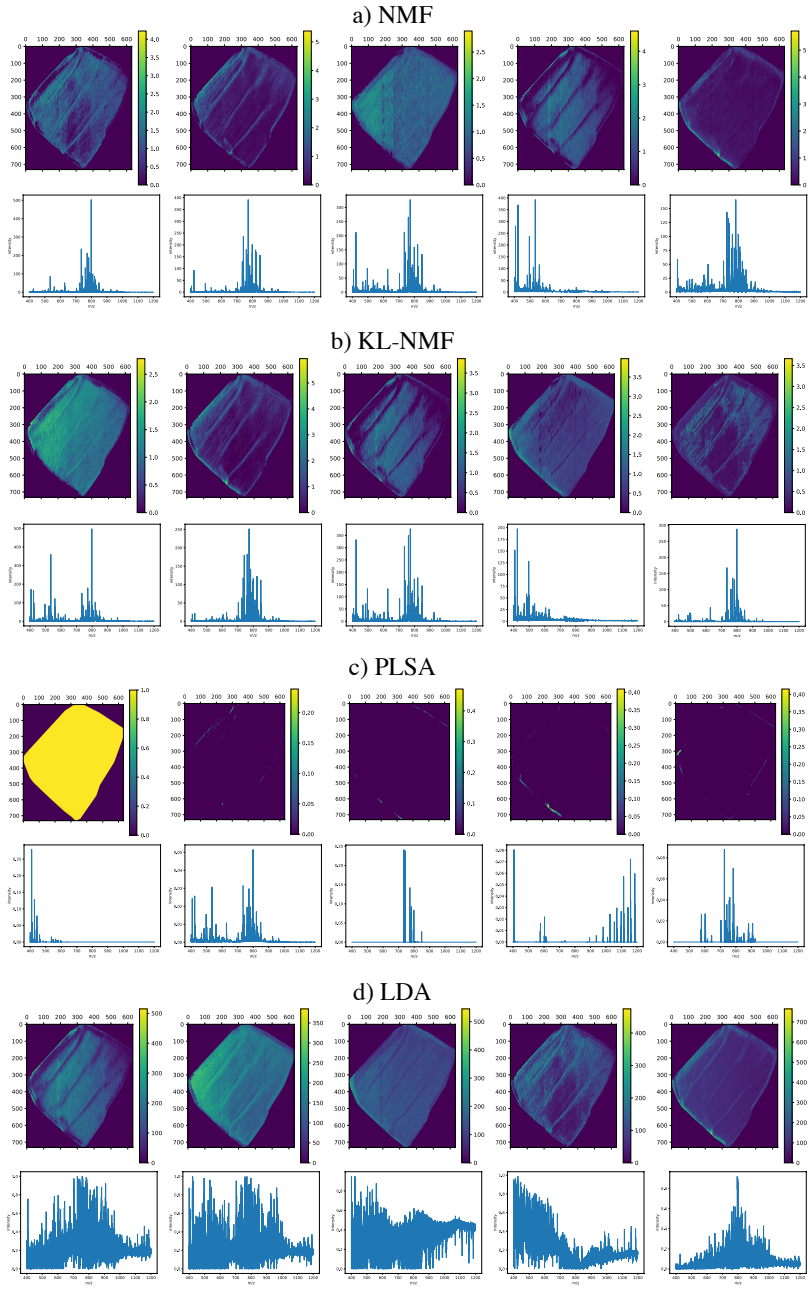


Figure 7: Visual results of the human-lymph-node sample P4. For each method: a) NMF, b) KL-NMF, c) PLSA, and d) LDA we have executed a decomposition in 5 components. The spatial distributions are given in the top rows and the spectral distributions in the bottom rows.

Appendix B

Supporting Information for Spatiochemical characterization of mouse pancreas MSI data using topological data analysis

Supporting Information: 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†,¶

†*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 at STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics*

Contents:

Page S2-5: Clustering obtained from the Mapper algorithm

Page S6-S11: Quantitative comparison

Page S12-S21: Matching ion images with obtained clusters

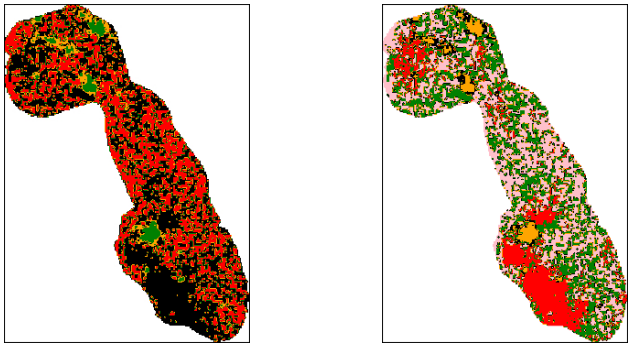
Page S22-S24: Results of algorithm on second dataset

Page S25-S30: Computational requirements

Page S31: Overview of the parameters in a Mapper experiment

Page S32: References

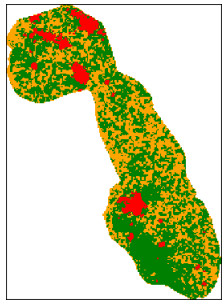
Clustering results main dataset



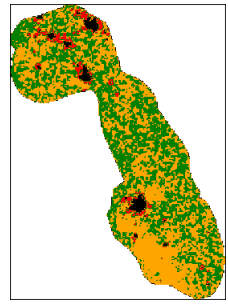
(a) 4 clusters, CH score = 3567

(b) 5 clusters, CH score = 3092

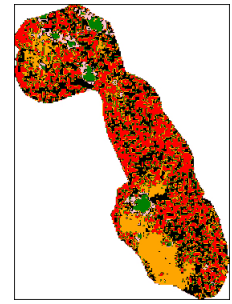
Figure S1: Clusters obtained from the Mapper graph using UMAP as filter function to reduce the original point space to 2 dimensions with 125 intervals and 90% overlap, using 4 and 5 clusters.



(a) 3 clusters, CH-score = 3154



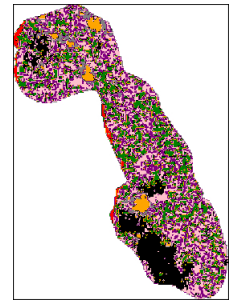
(b) 4 clusters, CH-score = 3491



(c) 5 clusters, CH-score = 3023



(d) 6 clusters, CH-score = 2728



(e) 7 clusters, CH-score = 3257

Figure S2: Clusters obtained from the Mapper graph using UMAP as filter function to reduce the original point space to 2 dimensions with 15 intervals and 25% overlap.

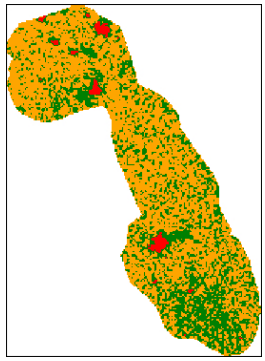
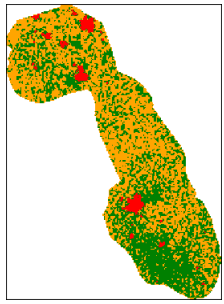
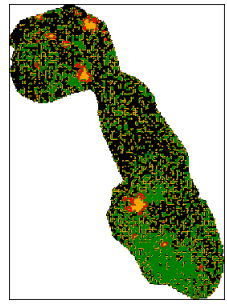


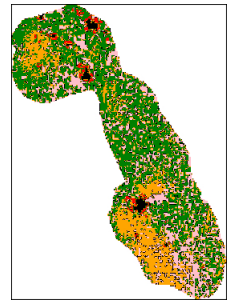
Figure S3: Clusters obtained from the Mapper graph using svd as filter function to reduce the original point space to 3 dimensions with 5 intervals and 50% overlap, using 3 clusters. CH-score = 6802



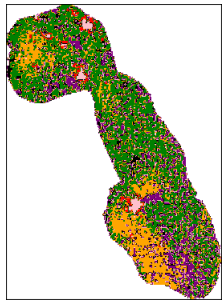
(a) 3 clusters, CH-score = 6231



(b) 4 clusters, CH-score = 5233



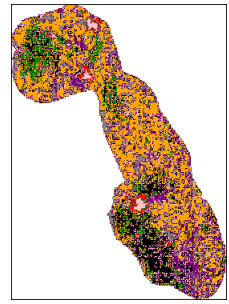
(c) 5 clusters, CH-score = 4725



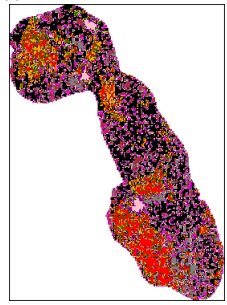
(d) 6 clusters, CH-score = 4593



(e) 7 clusters, CH-score = 3884



(f) 8 clusters, CH-score = 3550



(g) 9 clusters, CH-score = 3440

Figure S4: Clusters obtained from the Mapper graph using svd as filter function to reduce the original point space to 4 dimensions with 5 intervals and 55% overlap.

Quantitative comparison

This section is an addition to the quantitative comparison subsection from the "Results and Discussion". Many parameter values were tested, but only a small part of them could be shown in the main text. Here, an overview of the results is shown. Figures S5, S6, S7, S8 and S9 show an overview of the final scores of the algorithm without any human interaction. Each figure shows the final scores for a number of overlap percentages and intervals. During the algorithm each connected component is clustered in turn. At the start of the algorithm the connected components can be seen as initial clusters. If clustering the first component to 3 clusters results in a CH score higher than 2300 or higher than the initial score, this clustering becomes the 'current clustering' and its corresponding CH score becomes the 'current CH score'. Afterwards, this component gets clustered to 4 clusters. If this clustering results in a score higher than 2300 or higher than the 'current CH score', which is the CH score for 3 clusters or the initial CH score, this clustering becomes the 'current clustering' and its corresponding CH score becomes the 'current CH score'. Finally, this component gets divided into 5 clusters. Again, this clustering becomes the new 'current clustering' with corresponding 'current CH score' if its CH score is higher than the 'current CH score' or higher than 2300.

For the following component the same steps apply. The 'initial CH score' gets replaced by the final 'current CH score' from the previous component. These steps are repeated for each component. And the final 'current CH score' is reported in the figures below.

In the "Results and Discussion" section we looked at all the results, this means including the intermediate results. The highest CH scores mentioned in that section are mostly resulting from intermediate clusterings. As was mentioned before, a higher number of clusters can be informative even if the CH score decreases. This is the reason a higher number of clusters is accepted if it has CH score higher than 2300 even if its score is not higher than the previous score. Also for this reason we looked at higher amounts of clusters for some specific parameter settings.

It is clear by looking at the figures that the optimal overlap percentage is very dependent on the number of overlapping intervals. The higher the number of intervals, the higher the overlap percentage should be to obtain high scores. Parameter combinations using a high number of intervals together with a low overlap percentage result in a lot of small, scattered connected components in the Mapper graph. For the higher dimensions, less parameter combinations were tested because of their higher run times.

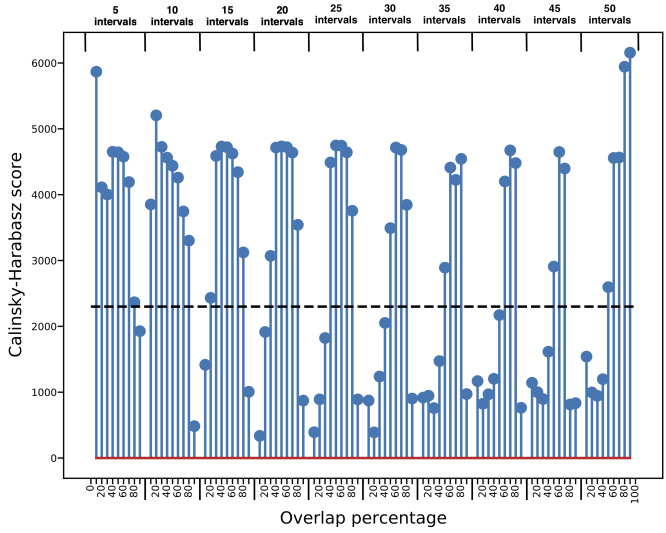


Figure S5: CH score computed for Mapper using dimensionality reduction to 2 dimensions and svd as filter function for a range of overlap percentages and intervals.
The dashed line represents the benchmark score of 2300.

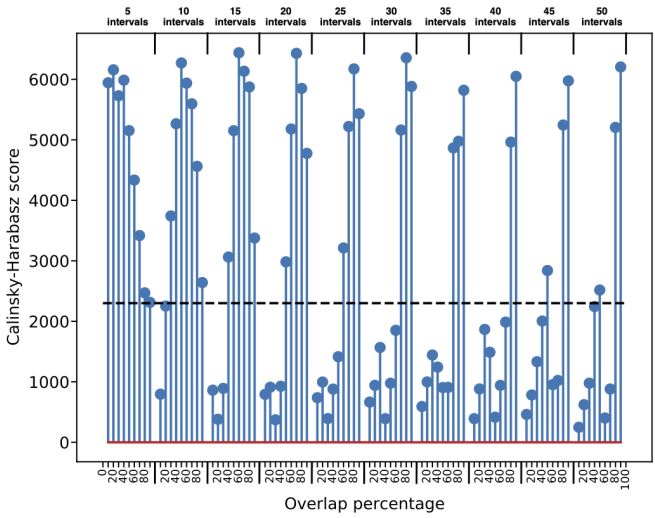


Figure S6: CH score computed for Mapper using dimensionality reduction to 3 dimensions and svd as filter function for a range of overlap percentages and intervals.
The dashed line represents the benchmark score of 2300.

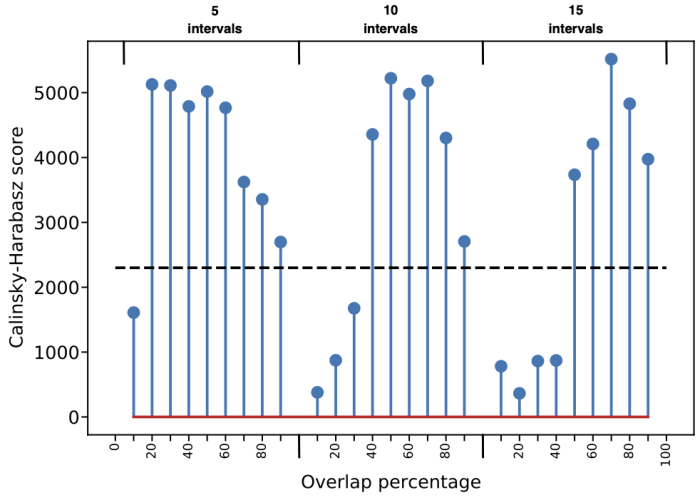


Figure S7: CH score computed for Mapper using dimensionality reduction to 4 dimensions and svd as filter function for a range of overlap percentages and intervals.
The dashed line represents the benchmark score of 2300.

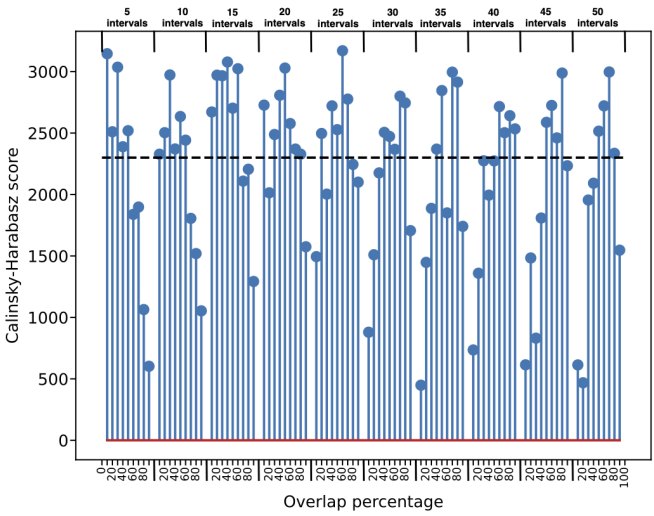


Figure S8: CH score computed for Mapper using dimensionality reduction to 2 dimensions and UMAP as filter function for a range of overlap percentages and intervals. The dashed line represents the benchmark score of 2300.

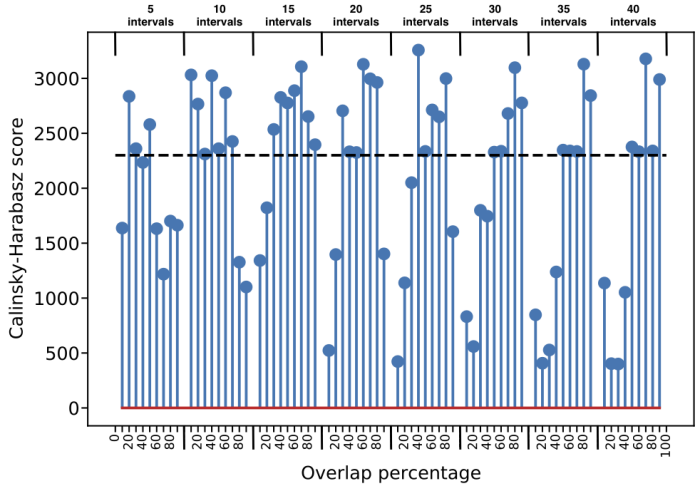


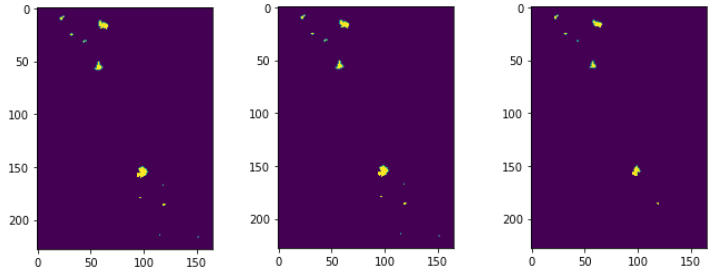
Figure S9: CH score computed for Mapper using dimensionality reduction to 3 dimensions and UMAP as filter function for a range of overlap percentages and intervals. The dashed line represents the benchmark score of 2300.

Matching ion images

In the main text, a comparison was made between the ion images and the obtained cluster for the core of the pancreatic islets. Here this comparison will be continued for some other clusters.

Cores of the pancreatic islets

In the main text a mapping was shown between the high intensity parts of the ion image for m/z-value 6006.67 and the cluster of the cores of the pancreatic islets. Since no straightforward choice can be made for the correct cut-off intensity for the mask some other figures were shown here with their corresponding Dice scores. The Dice scores are calculated for the mask of the ion image and the cluster of the cores of the islets.



(a) Intensity cut-off = 5, Dice score = 0.63

(b) Intensity cut-off = 15, Dice score = 0.89

(c) Intensity cut-off = 25, Dice score = 0.67

Figure S10: Overlap images between the ion image of m/z-value 6006.67 and the cluster containing the cores of the pancreatic islets, together with their corresponding Dice scores. Multiple intensity cut-offs were used.

Edges of the pancreatic islets

In Figure S12a, an ion image corresponding to insulin was shown. This image shows both the edges and the cores of the islets. In Figure S12c, we can see an ion image showing only the cores of the islets. In Figure S12b, an ion image can be seen where the m/z-value 14011 Da occurs almost everywhere but not in the cores of the pancreatic islets. The combination of these figures shows the reason that the edges of the islets were clustered separately from the cores. Unfortunately, in the high peaks, no ion image is found that shows a clear high intensity in only the edges and thus a spatial comparison is not easily made.

Another interesting thing that can be shown is the spatial distribution of glucagon. In Figure S11, the ion image corresponding to m/z-value 3484 can be shown, this m/z-value corresponds to glucagon. Glucagon is a hormone that lowers the blood sugar levels. It is produced in the pancreatic islets, but unlike insulin it is mainly produced in the non-β cells.[S1] In this dataset there is not a high abundance of glucagon, but it can be seen that it is present in the shell of one of the pancreatic islets.

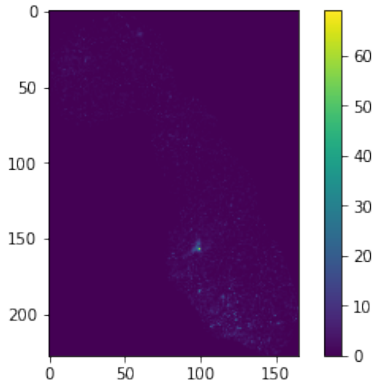
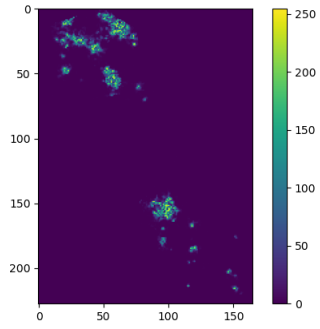
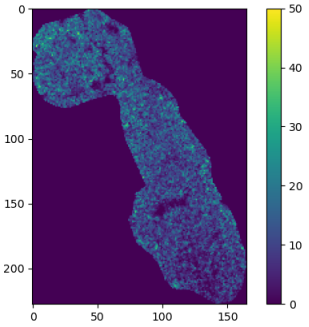


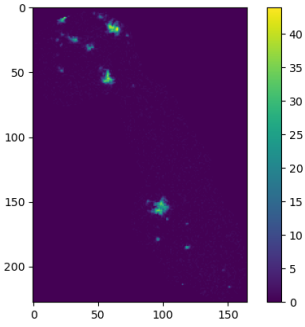
Figure S11: Ion image for m/z value 3484.40



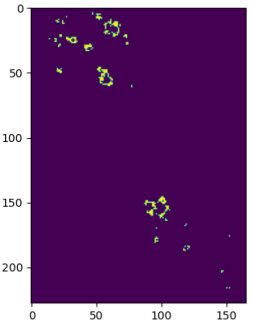
(a) Ion image for m/z-value 5800.77



(b) Ion image for m/z-value 14011.03



(c) Ion image for m/z-value 6006.67

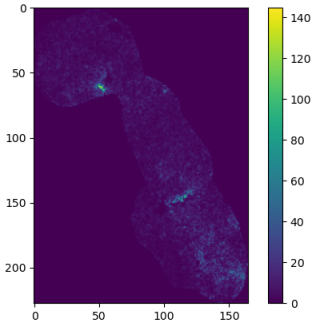


(d) Cluster corresponding to the edges of the pancreatic islets.

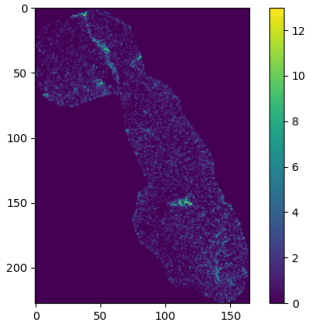
Figure S12

Region 2

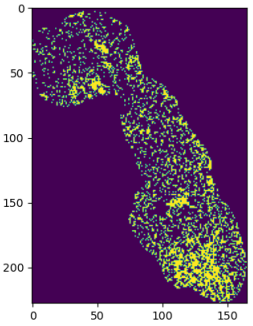
When looking at the highest peaks for *region 2*, which contains the bulk of the blood vessels, it seems that the m/z-values representing this region have a lower intensity than some m/z-values present in the complete pancreas. In Figure S12b, *region 2* seems much darker than the rest of the pancreas, nevertheless, this m/z-value represents the fourth highest peak in Figure S17. This difference in intensity most likely contributes to the separation of this cluster even though the cluster itself does not have a high intensity here. The fifth highest peak for this cluster belongs to m/z-value 4966.88, this clearly shows a small part of *region 2*. This ion image can be seen in Figure S13a, this means that a further clustering of this region would probably lead to this part becoming a separate cluster. Finally in Figure S13b, the ion image corresponding to m/z-value 2008.36 can be seen. This very clearly shows *region 2* which is positioned next to the main pancreatic islets. These islets produce hormones, which are secreted into the blood flow.[S2] This is only the 19th highest peak and more similar ion images can be found in lower peaks. Meaning that the m/z-values representing the blood vessels are slightly overshadowed by other m/z-values. The blood vessels are very small, with diameters reaching to 10 μm. As can be seen in Smets et al. (2020) and Smets et al. (2019), the spatial resolution of the data acquisition method is 50 μm. Because of this, the pixels or datapoints containing the blood vessels, will also contain other tissue parts. Let's call the datapoints from the dataset that are not contained in the islets, *region 2* or *region 3*, the 'main' part of the dataset. This would mean that the pixels containing blood vessels also contain, for a large part, tissue from the 'main' part of the pancreas. The blood vessels still influence the obtained mass spectra enough that the Mapper method can recognize them as a separate cluster, but the m/z-values that are specific for the blood vessels will be overshadowed by other m/z-values from the 'main' part of the dataset. Because of this, the cluster containing blood vessels suffers a lot from the high noise levels in this dataset.



(a) Ion image for m/z-value 4966.88



(b) Ion image for m/z-value 2008.36

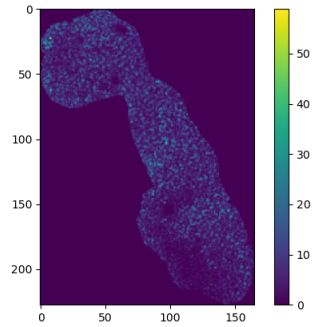


(c) Cluster corresponding to *region* 2.

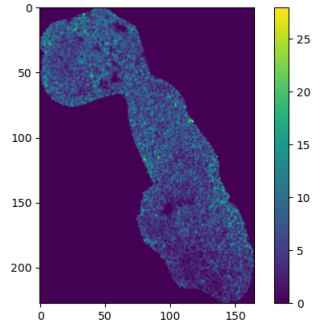
Figure S13

Region 3

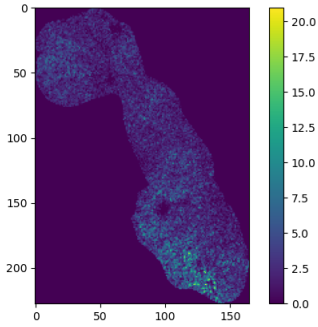
For *region 3*, the ion images are the least clear. When analysing a large amount of ion images, it can clearly be seen that the intensities differ, but based on the ion images of the high peaks this would not look like a different cluster. In Figure S14a, S14b and S14c, some examples are shown of ion images that show slight differences in intensities for this part of the pancreas. In Figure S14d, the corresponding cluster is shown.



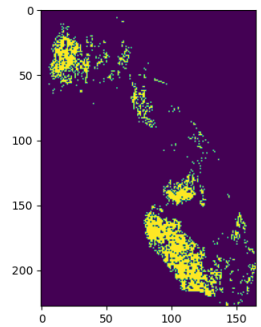
(a) Ion image for m/z-value 3139.52



(b) Ion image for m/z-value 7006.58



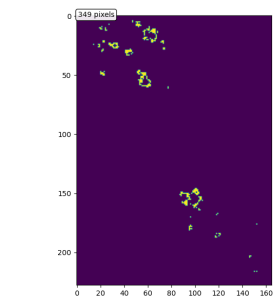
(c) Ion image for m/z-value 6727.32



(d) Cluster corresponding to *region 3*.

Figure S14

Finally, in Figures S15, S16, S17 and S18, the clusters can be seen with their corresponding mean, maximum and minimum mass spectra and the 60 highest m/z-values. Figures S15, S16 and S17 were obtained using svd, 3 dimensions, 5 intervals, 50% overlap and 5 clusters. Figure S18 was obtained using svd, 4 dimensions, 5 intervals, 55 % overlap and 5 clusters.



60 largest m/z values

5800.77	11067.95	7885.51
6650.11	6727.32	10946.99
6278.21	6354.13	13775.53
3458.66	8814.63	7212.48
3923.23	5575.57	2483.21
14011.03	5710.69	8948.47
6210.0	8074.68	9402.74
6006.67	11349.78	2300.48
3327.4	6578.05	8369.38
4336.31	3535.88	13942.83
6122.49	14086.96	2142.19
8567.55	3665.85	5172.78
4966.88	2216.83	7113.39
6857.3	9981.83	3005.68
3140.81	4750.69	7979.45
6980.84	7046.47	3243.76
6484.11	10508.16	2366.11
2901.45	5445.6	2579.73
5877.99	2008.36	8737.42
11147.74	2075.27	4167.73

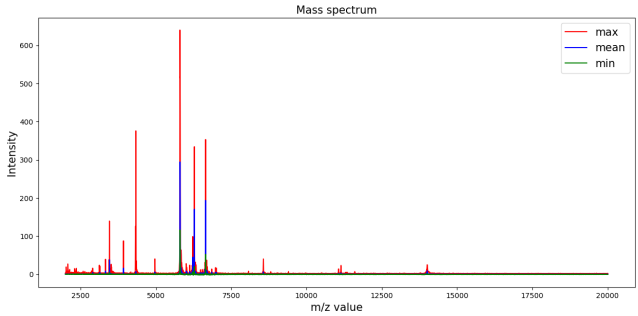
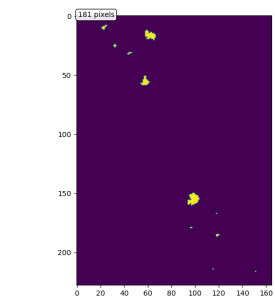


Figure S15: Edges of the pancreatic islets together with the mean, max and min mass spectra from the cluster and the 60 highest m/z peaks.



60 largest m/z values

5800.77	6484.11	2009.64
6650.11	11147.74	4750.69
6278.21	6726.04	2216.83
3923.23	3525.58	8813.35
6006.67	11066.66	2617.05
3458.66	6351.56	7100.52
5877.99	5535.68	2483.21
2901.45	3005.68	4452.13
3124.08	11308.6	14076.66
4336.31	5641.2	7212.48
3327.4	5072.4	9981.83
6210.0	6580.62	7885.51
11600.72	8075.97	3596.36
6122.49	3664.56	13776.82
5711.98	10509.45	4243.66
14009.75	5219.11	2151.2
4966.88	11668.92	2317.21
8567.55	5446.88	13941.54
6857.3	11533.8	9401.45
6980.84	4169.02	4570.52

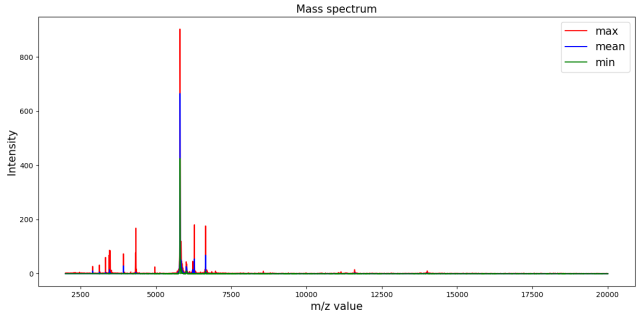
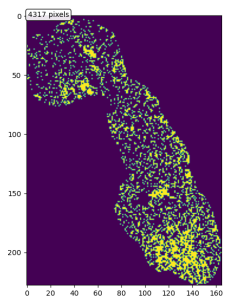


Figure S16: Cores of the pancreatic islets together with the mean, max and min mass spectra from the cluster and the 60 highest m/z peaks.



60 largest m/z values

6650.11	8813.35	5675.95
6278.21	2216.83	7884.22
3458.66	6431.34	6047.85
14011.03	8074.68	13774.25
4966.88	5575.57	3062.31
6211.29	11351.06	2831.96
3326.12	3665.85	2960.64
8567.55	2077.85	5445.6
7006.58	2483.21	2637.64
6857.3	2288.89	3243.76
6122.49	9979.26	7113.39
11146.45	14084.38	3758.51
6547.16	4284.84	4210.2
5800.77	4750.69	4003.01
3140.81	2355.81	9402.74
11067.95	2148.63	4839.48
6727.32	3923.23	4408.38
3537.16	2703.27	10508.16
2008.36	2557.85	5171.49
6354.13	7212.48	6933.22

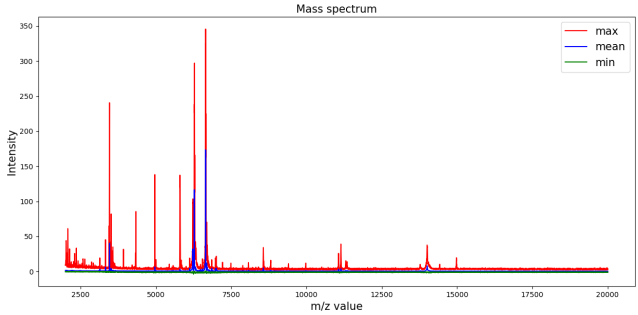
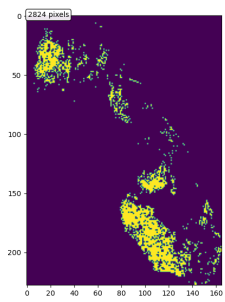


Figure S17: *Region 2* together with the mean, max and min mass spectra from the cluster and the 60 highest m/z peaks.



60 largest m/z values

6651.4	8075.97	2447.18
6279.49	3537.16	5677.23
3459.95	14080.52	7118.53
6211.29	5575.57	9405.31
14012.32	6563.89	8948.47
8568.84	6049.14	8369.38
3327.4	9981.83	4840.77
6857.3	7046.47	7978.17
6123.78	2218.12	2092.0
4966.88	3924.51	14220.79
6980.84	3665.85	2317.21
6485.39	13778.11	8738.71
3140.81	4286.12	3243.76
5800.77	2009.64	14425.41
11149.02	7885.51	2960.64
6728.61	4751.97	2704.56
11069.24	10509.45	3062.31
6355.42	7213.76	13927.39
8815.92	5446.88	2831.96
11352.35	10948.27	5172.78

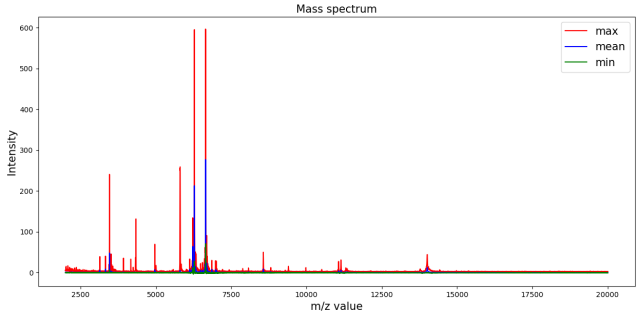
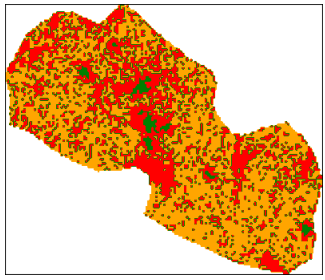
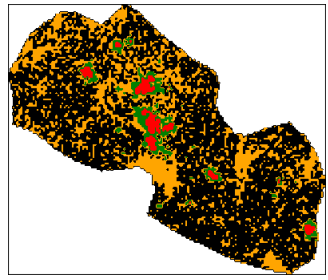


Figure S18: *Region 3* together with the mean, max and min mass spectra from the cluster and the 60 highest m/z peaks.

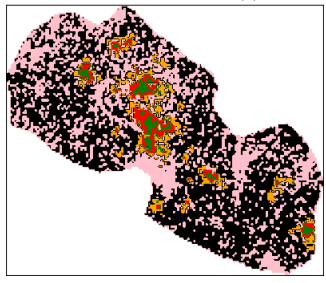
Clustering results second dataset



(a) 3 clusters, CH-score = 7081

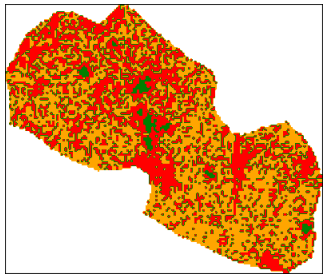


(b) 4 clusters, CH-score = 6194

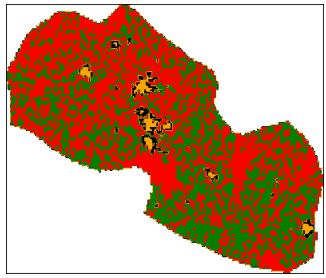


(c) 5 clusters, CH-score = 4630

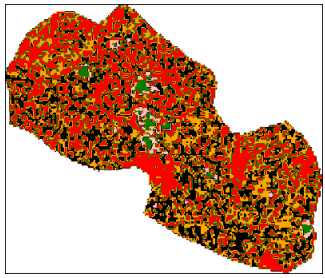
Figure S19: Clusters obtained from the Mapper graph using svd as filter function to reduce the original points space to 2 dimensions with 25 intervals and 90% overlap for the second dataset.



(a) 3 clusters, CH-score = 6825

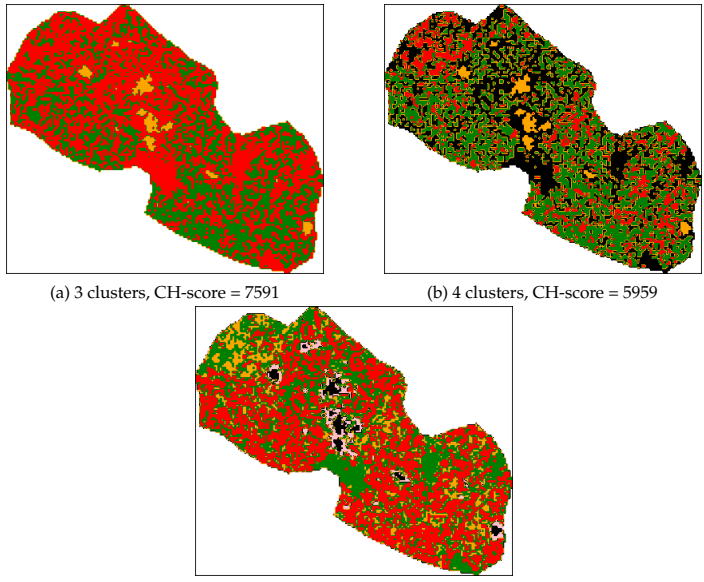


(b) 4 clusters, CH-score = 6076



(c) 5 clusters, CH-score = 5645

Figure S20: Clusters obtained from the Mapper graph using svd as filter function to reduce the original points space to 3 dimensions with 75 intervals and 90% overlap for the second dataset.



(a) 3 clusters, CH-score = 7591

(b) 4 clusters, CH-score = 5959

(c) 5 clusters, CH-score = 5402

Figure S21: Clusters obtained from the Mapper graph using svd as filter function to reduce the original points space to 4 dimensions with 15 intervals and 70% overlap for the second dataset.

Computational requirements

In this part of our work, we elaborate on the run time of the experiments, since this is an important limitation for real life applications. The time comparison between our benchmark method and the Mapper method was done on a Macbook air 2020 M1 chip, with 8Gb ram. For the benchmark, the parameters described in Smets et al. (2020)[S3] were used. Averaged over 5 runs, the run time was 30.03 seconds. For the Mapper method the whole algorithm was timed, including making the Mapper graph and clustering the connected components. The same algorithm was used as for the quantitative comparison.

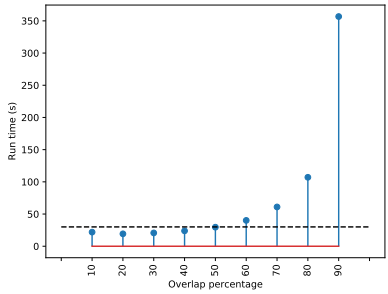


Figure S22: Run time computed for Mapper using dimensionality reduction to 2 dimensions, svd as filter function and using 5 intervals. Run time computation was executed on a Macbook air M1 2020.
The dashed line shows the average timing of the benchmark method (30s).

The Mapper method using svd as filter function to reduce the space to 2 dimensions was timed using 5 intervals with 10, 20, 30, ... 90% overlap percentages. As can be seen in Figure S22, for small overlap the method is slightly faster than the benchmark, which is shown as a dashed line on the figure. Increasing the overlap, however, significantly increases the run time.

To get a more complete overview of the evolution of the run times a larger visualization was made using the Flemish supercomputer (VSC). Running this code using the VSC took

longer than running it on the laptop. In Figures S23 and S24, the run times can be seen on the VSC for dimensionality reduction to 2 dimensions using svd and UMAP as filter function respectively. In these figures, it becomes clear that the higher overlap percentages cause significantly higher run times, which increase even more when higher numbers of intervals are used. It can also be seen that the run time is higher when using UMAP as filter function than when using svd as filter function.

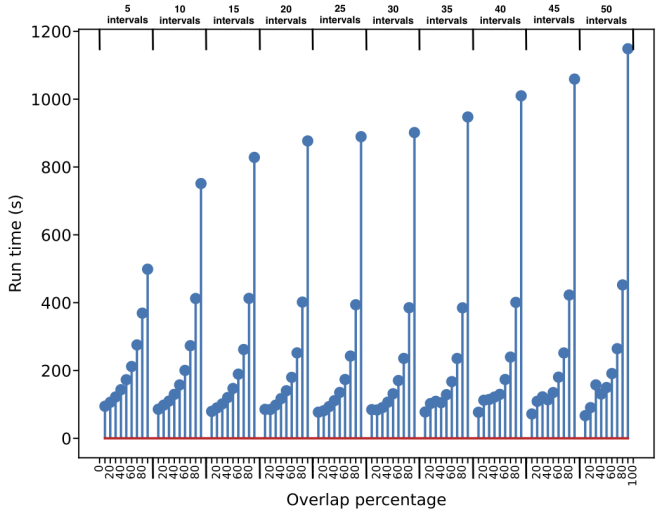


Figure S23: Run time computed on VSC for Mapper using dimensionality reduction to 2 dimensions and svd as filter function for a range of overlap percentages and intervals.

A similar evaluation of run times was done for dimensionality reduction to 3 dimensions with svd and UMAP as filter functions. This can be seen in Figure S25 and S26. The run times are higher than the run times for 2 dimensions when a high overlap percentage is used. For lower overlap percentages, however, the run times are very similar to those

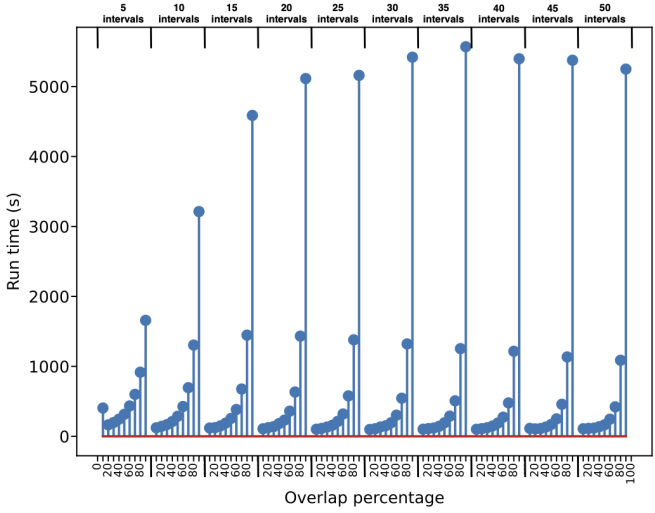


Figure S24: Run time computed on VSC for Mapper using dimensionality reduction to 2 dimensions and UMAP as filter function for a range of overlap percentages and intervals.

for 2 dimensions.

Finally, in Figure S27, the run times can be seen for dimensionality reduction to 4 dimensions using svd as filter function. Because of higher run times less parameter combinations are shown here.

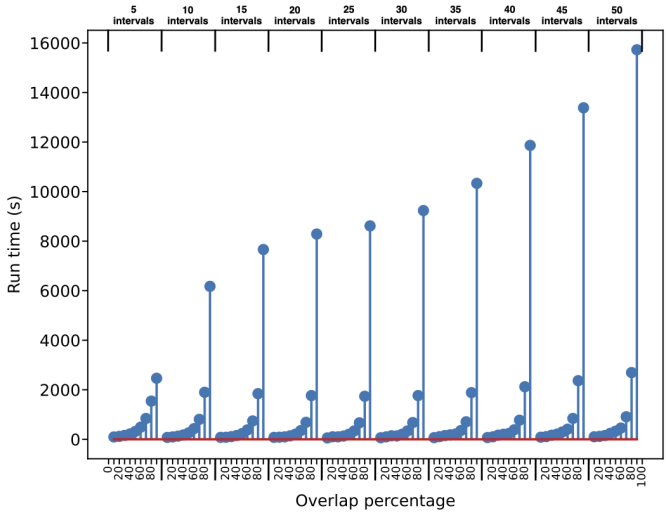


Figure S25: Run time computed on VSC for Mapper using dimensionality reduction to 3 dimensions and svd as filter function for a range of overlap percentages and intervals.

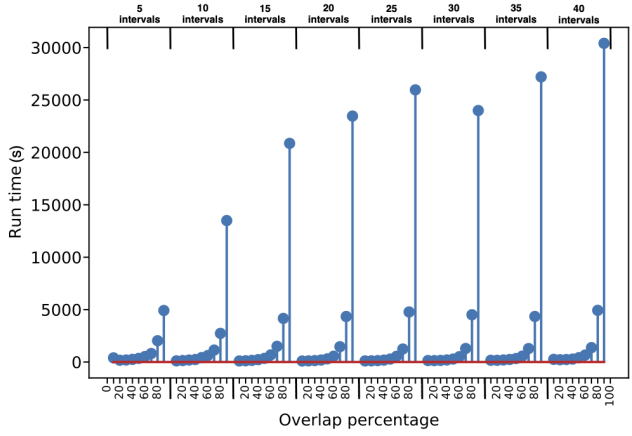


Figure S26: Run time computed on VSC for Mapper using dimensionality reduction to 3 dimensions and UMAP as filter function for a range of overlap percentages and intervals.

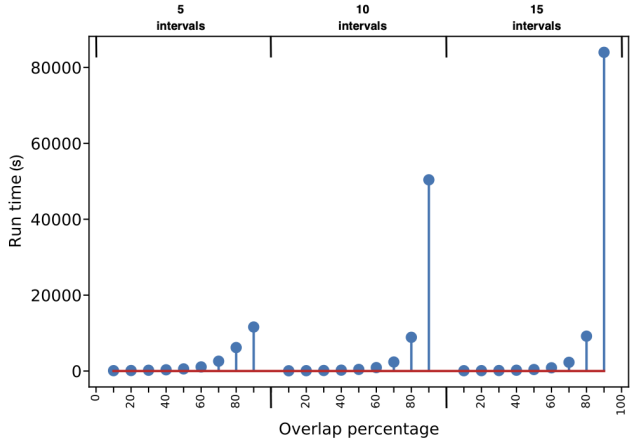


Figure S27: Run time computed on VSC for Mapper using dimensionality reduction to 4 dimensions and svd as filter function for a range of overlap percentages and intervals.

Parameters Mapper graph

Table S1: Table summarizing the parameters that can be adjusted in a Mapper experiment together with the effects they have on the resulting Mapper graph.

Mapper parameters	Function	Effect on Mapper graph
Filter function	Method that will map the original point cloud to a lower dimension	Effects: • Run time • Memory requirements • Distribution of data points in new space
New dimensionality	Dimensionality to which the filter function maps the data points	Effects: • Run time • Which structures can be found in the data set
Number of overlapping intervals	The lower dimensional space is divided into a specific number of intervals	More intervals $\rightarrow$ More, smaller clusters $\rightarrow$ A more detailed overview of the data points
Overlap percentage	Determines how much the intervals overlap	Higher overlap percentage $\rightarrow$ More connectivity between the clusters in the Mapper graph
Clustering method	Method that will be used (in the original high dimensional space) to cluster data points that are located in the same intervals	Effects: • Run time • Memory requirements • Which kind of clusters can be found • Determines which data points will be clustered

References

(S1) Moede, T.; Leibiger, I. B.; Berggren, P. O. Alpha cell regulation of beta cell function. *Diabetologia* **2020**, *63*, 2064–2075.

(S2) Campbell, J. E.; Newgard, C. B. Mechanisms controlling pancreatic islet cell function in insulin secretion. *Nat. Rev. Mol. Cell. Biol.* **2021**, *22*, 142–158.

(S3) 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. *Anal. Chem.* **2020**, *92*, 5240–5248.

Appendix C

Supporting Information for Hierarchical bi-clustering of mouse pancreas MSI data using recursive rank-2 NMF

Supporting Information

Hierarchical bi-clustering of 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

Email: melanie.nijs@esat.kuleuven

Table of Contents

S1. RUNTIMES 2

S2. ISOTOPE SELECTION 3

S3. PANCREAS M1 SAMPLE ANALYSIS USING RR2-NMF 7

A. RR2-NMF UNTIL DEPTH = 2 7
B. RR2-NMF UNTIL CONVERGENCE WITH THRESHOLD = 0.2 13
C. ION IMAGES OF M/Z-VALUES OF INTEREST IN THE ISLETS OF LANGERHANS 23

S4. PANCREAS M2 SAMPLE ANALYSIS USING RR2-NMF 24

A. RR2-NMF UNTIL DEPTH = 2 24
B. RR2-NMF UNTIL CONVERGENCE WITH THRESHOLD = 0.2 30
C. ION IMAGES OF M/Z-VALUES OF INTEREST IN THE ISLETS OF LANGERHANS 40

S5. PANCREAS M3 SAMPLE ANALYSIS USING RR2-NMF 41

A. RR2-NMF UNTIL DEPTH = 2 41
B. RR2-NMF UNTIL CONVERGENCE WITH THRESHOLD = 0.2 47
C. ION IMAGES OF M/Z-VALUES OF INTEREST IN THE ISLETS OF LANGERHANS 59

S1. Runtimes

All experiments below have been run on a 2019 MacBook Pro with 2,6 GHz 6-Core Intel Core i7 and 16 GB RAM. We run each sample and plot and save every level of recursion. These options can be deactivated and increase runtime. We measure both CPU and wall times, which can improve depending on the hardware used.

Table S1 Overview timings for rr2-NMF across samples M1, M2, and M3.

Sample	**M1**	**M2**	**M3**
Size [GB]	1.19	1.66	0.78
Timing CPU/wall [min:s] depth	12:9 / 7:03	15:21 / 4:16	6:25 / 1:47
Timing CPU/wall [min:s] convergence	22:07 / 5:48	32:37 / 8:52	15:06 / 4:49

For bigger samples, a parallel implementation is available to improve runtime.

S2. Isotope selection

Carbon has two stable isotopes: carbon-12 (^{12}C) and carbon-13 (^{13}C). The most abundant isotope is ^{12}C, which has a natural abundance of about 98.9%. ^{13}C, on the other hand, has a natural abundance of approximately 1.1%. The difference in mass between these isotopes is due to the presence of an additional neutron in the nucleus of ^{13}C.When a molecule containing carbon atoms is analyzed using mass spectrometry, the presence of ^{13}C isotopes can cause a shift in the observed m/z values compared to the expected m/z values based on the monoisotopic masses of the elements involved. This is because the presence of ^{13}C isotopes increases the overall mass of the molecule, leading to a higher m/z value.

Using NMF, we automatically focus on the highest peaks, i.e. the average masses across all isotopes. When peaks lie closely together, one can image the distribution of the isotopes overlaps and creates a joint peak containing both. Here we explain how we select the most distinct isotopes of Ins1 and Ins2 for the visualization of the spatial abundances in the ion images.

First, take a closer look at the spectrum of cluster "*M2_convergence_2*" (shown below in Figure S1). This cluster describes the islets of Langerhans and contains both insulins.

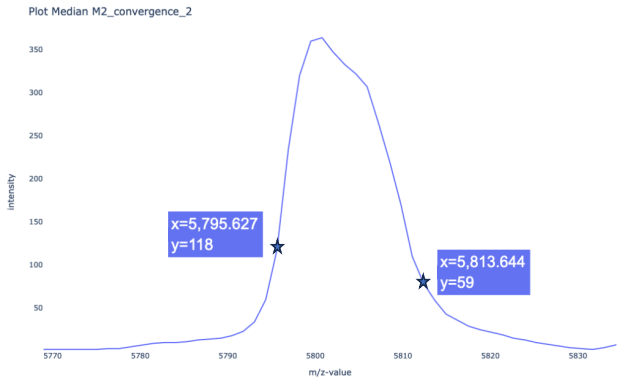


Figure S 1. Zoomed-in spectrum of cluster "M2_convergence_2" in range 5770-5830 Da containing insulin 1 and 2, resp. 5808 Da and 5800 Da.

For the peak values of insulin 1 and 2, rr2-NMF gives us respectively 5800 Da and 5808 Da. Because of the large number of carbon atoms present in this protein, the statistical probability of some of them having isotope C^{13} instead of C^{12} increases considerably. This makes that the peak value shifts from the monoisotopic mass to that of the average across all isotopes. We see that on average, the peak shifts 1 Da for each 1 kDa increment in mass.

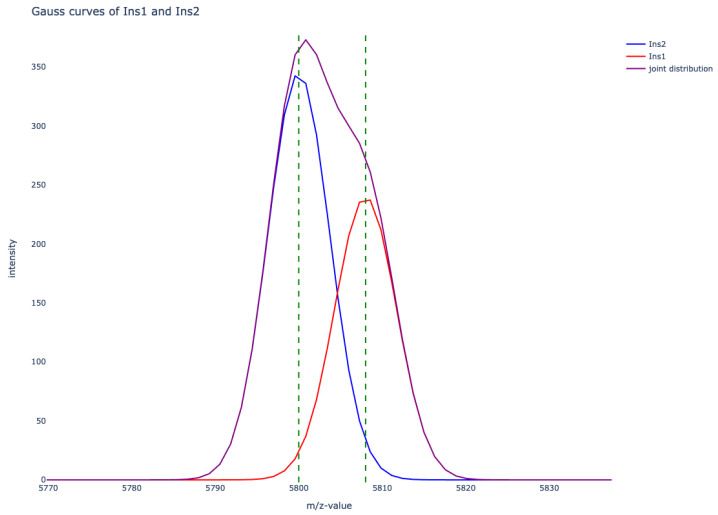


Figure S 2. Zoomed-in spectrum of Gaussian curves of Ins1 and Ins2 in range 5770-5830 Da, resp. 5808 Da and 5800 Da in dashed green line.

Figure S2 demonstrates how the peaks of Ins1 and Ins2 overlap by plotting the expected Gaussian curves at each value, demonstrating the expected distribution of the isotopes. Note that in the real data, other metabolites could be in this range too. We see how the isotopes of Ins1 can influence an ion image at m/z = 5800 Da and vice versa for m/z=5808, it is also important to notice the difference in abundance between Ins2 and Ins1. To select the most representative ion images corresponding to each insulin, we select the ones that are removed respectively – and + 5 Da of the peaks at 5800 and 5808 Da. See the ion images below. As similar analysis is done for the double charged state, now – and + 2.5 Da, as showed below and in the main manuscript.

Figure S3 shows how the intensity patterns across their spatial abundances become more distinct when selecting the most representative isotopes. For the double charged states, however, we are unable to find an isotope at $Ins1^{2+}$+2.5 Da. The ion image at the bottom right in Figure S4 shows the background noise signal. We see however a difference in intensity between the peak value of $Ins1^{2+}$ (top right) and the isotope $Ins2^{2+}$ – 2.5 Da (bottom left).

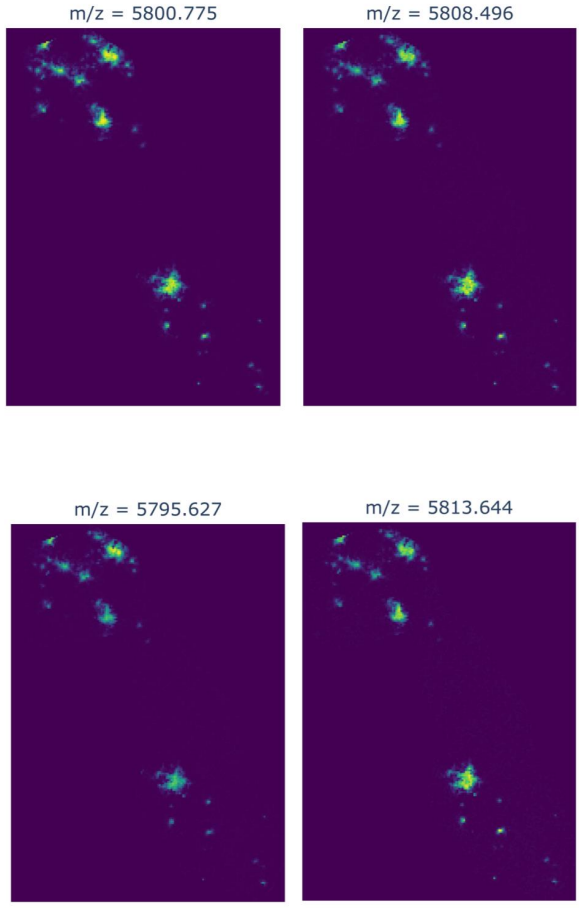


Figure S 3. Ion images of Ins2 (left) and Ins1 (right). Top: peak values, bottom: resp. - and + 5 Da.

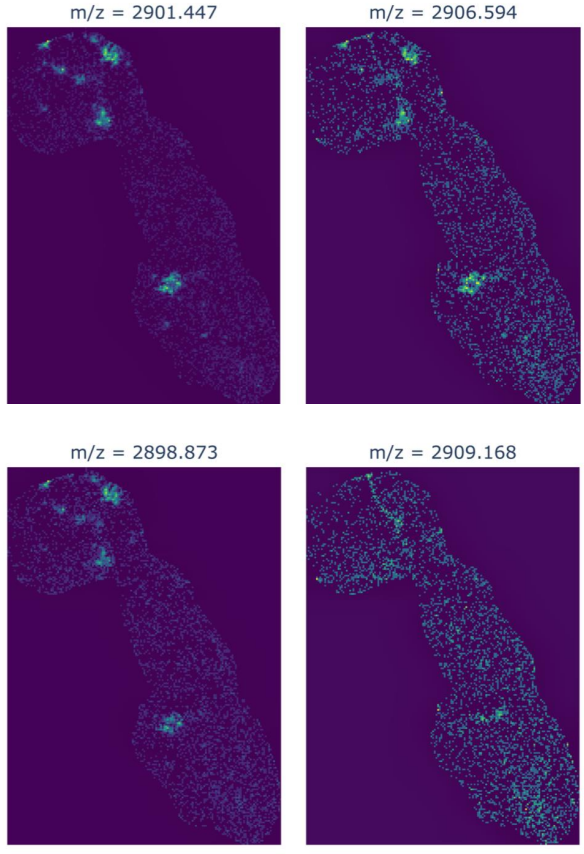
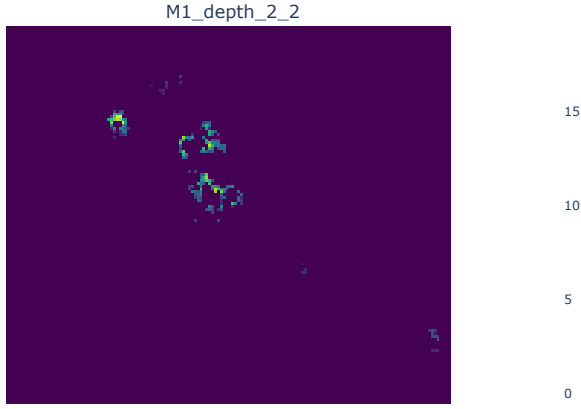
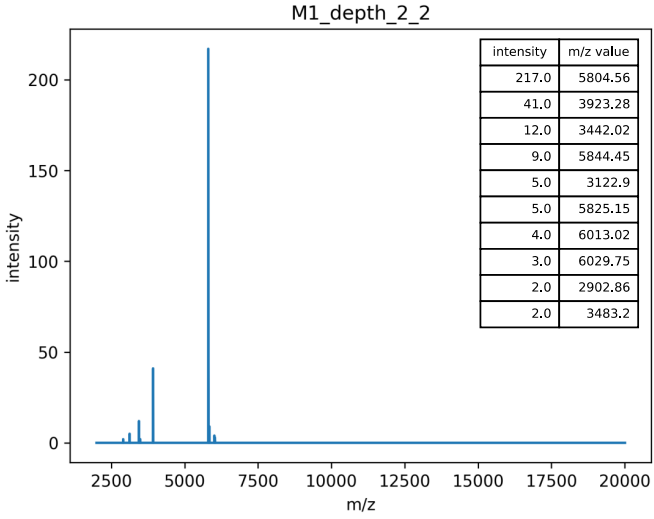


Figure S 4. Ion images of the double charged stated Ins2 (left) and Ins1 (right). Top: peak values, bottom: resp. - and + 2.5 Da.

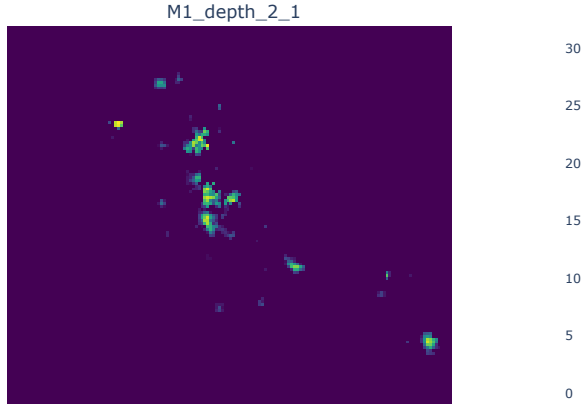
S3. Pancreas M1 sample analysis using rr2-NMF

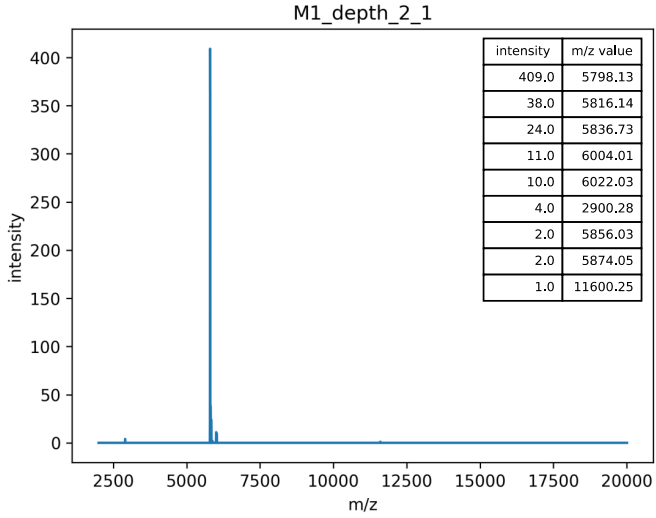
a. rr2-NMF until depth = 2



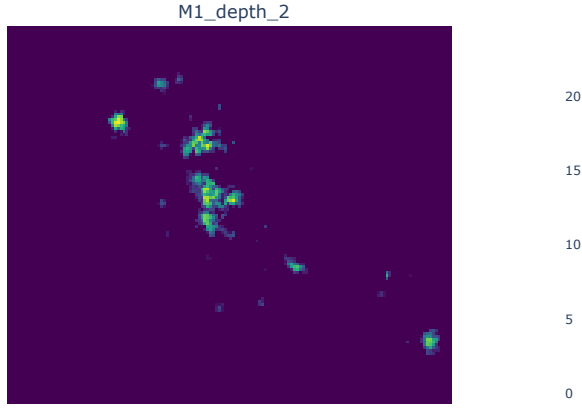


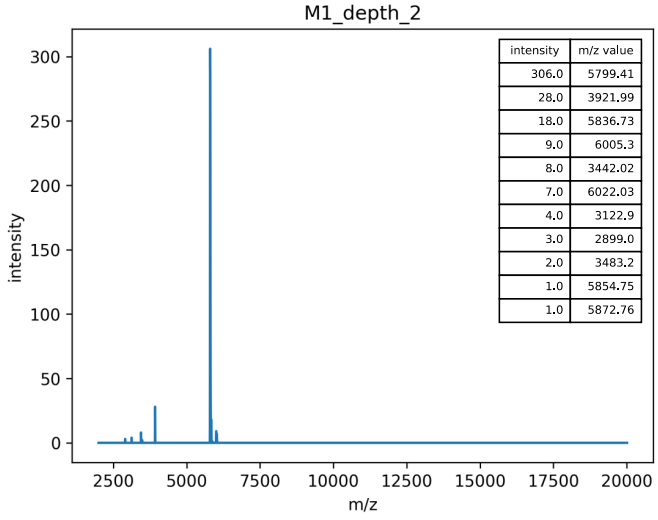
intensity	m/z value
217.0	5804.56
41.0	3923.28
12.0	3442.02
9.0	5844.45
5.0	3122.9
5.0	5825.15
4.0	6013.02
3.0	6029.75
2.0	2902.86
2.0	3483.2



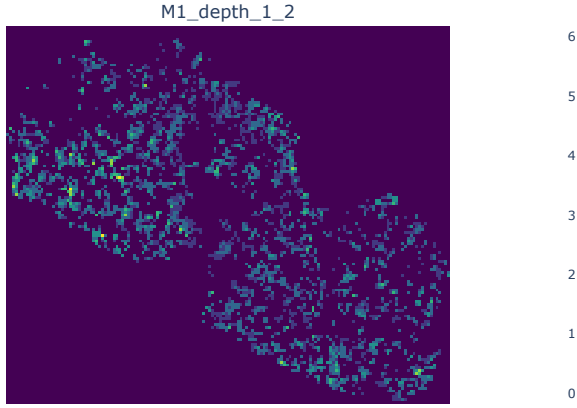


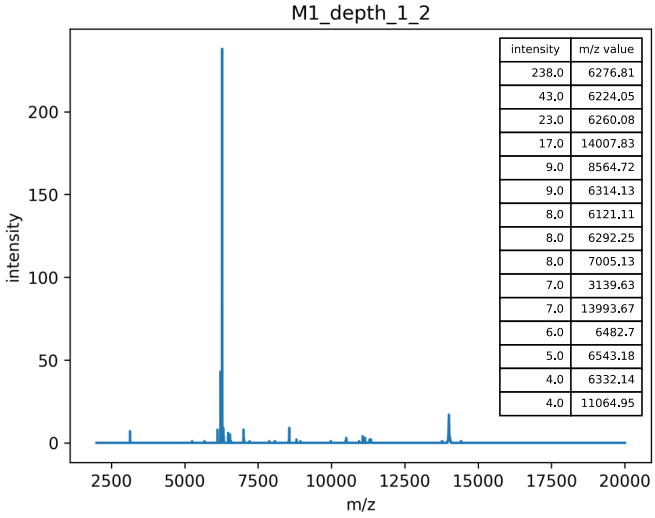
intensity	m/z value
409.0	5798.13
38.0	5816.14
24.0	5836.73
11.0	6004.01
10.0	6022.03
4.0	2900.28
2.0	5856.03
2.0	5874.05
1.0	11600.25



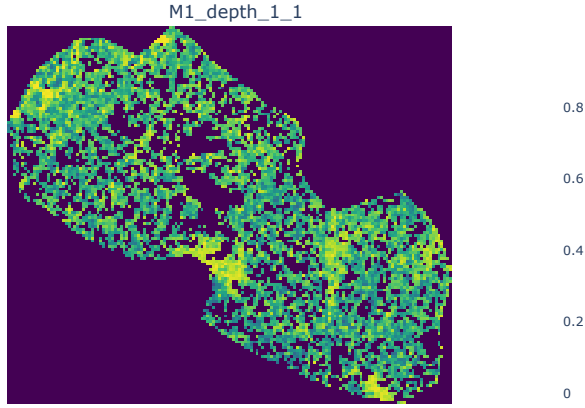


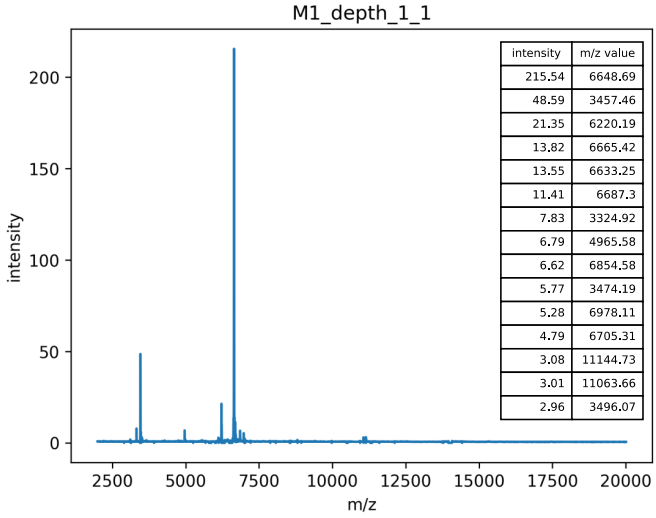
intensity	m/z value
306.0	5799.41
28.0	3921.99
18.0	5836.73
9.0	6005.3
8.0	3442.02
7.0	6022.03
4.0	3122.9
3.0	2899.0
2.0	3483.2
1.0	5854.75
1.0	5872.76



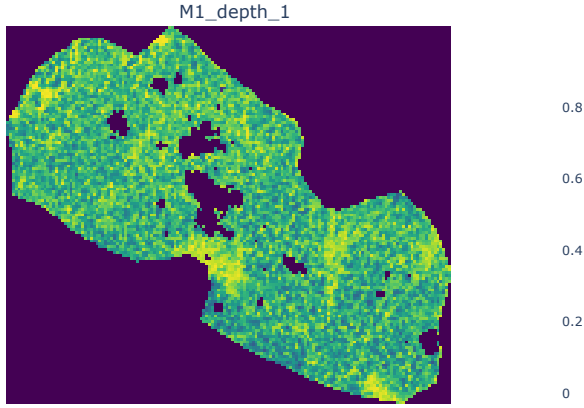


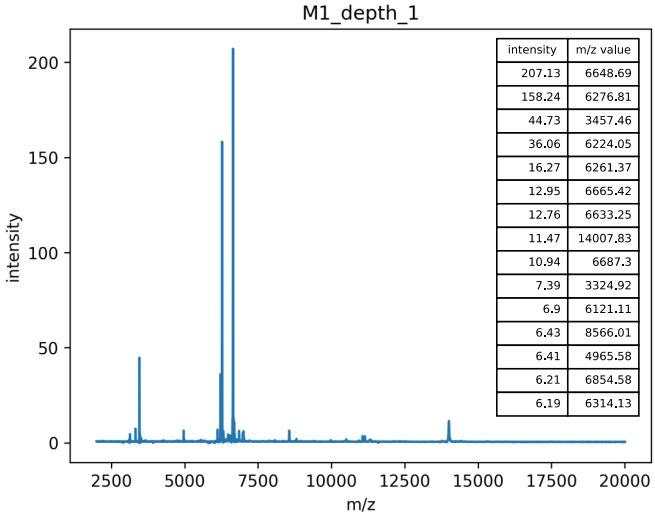
intensity	m/z value
238.0	6276.81
43.0	6224.05
23.0	6260.08
17.0	14007.83
9.0	8564.72
9.0	6314.13
8.0	6121.11
8.0	6292.25
8.0	7005.13
7.0	3139.63
7.0	13993.67
6.0	6482.7
5.0	6543.18
4.0	6332.14
4.0	11064.95





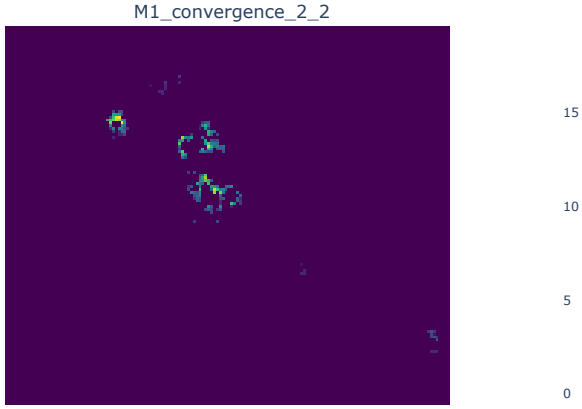
intensity	m/z value
215.54	6648.69
48.59	3457.46
21.35	6220.19
13.82	6665.42
13.55	6633.25
11.41	6687.3
7.83	3324.92
6.79	4965.58
6.62	6854.58
5.77	3474.19
5.28	6978.11
4.79	6705.31
3.08	11144.73
3.01	11063.66
2.96	3496.07

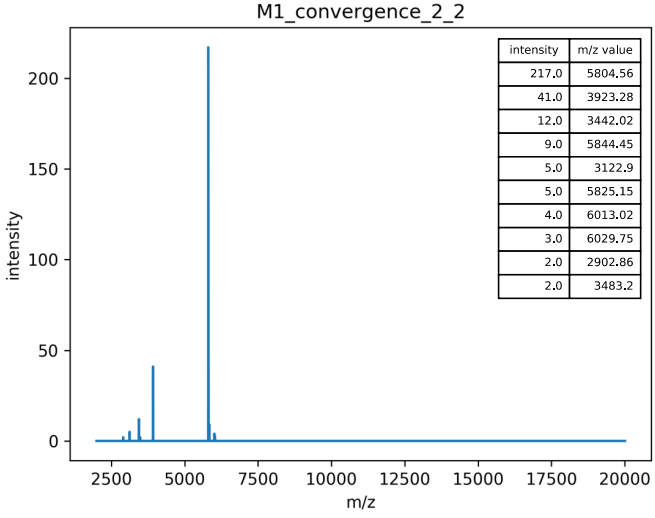




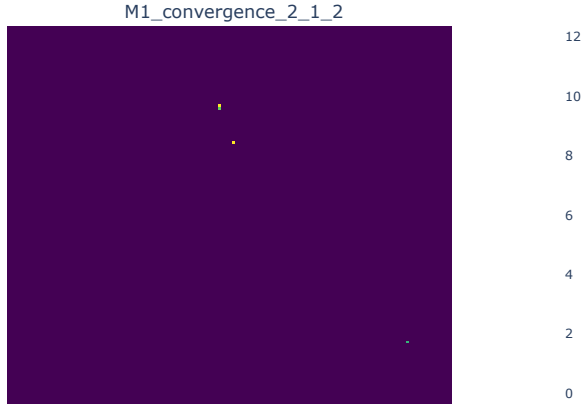
intensity	m/z value
207.13	6648.69
158.24	6276.81
44.73	3457.46
36.06	6224.05
16.27	6261.37
12.95	6665.42
12.76	6633.25
11.47	14007.83
10.94	6687.3
7.39	3324.92
6.9	6121.11
6.43	8566.01
6.41	4965.58
6.21	6854.58
6.19	6314.13

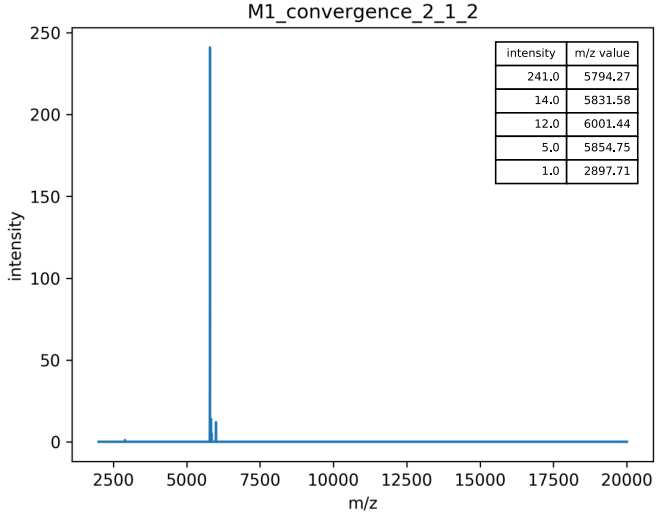
b. rr2-NMF until convergence with threshold = 0.2

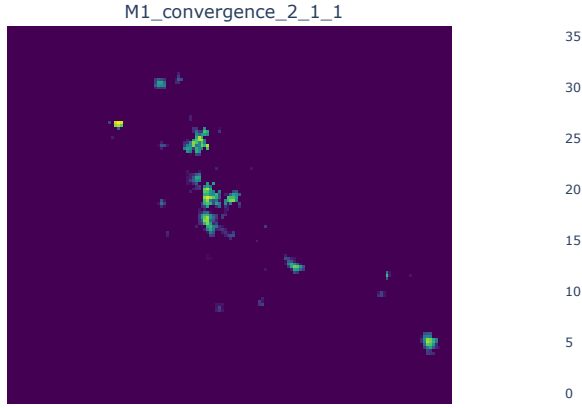


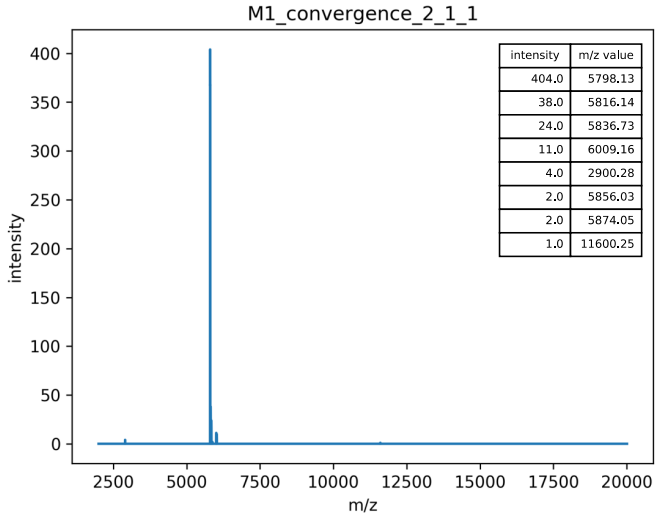


intensity	m/z value
217.0	5804.56
41.0	3923.28
12.0	3442.02
9.0	5844.45
5.0	3122.9
5.0	5825.15
4.0	6013.02
3.0	6029.75
2.0	2902.86
2.0	3483.2

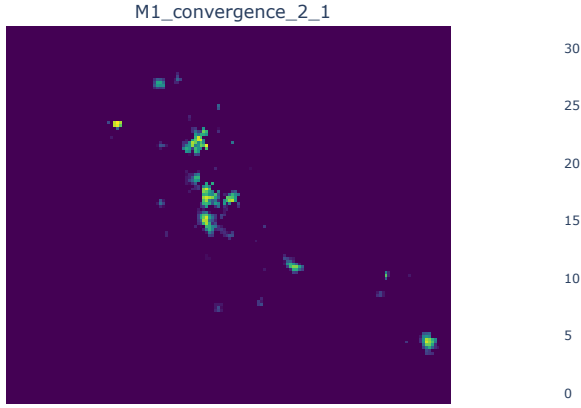

M1_convergence_2_1_2
12
10
8
6
4
2
0

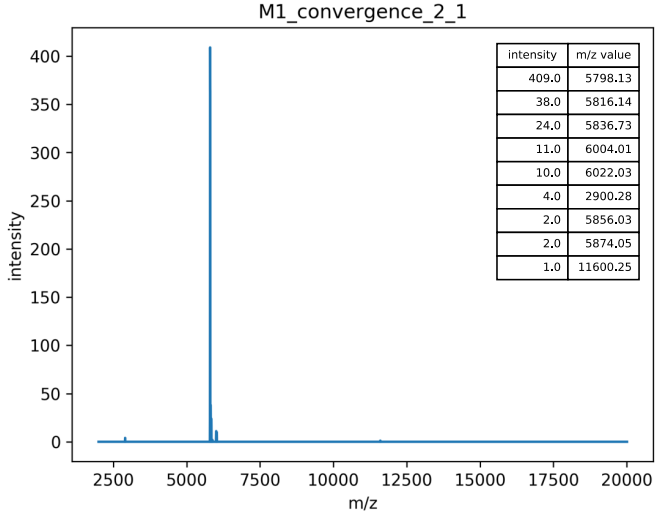

M1_convergence_2_1_2
intensity
m/z value
241.0 5794.27
14.0 5831.58
12.0 6001.44
5.0 5854.75
1.0 2897.71
intensity
m/z
250
200
150
100
50
0
2500
5000
7500
10000
12500
15000
17500
20000



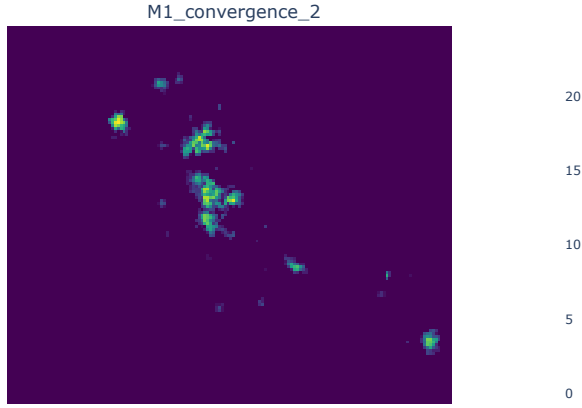


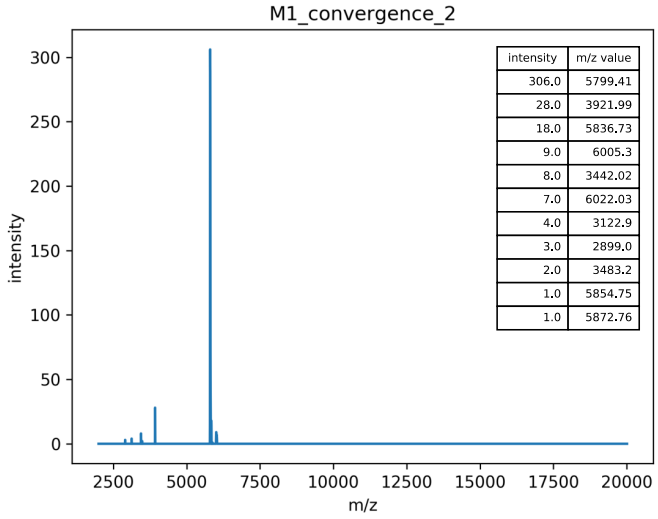
intensity	m/z value
404.0	5798.13
38.0	5816.14
24.0	5836.73
11.0	6009.16
4.0	2900.28
2.0	5856.03
2.0	5874.05
1.0	11600.25



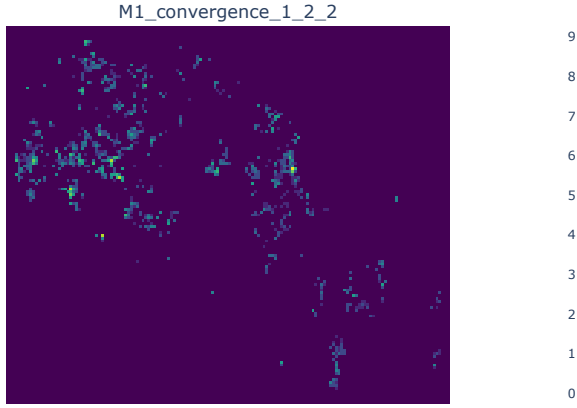


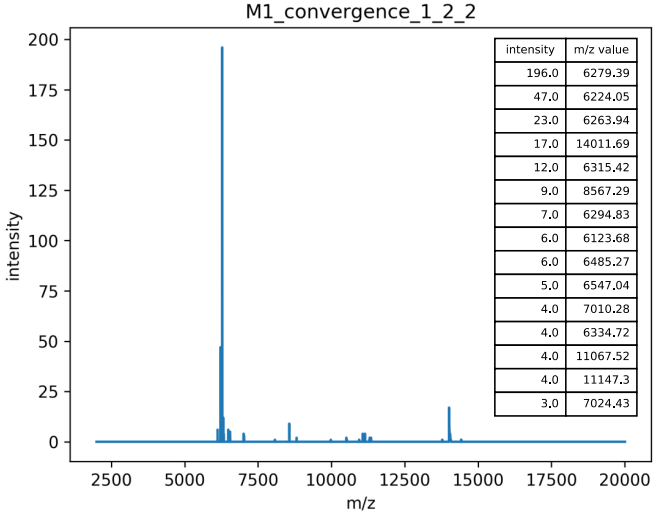
intensity	m/z value
409.0	5798.13
38.0	5816.14
24.0	5836.73
11.0	6004.01
10.0	6022.03
4.0	2900.28
2.0	5856.03
2.0	5874.05
1.0	11600.25



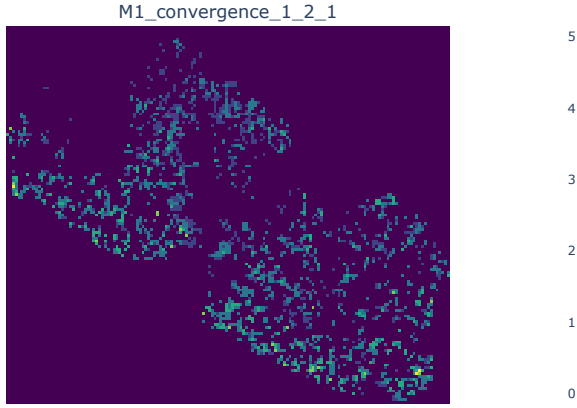


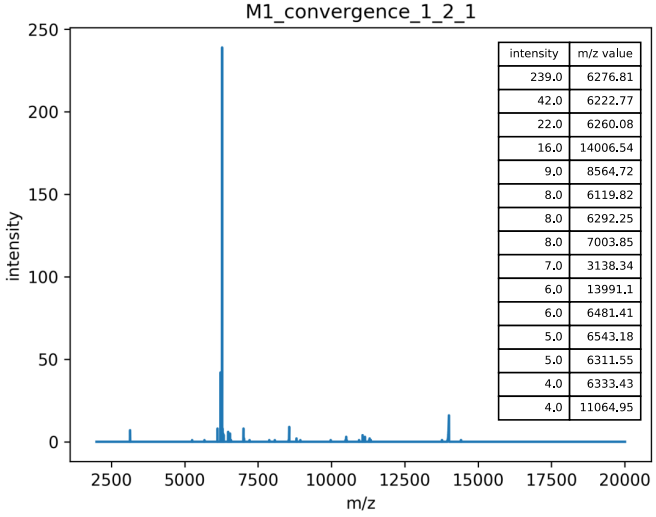
intensity	m/z value
306.0	5799.41
28.0	3921.99
18.0	5836.73
9.0	6005.3
8.0	3442.02
7.0	6022.03
4.0	3122.9
3.0	2899.0
2.0	3483.2
1.0	5854.75
1.0	5872.76



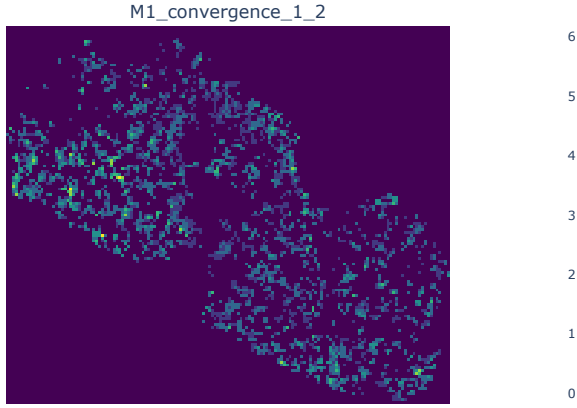


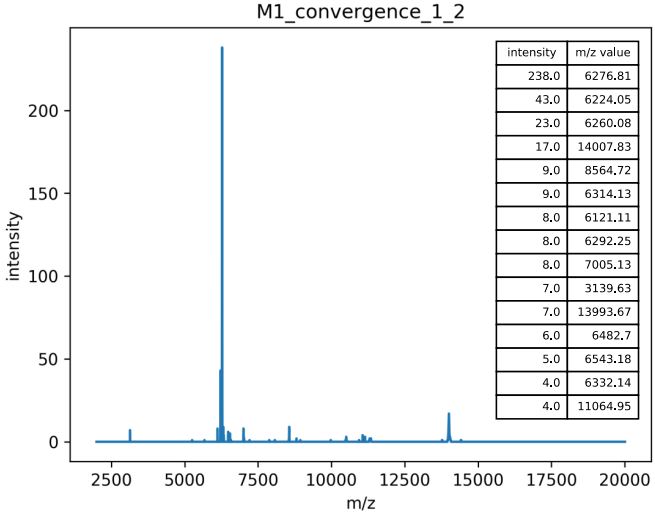
intensity	m/z value
196.0	6279.39
47.0	6224.05
23.0	6263.94
17.0	14011.69
12.0	6315.42
9.0	8567.29
7.0	6294.83
6.0	6123.68
6.0	6485.27
5.0	6547.04
4.0	7010.28
4.0	6334.72
4.0	11067.52
4.0	11147.3
3.0	7024.43



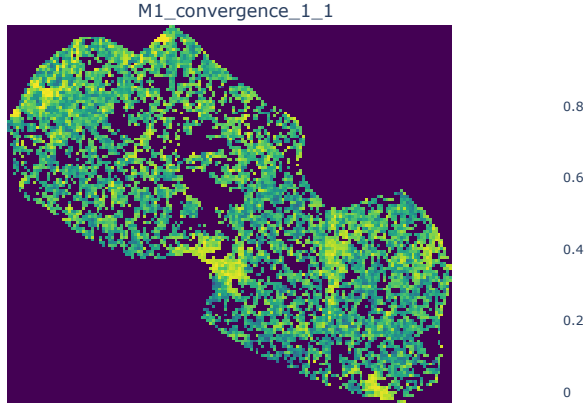


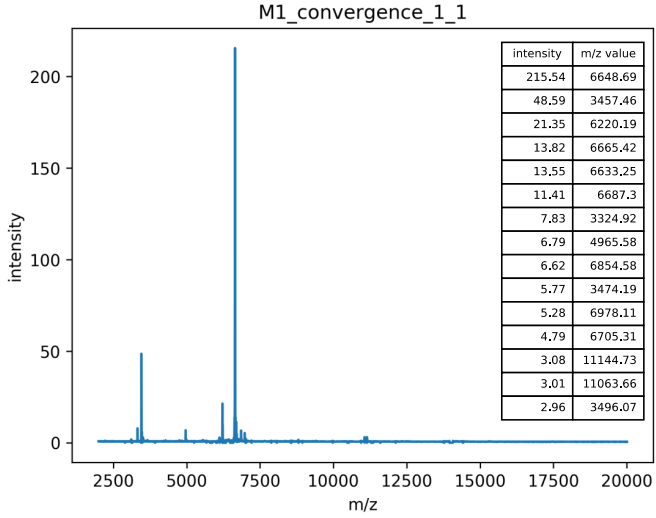
intensity	m/z value
239.0	6276.81
42.0	6222.77
22.0	6260.08
16.0	14006.54
9.0	8564.72
8.0	6119.82
8.0	6292.25
8.0	7003.85
7.0	3138.34
6.0	13991.1
6.0	6481.41
5.0	6543.18
5.0	6311.55
4.0	6333.43
4.0	11064.95



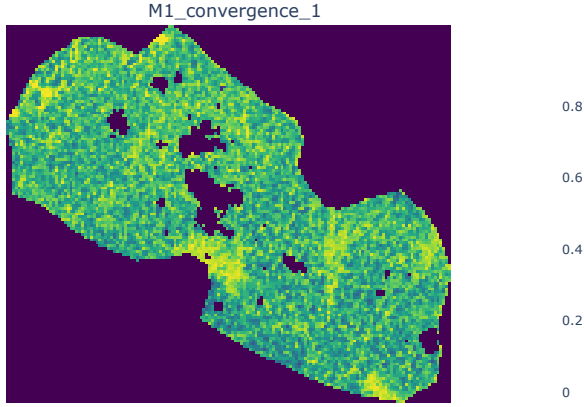


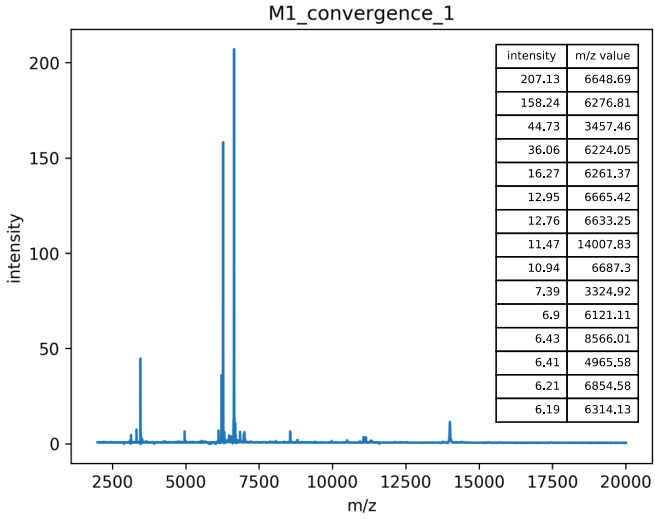
intensity	m/z value
238.0	6276.81
43.0	6224.05
23.0	6260.08
17.0	14007.83
9.0	8564.72
9.0	6314.13
8.0	6121.11
8.0	6292.25
8.0	7005.13
7.0	3139.63
7.0	13993.67
6.0	6482.7
5.0	6543.18
4.0	6332.14
4.0	11064.95





intensity	m/z value
215.54	6648.69
48.59	3457.46
21.35	6220.19
13.82	6665.42
13.55	6633.25
11.41	6687.3
7.83	3324.92
6.79	4965.58
6.62	6854.58
5.77	3474.19
5.28	6978.11
4.79	6705.31
3.08	11144.73
3.01	11063.66
2.96	3496.07





intensity	m/z value
207.13	6648.69
158.24	6276.81
44.73	3457.46
36.06	6224.05
16.27	6261.37
12.95	6665.42
12.76	6633.25
11.47	14007.83
10.94	6687.3
7.39	3324.92
6.9	6121.11
6.43	8566.01
6.41	4965.58
6.21	6854.58
6.19	6314.13

c. ion images of m/z-values of interest in the Islets of Langerhans

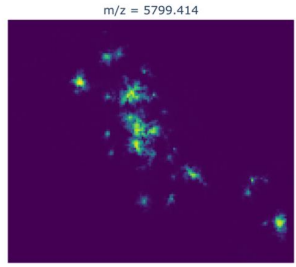


Figure S 5. Ion image of Ins2 for M1.

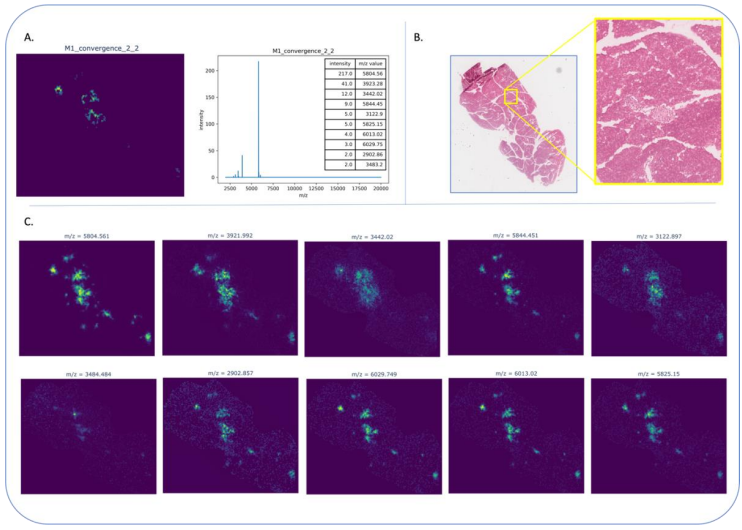
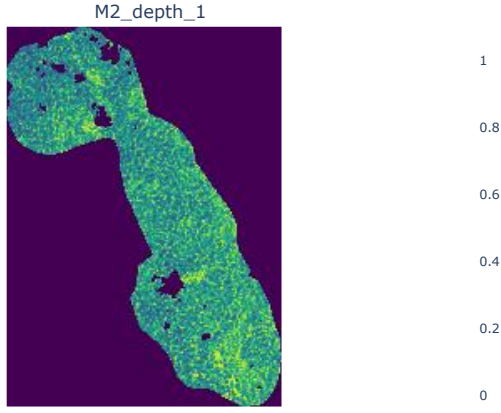
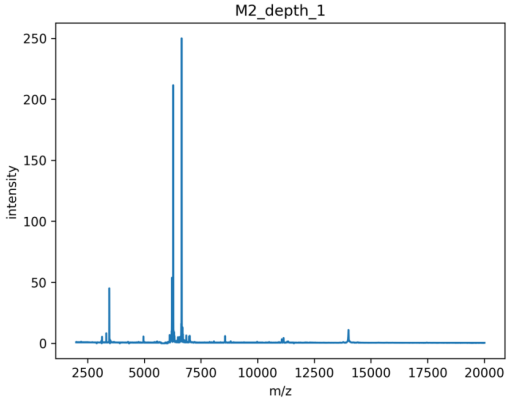


Figure S 6. Spatial correspondence of cluster "M1_convergence_2_2" and the ion images of the corresponding m/z-peaks. The cluster shows also for this sample the relation between Ins1 (m/z = 5804.56 Da and m/z = 2902.86 Da) and the outer periphery of the islets of Langerhans in mice. (A) Result cluster obtained with rr2-NMF. (B) H&E staining of the M1 sample. (C) Ion images of the peak values corresponding to this cluster.

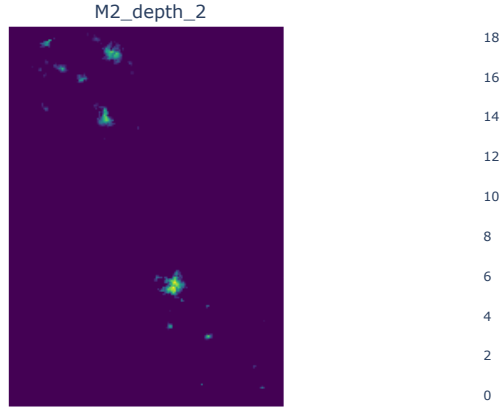
S4. Pancreas M2 sample analysis using rr2-NMF

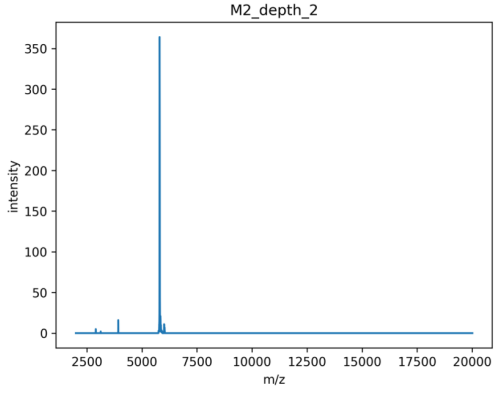
a. rr2-NMF until depth = 2



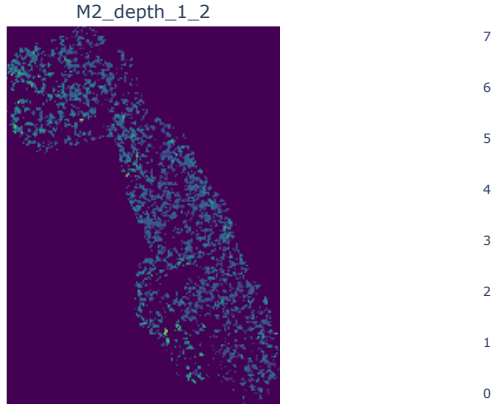


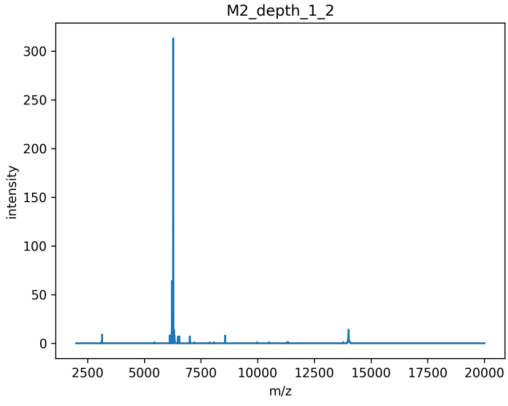
intensity	m/z value
249.96	6650.11
211.54	6278.21
53.78	6224.16
45.06	3458.66
19.27	6262.76
15.92	6633.38
13.11	6688.72
11.98	6668.13
10.95	14011.03
10.45	6293.65
9.94	6210.0
9.26	6316.81
8.23	3326.12
7.49	6706.73
6.8	6122.49



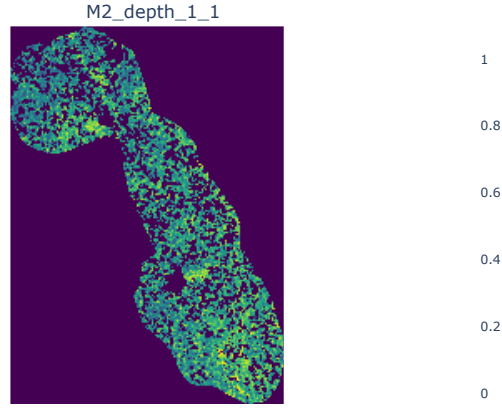


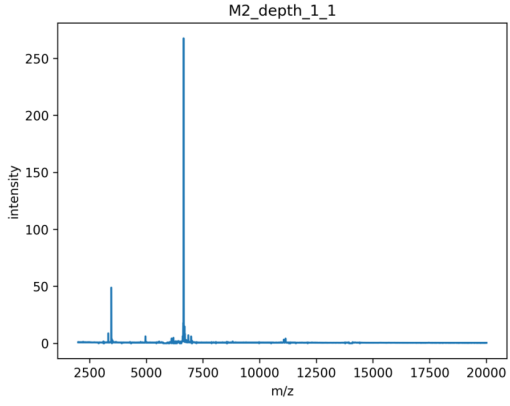
intensity	m/z value
364.0	5800.77
21.0	5839.38
16.0	3923.23
11.0	6006.67
10.0	5782.76
10.0	5858.68
7.0	6023.4
5.0	2901.45
4.0	5875.41
3.0	5757.02
3.0	5896.0
2.0	3122.79
2.0	5915.31
1.0	5977.08



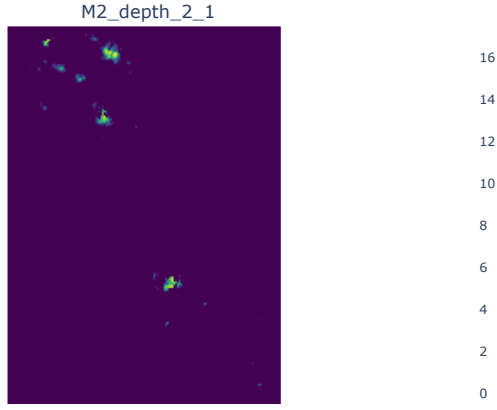


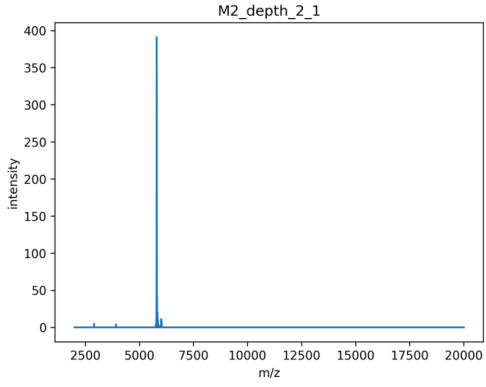
intensity	m/z value
313.0	6278.21
64.0	6224.16
26.0	6261.48
15.0	6293.65
14.0	14009.75
13.0	6315.52
11.0	6208.71
9.0	3139.52
8.0	8567.55
8.0	6122.49
8.0	6333.54
7.0	7006.58
7.0	6545.87
7.0	6484.11
5.0	13993.02



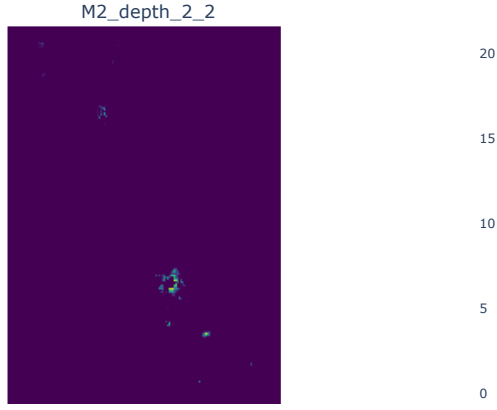


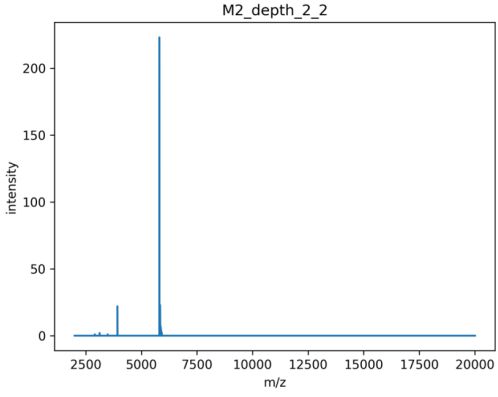
intensity	m/z value
267.62	6650.11
48.86	3458.66
17.8	6633.38
14.83	6688.72
13.45	6668.13
8.74	3326.12
8.36	6706.73
7.06	6857.3
6.25	6608.93
6.08	4966.88
6.01	6980.84
4.8	6199.71
4.23	6119.92
4.14	11147.74
3.84	3475.39





intensity	m/z value
391.0	5800.77
43.0	5814.93
20.0	5838.09
11.0	5782.76
11.0	6005.39
10.0	5857.4
8.0	6023.4
5.0	2901.45
4.0	3920.65
3.0	5755.73
3.0	5896.0
1.0	5910.16
1.0	5977.08

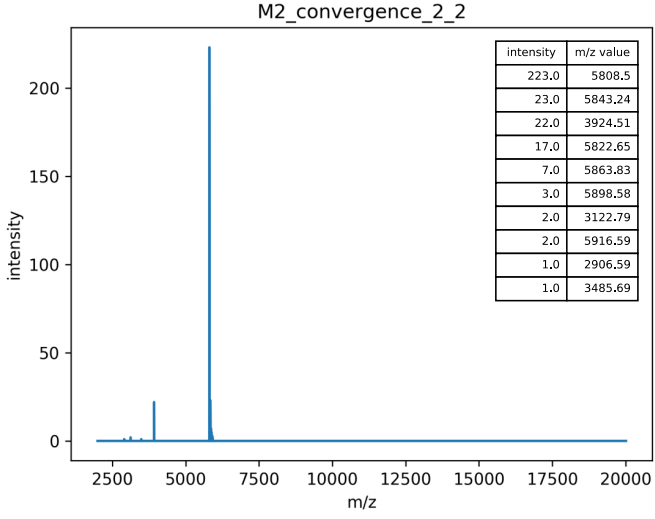




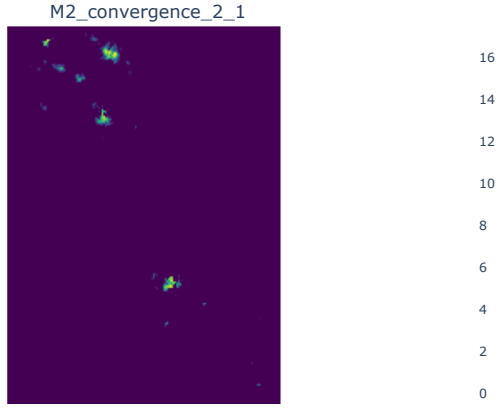
intensity	m/z value
223.0	5808.5
23.0	5843.24
22.0	3924.51
17.0	5822.65
7.0	5863.83
3.0	5898.58
2.0	3122.79
2.0	5916.59
1.0	2906.59
1.0	3485.69

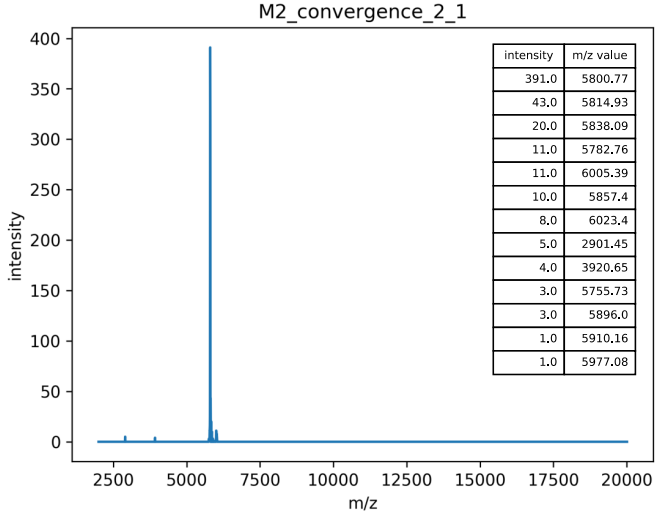
b. rr2-NMF until convergence with threshold = 0.2



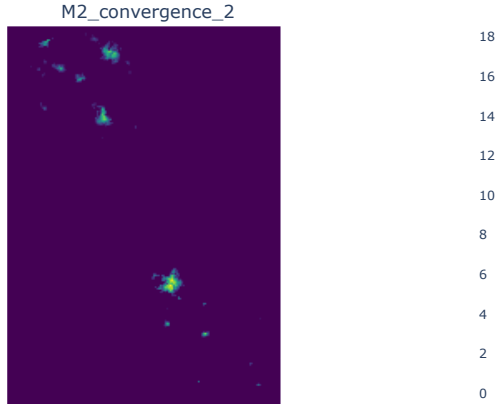


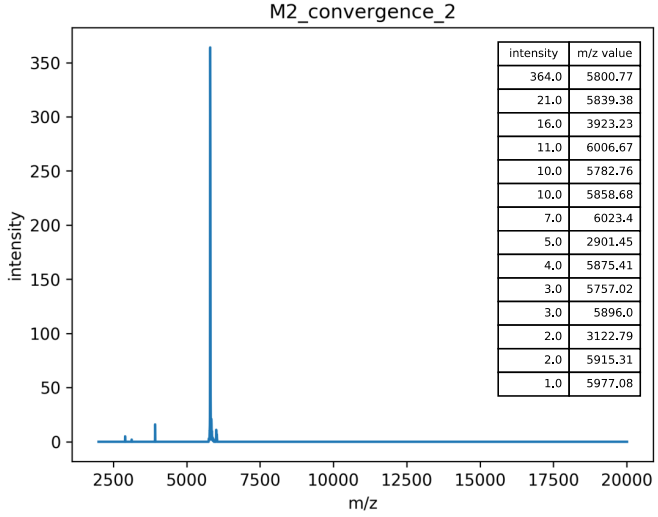
intensity	m/z value
223.0	5808.5
23.0	5843.24
22.0	3924.51
17.0	5822.65
7.0	5863.83
3.0	5898.58
2.0	3122.79
2.0	5916.59
1.0	2906.59
1.0	3485.69



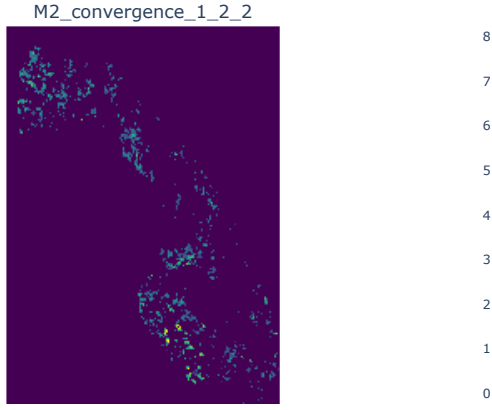


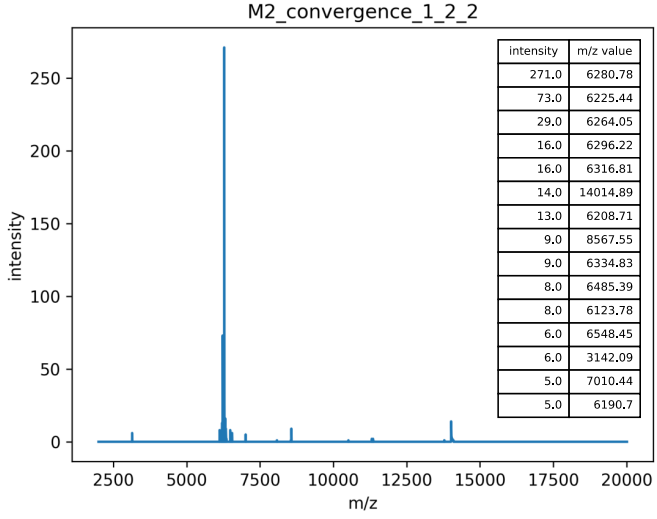
intensity	m/z value
391.0	5800.77
43.0	5814.93
20.0	5838.09
11.0	5782.76
11.0	6005.39
10.0	5857.4
8.0	6023.4
5.0	2901.45
4.0	3920.65
3.0	5755.73
3.0	5896.0
1.0	5910.16
1.0	5977.08



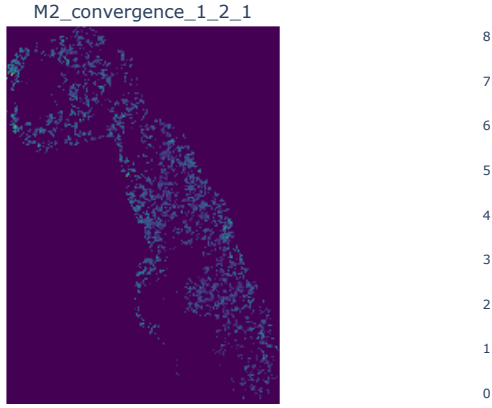


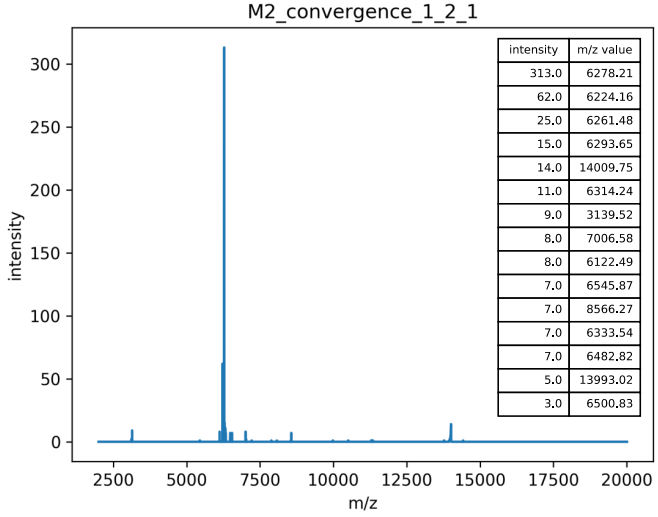
intensity	m/z value
364.0	5800.77
21.0	5839.38
16.0	3923.23
11.0	6006.67
10.0	5782.76
10.0	5858.68
7.0	6023.4
5.0	2901.45
4.0	5875.41
3.0	5757.02
3.0	5896.0
2.0	3122.79
2.0	5915.31
1.0	5977.08



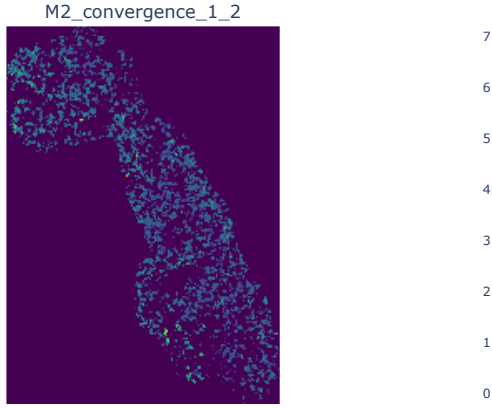


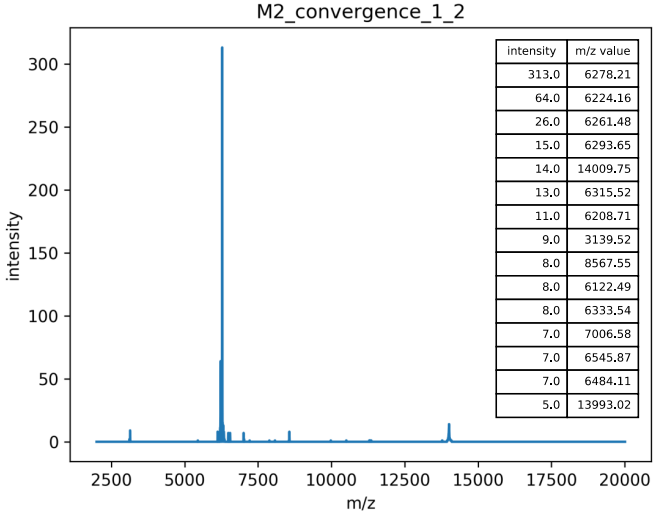
intensity	m/z value
271.0	6280.78
73.0	6225.44
29.0	6264.05
16.0	6296.22
16.0	6316.81
14.0	14014.89
13.0	6208.71
9.0	8567.55
9.0	6334.83
8.0	6485.39
8.0	6123.78
6.0	6548.45
6.0	3142.09
5.0	7010.44
5.0	6190.7



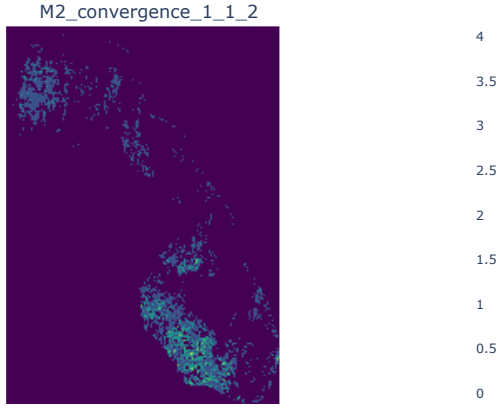


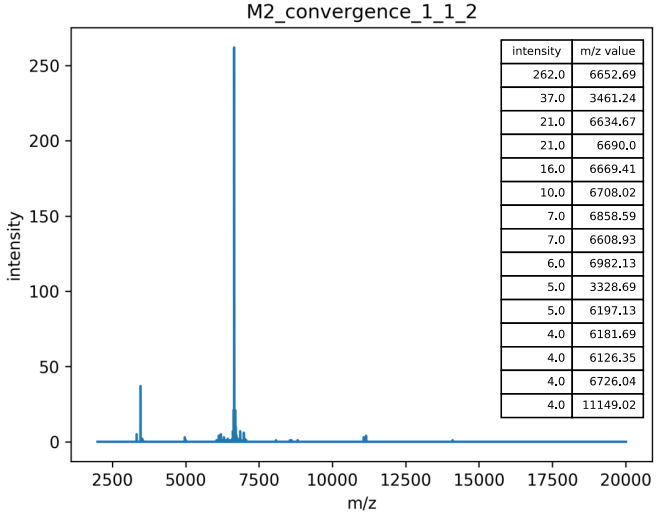
intensity	m/z value
313.0	6278.21
62.0	6224.16
25.0	6261.48
15.0	6293.65
14.0	14009.75
11.0	6314.24
9.0	3139.52
8.0	7006.58
8.0	6122.49
7.0	6545.87
7.0	8566.27
7.0	6333.54
7.0	6482.82
5.0	13993.02
3.0	6500.83



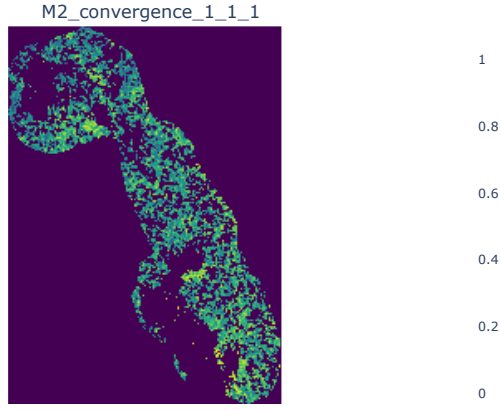


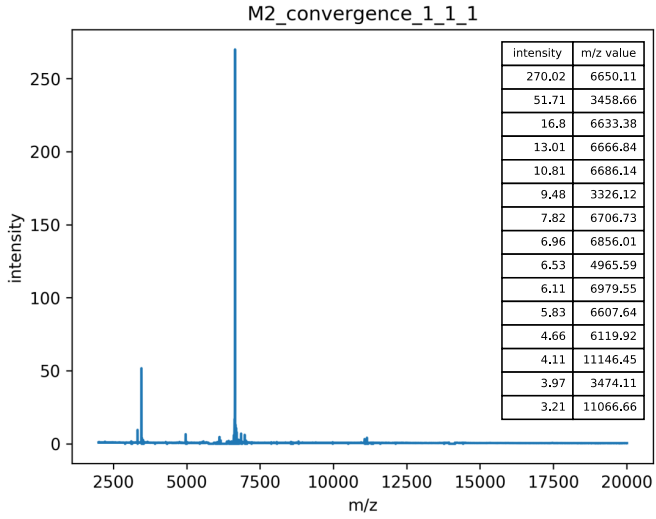
intensity	m/z value
313.0	6278.21
64.0	6224.16
26.0	6261.48
15.0	6293.65
14.0	14009.75
13.0	6315.52
11.0	6208.71
9.0	3139.52
8.0	8567.55
8.0	6122.49
8.0	6333.54
7.0	7006.58
7.0	6545.87
7.0	6484.11
5.0	13993.02



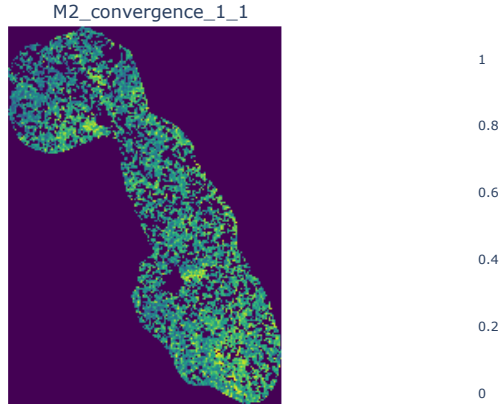


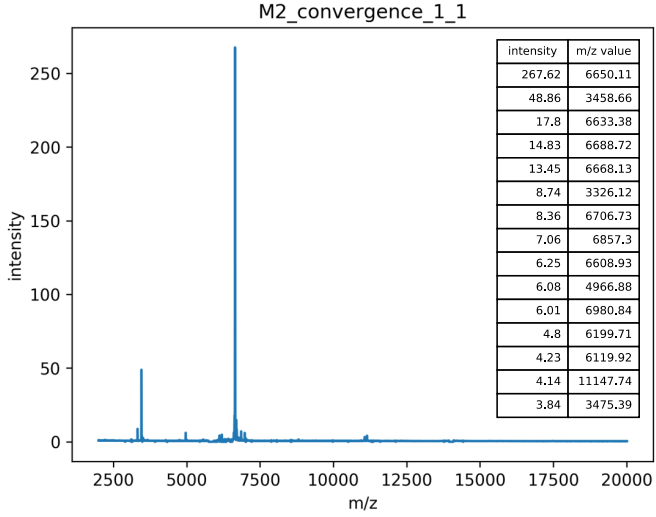
intensity	m/z value
262.0	6652.69
37.0	3461.24
21.0	6634.67
21.0	6690.0
16.0	6669.41
10.0	6708.02
7.0	6858.59
7.0	6608.93
6.0	6982.13
5.0	3328.69
5.0	6197.13
4.0	6181.69
4.0	6126.35
4.0	6726.04
4.0	11149.02



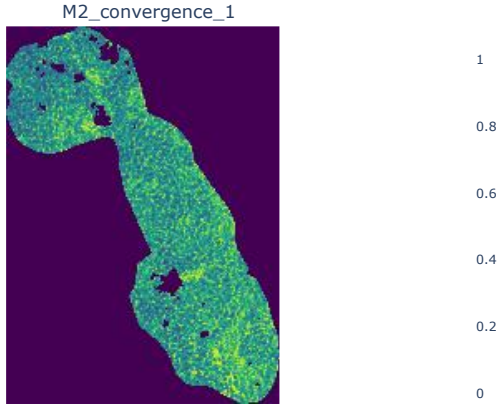


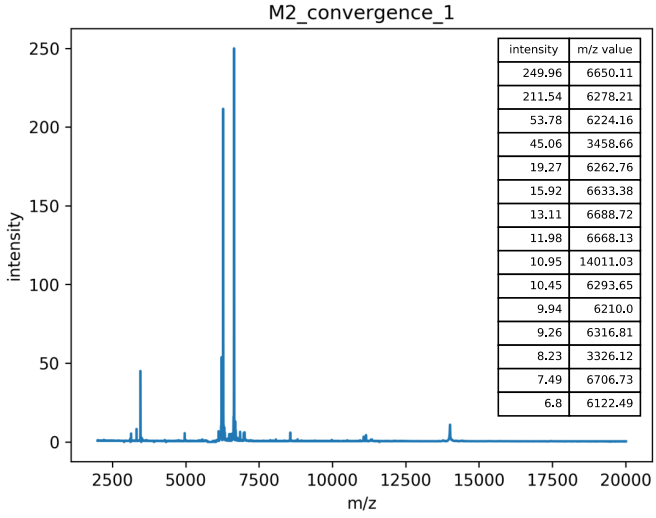
intensity	m/z value
270.02	6650.11
51.71	3458.66
16.8	6633.38
13.01	6666.84
10.81	6686.14
9.48	3326.12
7.82	6706.73
6.96	6856.01
6.53	4965.59
6.11	6979.55
5.83	6607.64
4.66	6119.92
4.11	11146.45
3.97	3474.11
3.21	11066.66





intensity	m/z value
267.62	6650.11
48.86	3458.66
17.8	6633.38
14.83	6688.72
13.45	6668.13
8.74	3326.12
8.36	6706.73
7.06	6857.3
6.25	6608.93
6.08	4966.88
6.01	6980.84
4.8	6199.71
4.23	6119.92
4.14	11147.74
3.84	3475.39





intensity	m/z value
249.96	6650.11
211.54	6278.21
53.78	6224.16
45.06	3458.66
19.27	6262.76
15.92	6633.38
13.11	6688.72
11.98	6668.13
10.95	14011.03
10.45	6293.65
9.94	6210.0
9.26	6316.81
8.23	3326.12
7.49	6706.73
6.8	6122.49

c. ion images of m/z-values of interest in the Islets of Langerhans

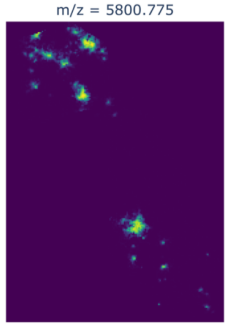


Figure S 7. Ion image of Ins2 for M2.

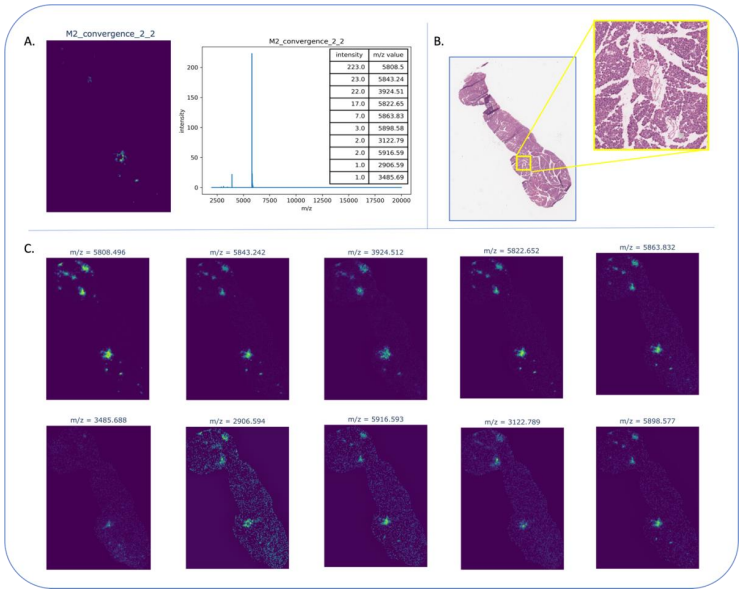
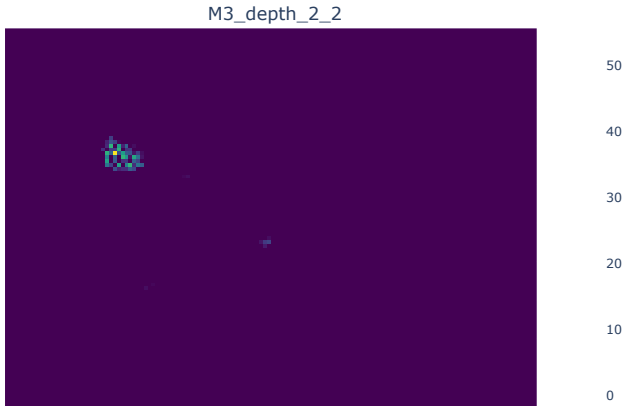
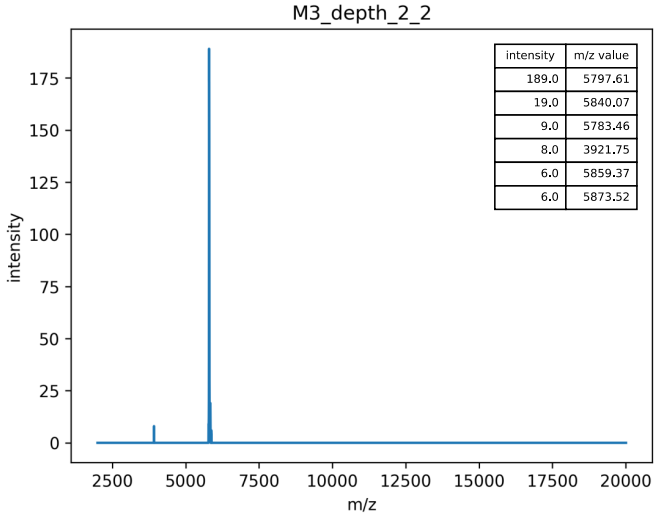


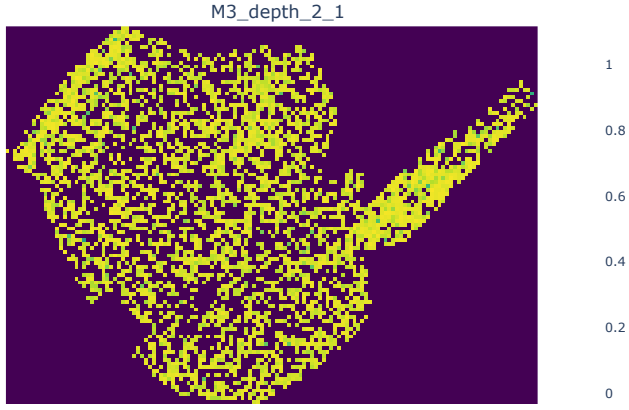
Figure S 8. Spatial correspondence of cluster "M2_convergence_2_2" and the ion images of the corresponding m/z-peaks. The cluster shows relation between Ins1 (m/z = 5808.5 Da and m/z = 2906.59 Da) and the outer periphery of the islets of Langerhans in mice. (A) Result cluster obtained with rr2-NMF. (B) H&E staining of the M2 sample. (C) Ion images of the peak values corresponding to this cluster.

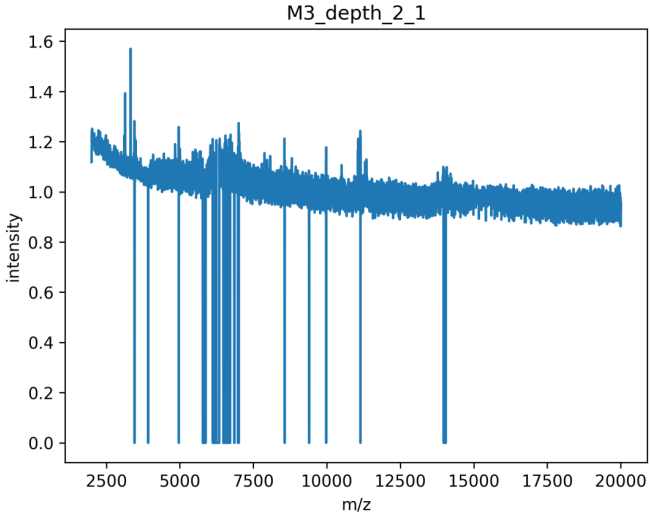
S5. Pancreas M3 sample analysis using rr2-NMF

a. rr2-NMF until depth = 2

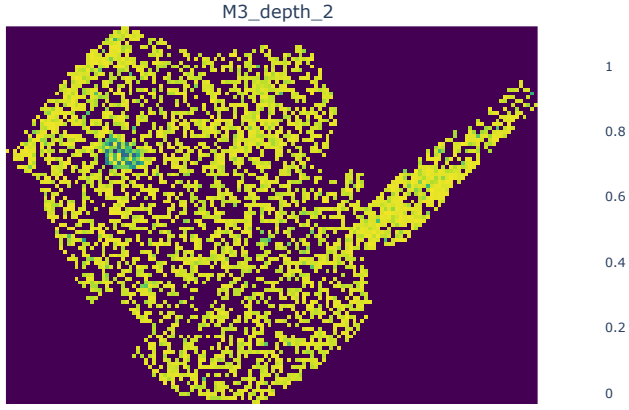


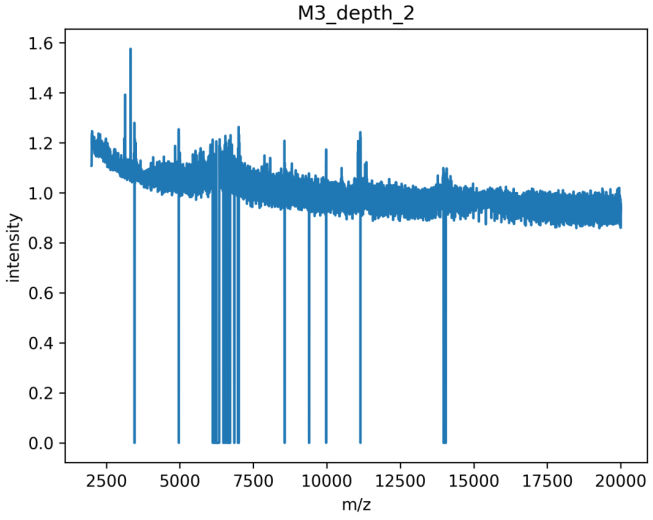




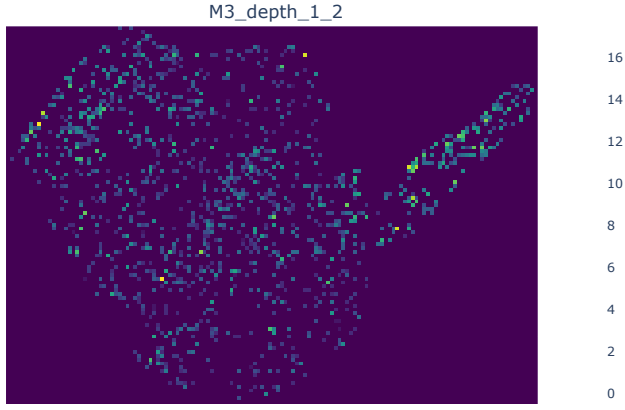


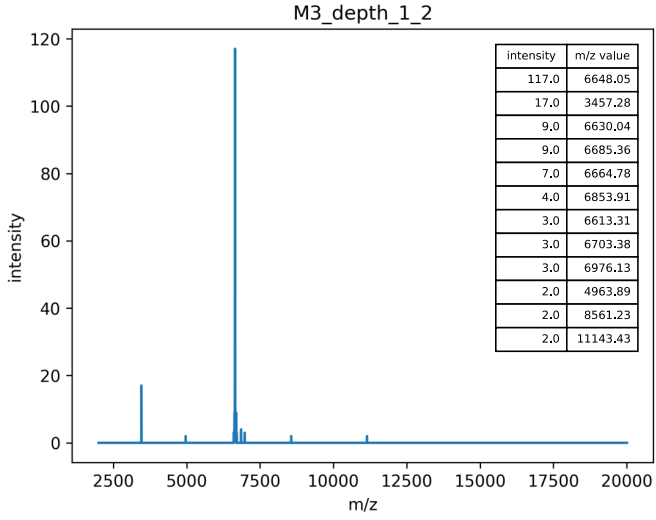
intensity	m/z value
1.57	3324.77
1.39	3138.21
1.28	3453.42
1.27	7001.87
1.26	4962.61
1.25	2012.43
1.24	2233.73
1.24	11142.14
1.24	2286.48
1.23	2036.88
1.23	2080.62
1.23	6725.25
1.22	2484.62
1.22	3111.19
1.22	6550.27



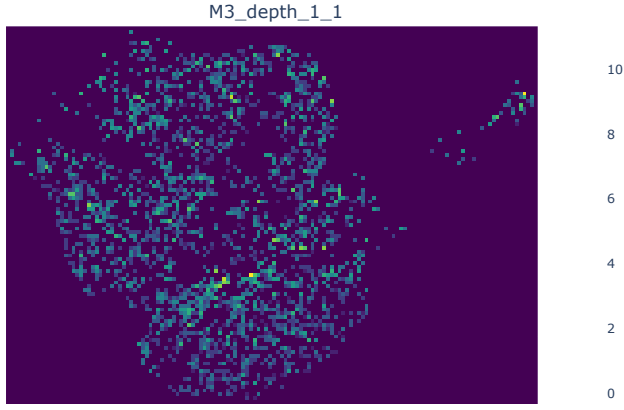


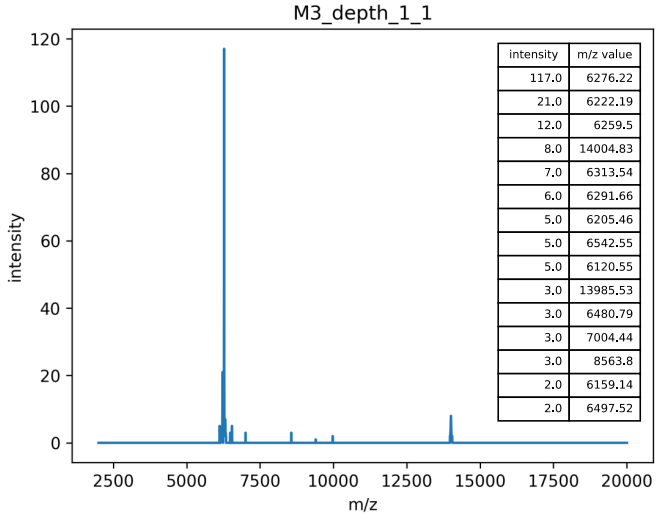
intensity	m/z value
1.58	3324.77
1.39	3139.49
1.28	3453.42
1.26	7001.87
1.25	4962.61
1.25	2012.43
1.24	11142.14
1.24	2233.73
1.23	2286.48
1.23	6725.25
1.23	6550.27
1.23	2036.88
1.22	2071.62
1.22	2085.77
1.22	3111.19



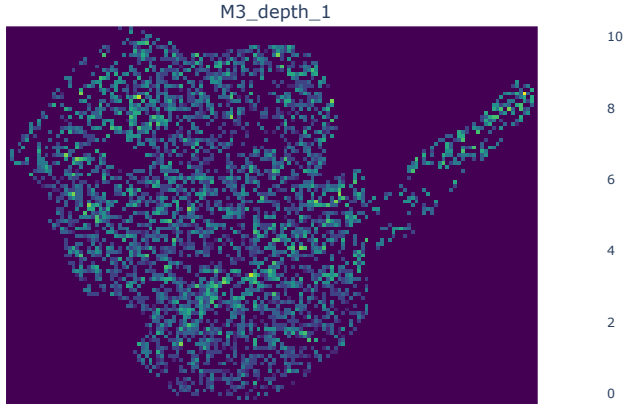


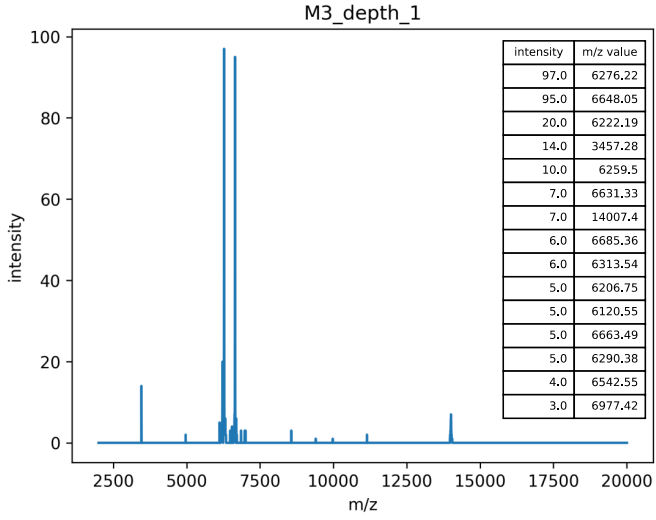
intensity	m/z value
117.0	6648.05
17.0	3457.28
9.0	6630.04
9.0	6685.36
7.0	6664.78
4.0	6853.91
3.0	6613.31
3.0	6703.38
3.0	6976.13
2.0	4963.89
2.0	8561.23
2.0	11143.43





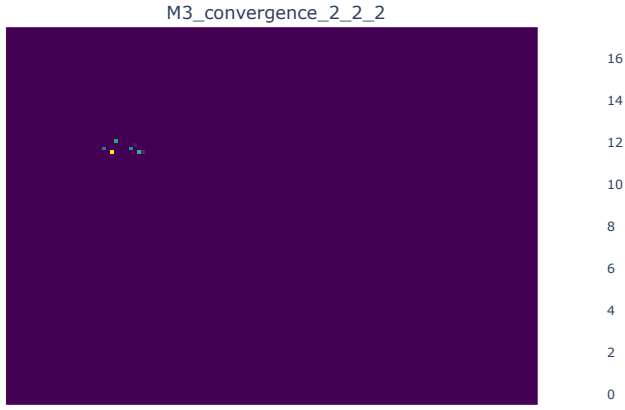
intensity	m/z value
117.0	6276.22
21.0	6222.19
12.0	6259.5
8.0	14004.83
7.0	6313.54
6.0	6291.66
5.0	6205.46
5.0	6542.55
5.0	6120.55
3.0	13985.53
3.0	6480.79
3.0	7004.44
3.0	8563.8
2.0	6159.14
2.0	6497.52

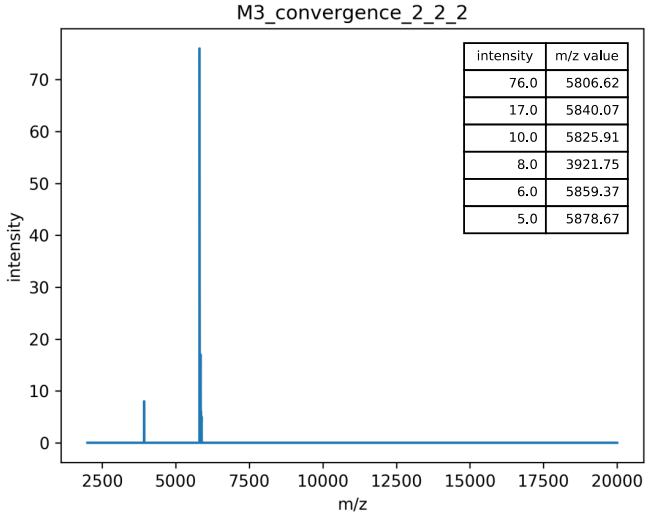




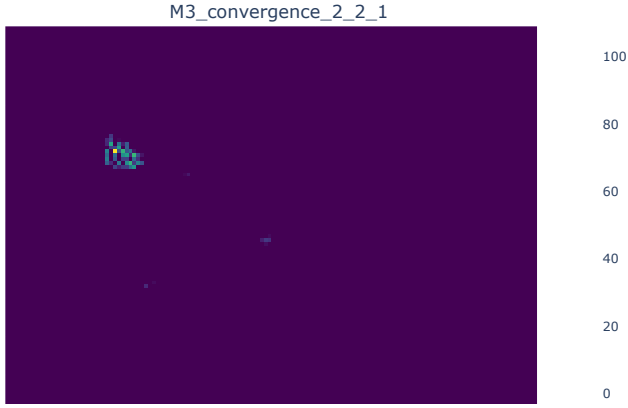
intensity	m/z value
97.0	6276.22
95.0	6648.05
20.0	6222.19
14.0	3457.28
10.0	6259.5
7.0	6631.33
7.0	14007.4
6.0	6685.36
6.0	6313.54
5.0	6206.75
5.0	6120.55
5.0	6663.49
5.0	6290.38
4.0	6542.55
3.0	6977.42

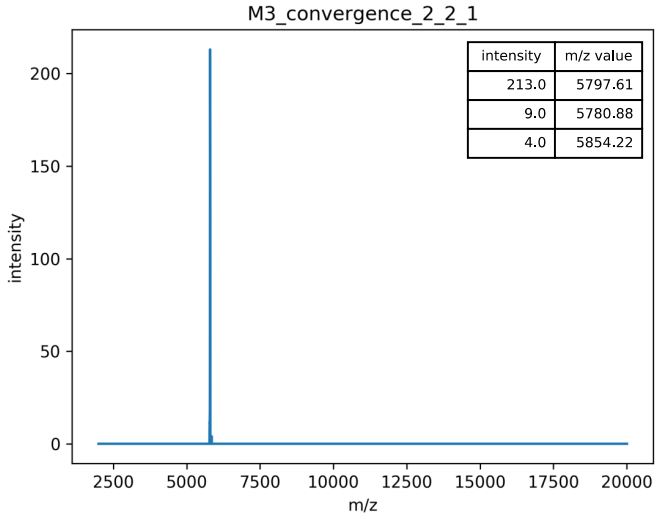
b. rr2-NMF until convergence with threshold = 0.2

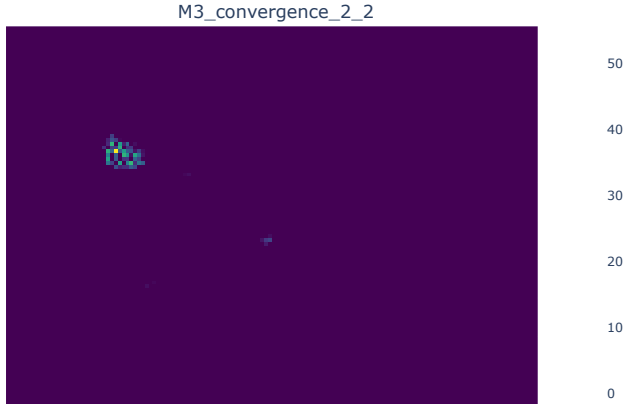


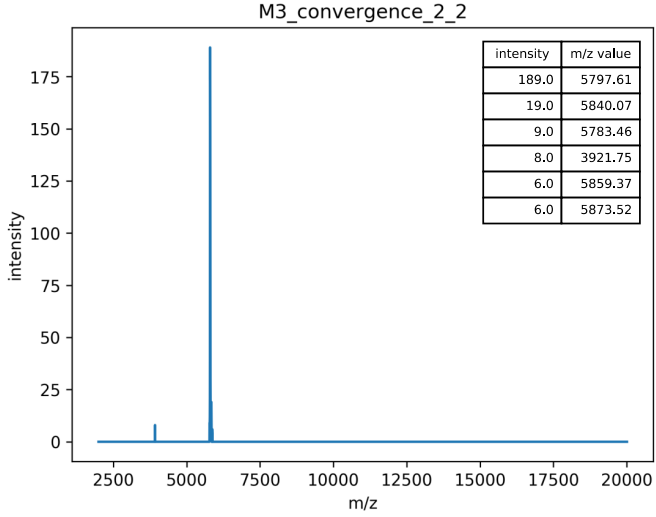


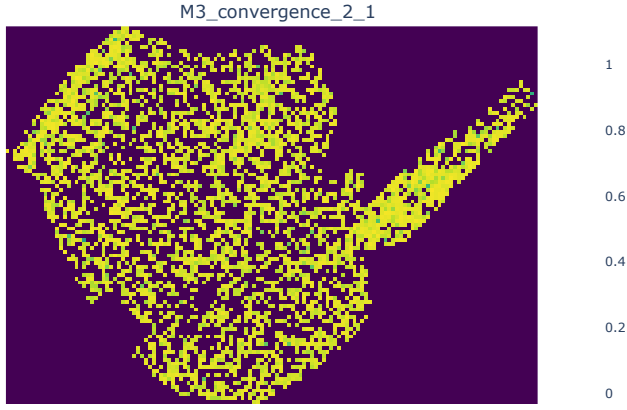
intensity	m/z value
76.0	5806.62
17.0	5840.07
10.0	5825.91
8.0	3921.75
6.0	5859.37
5.0	5878.67

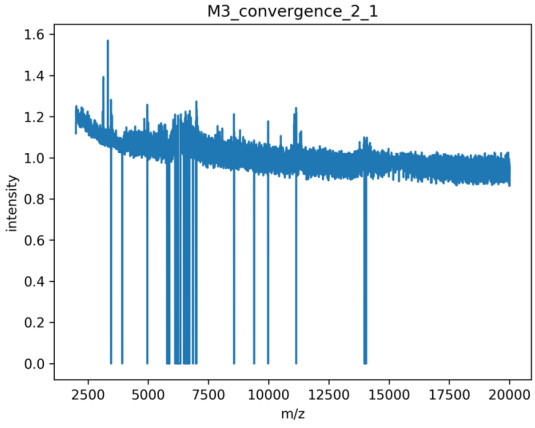

M3_convergence_2_2_1
100
80
60
40
20
0


M3_convergence_2_2_1
intensity
m/z value
213.0
5797.61
9.0
5780.88
4.0
5854.22
200
150
100
50
0
intensity
2500
5000
7500
10000
12500
15000
17500
20000
m/z

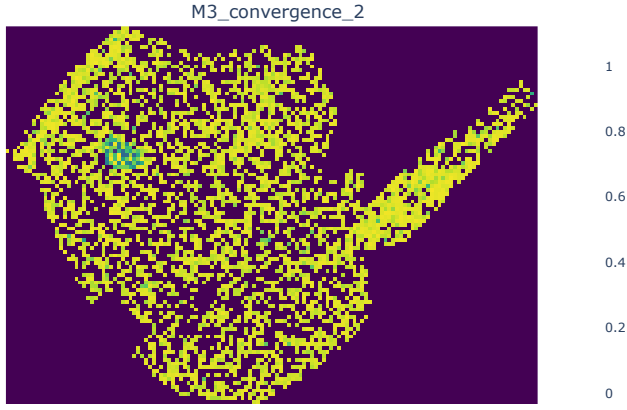

M3_convergence_2_2
50
40
30
20
10
0

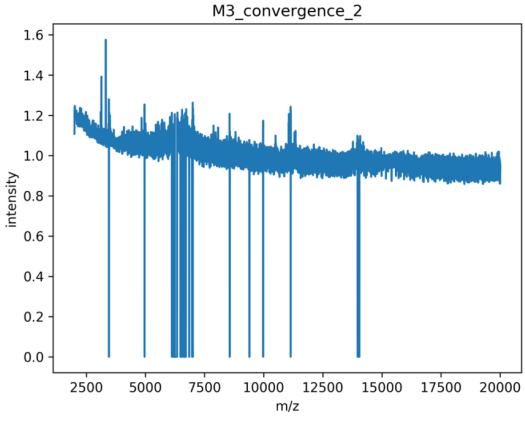

M3_convergence_2_2
intensity
m/z
intensity
m/z value
189.0
5797.61
19.0
5840.07
9.0
5783.46
8.0
3921.75
6.0
5859.37
6.0
5873.52



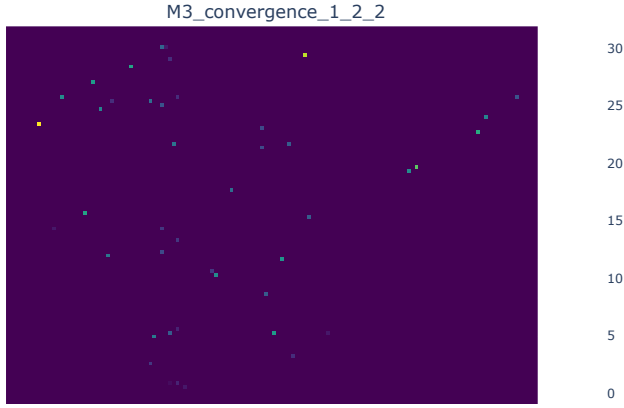


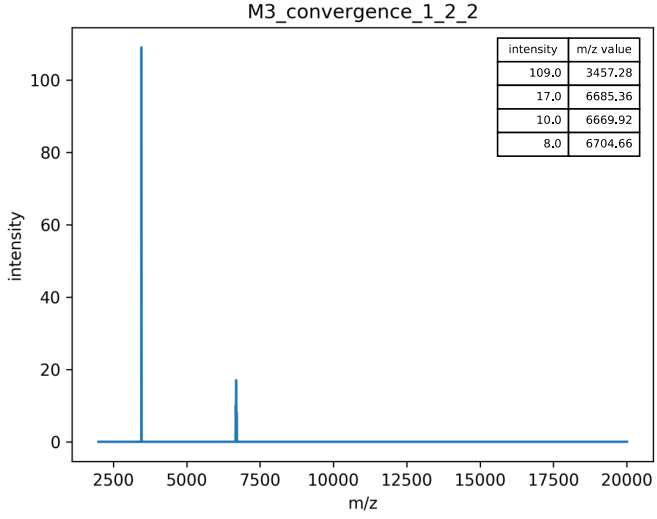
intensity	m/z value
1.57	3324.77
1.39	3138.21
1.28	3453.42
1.27	7001.87
1.26	4962.61
1.25	2012.43
1.24	2233.73
1.24	11142.14
1.24	2286.48
1.23	2036.88
1.23	2080.62
1.23	6725.25
1.22	2484.62
1.22	3111.19
1.22	6550.27

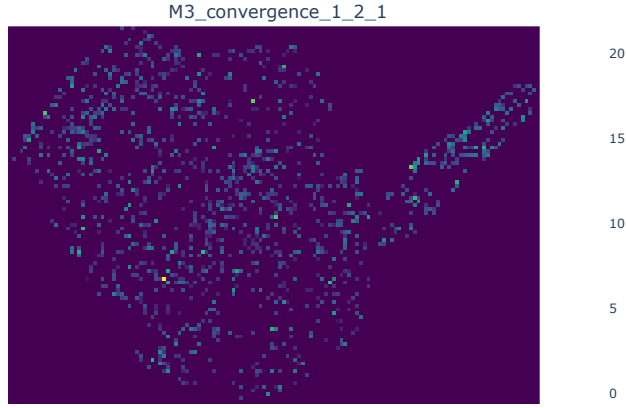


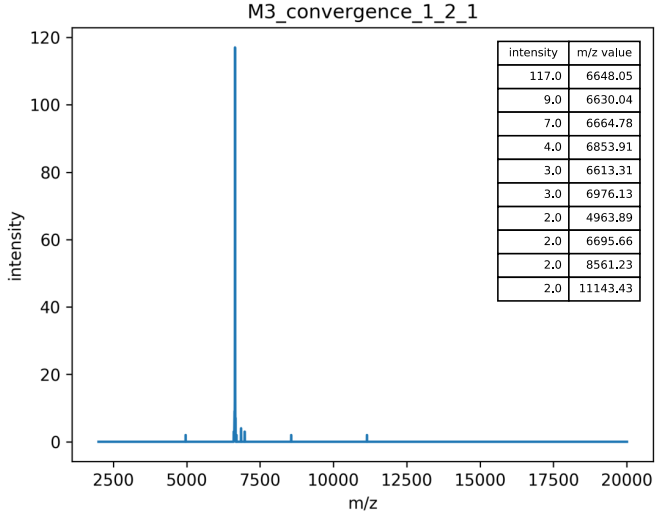


intensity	m/z value
1.58	3324.77
1.39	3139.49
1.28	3453.42
1.26	7001.87
1.25	4962.61
1.25	2012.43
1.24	11142.14
1.24	2233.73
1.23	2286.48
1.23	6725.25
1.23	6550.27
1.23	2036.88
1.22	2071.62
1.22	2085.77
1.22	3111.19

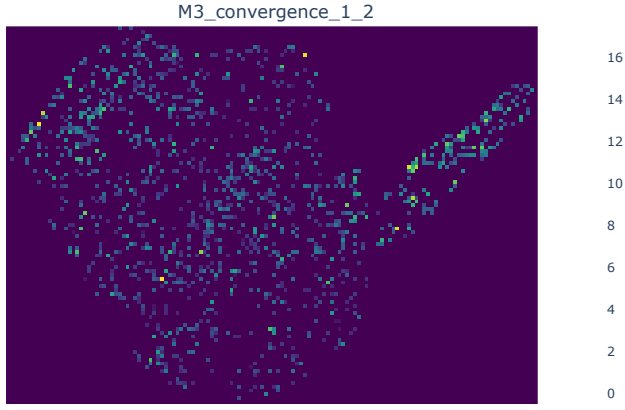

M3_convergence_1_2_2
30
25
20
15
10
5
0

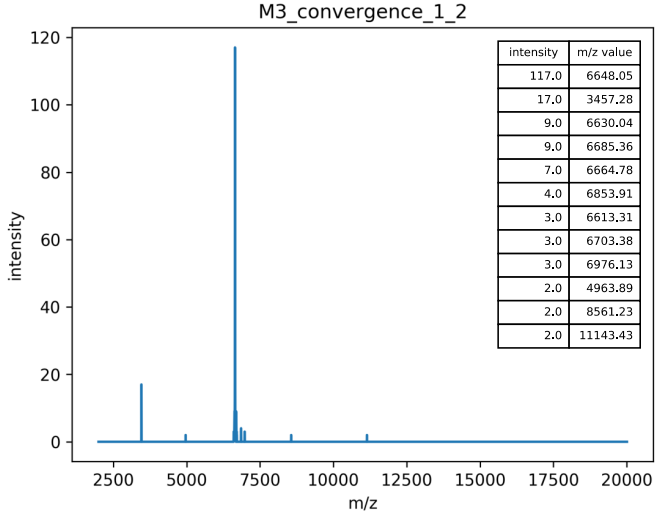

M3_convergence_1_2_2
intensity
m/z value
109.0
3457.28
17.0
6685.36
10.0
6669.92
8.0
6704.66
intensity
100
80
60
40
20
0
2500
5000
7500
10000
12500
15000
17500
20000
m/z



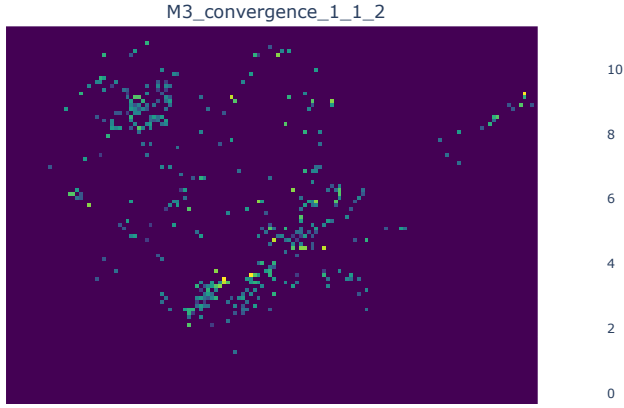


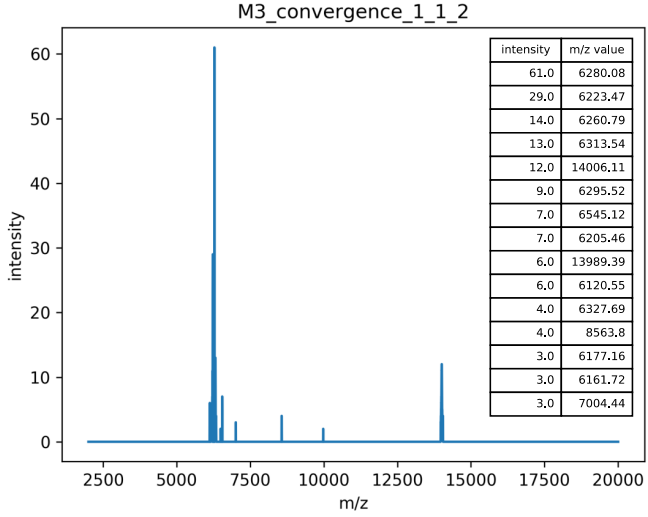
intensity	m/z value
117.0	6648.05
9.0	6630.04
7.0	6664.78
4.0	6853.91
3.0	6613.31
3.0	6976.13
2.0	4963.89
2.0	6695.66
2.0	8561.23
2.0	11143.43



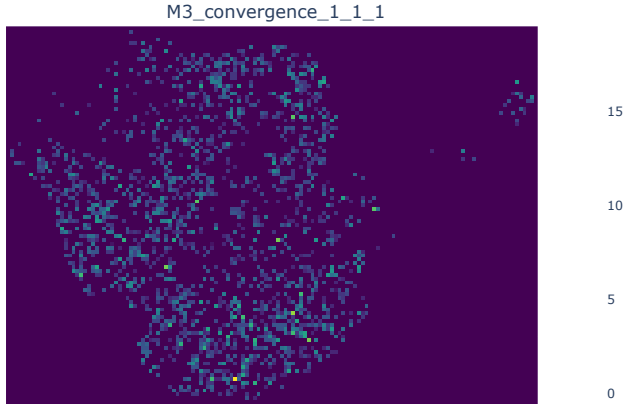


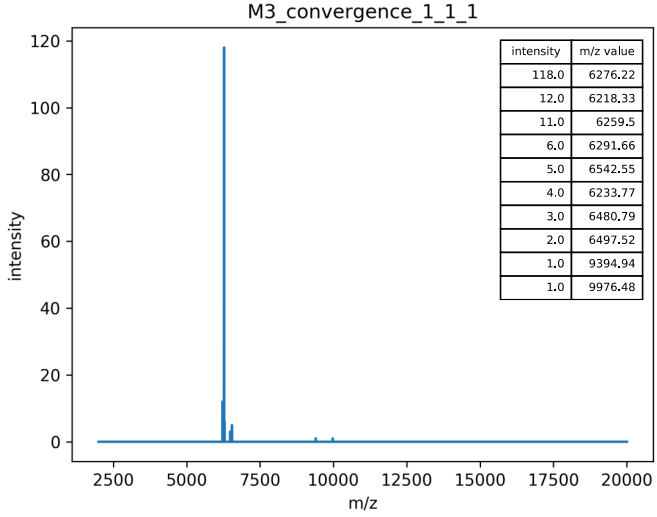
intensity	m/z value
117.0	6648.05
17.0	3457.28
9.0	6630.04
9.0	6685.36
7.0	6664.78
4.0	6853.91
3.0	6613.31
3.0	6703.38
3.0	6976.13
2.0	4963.89
2.0	8561.23
2.0	11143.43



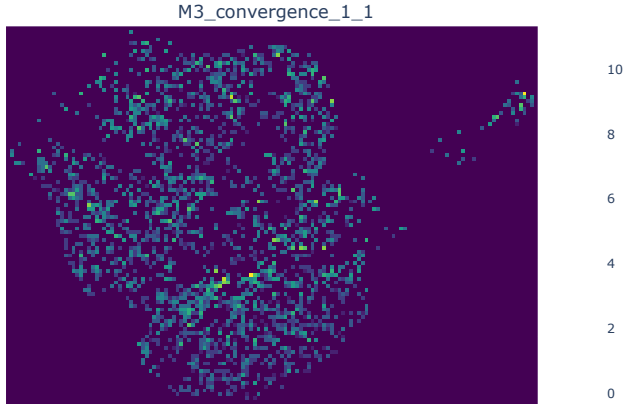


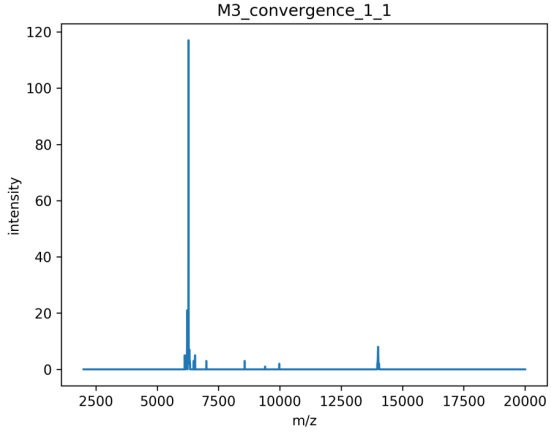
intensity	m/z value
61.0	6280.08
29.0	6223.47
14.0	6260.79
13.0	6313.54
12.0	14006.11
9.0	6295.52
7.0	6545.12
7.0	6205.46
6.0	13989.39
6.0	6120.55
4.0	6327.69
4.0	8563.8
3.0	6177.16
3.0	6161.72
3.0	7004.44



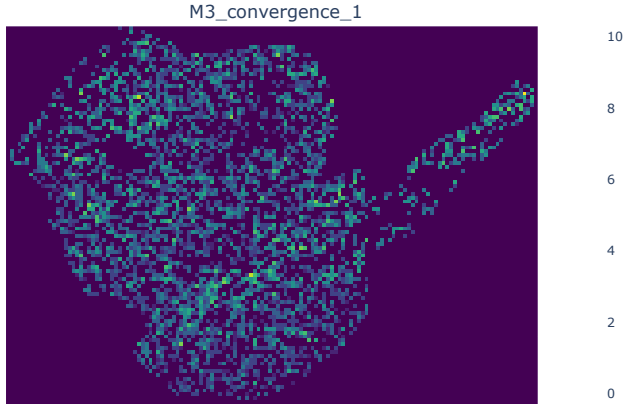


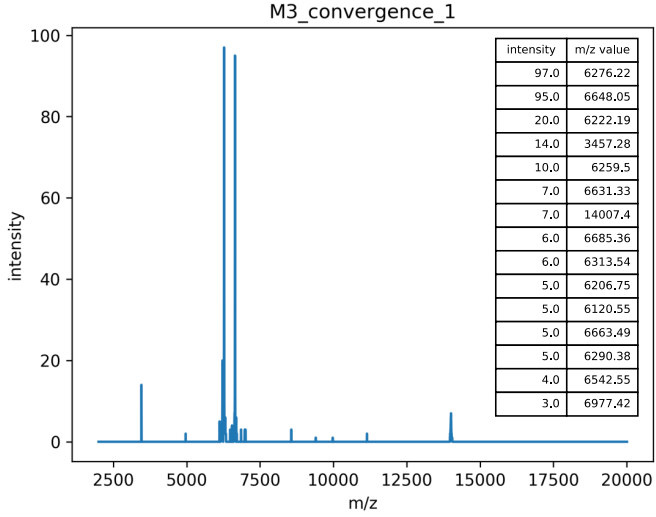
intensity	m/z value
118.0	6276.22
12.0	6218.33
11.0	6259.5
6.0	6291.66
5.0	6542.55
4.0	6233.77
3.0	6480.79
2.0	6497.52
1.0	9394.94
1.0	9976.48





intensity	m/z value
117.0	6276.22
21.0	6222.19
12.0	6259.5
8.0	14004.83
7.0	6313.54
6.0	6291.66
5.0	6205.46
5.0	6542.55
5.0	6120.55
3.0	13985.53
3.0	6480.79
3.0	7004.44
3.0	8563.8
2.0	6159.14
2.0	6497.52





intensity	m/z value
97.0	6276.22
95.0	6648.05
20.0	6222.19
14.0	3457.28
10.0	6259.5
7.0	6631.33
7.0	14007.4
6.0	6685.36
6.0	6313.54
5.0	6206.75
5.0	6120.55
5.0	6663.49
5.0	6290.38
4.0	6542.55
3.0	6977.42

c. ion images of m/z-values of interest in the Islets of Langerhans

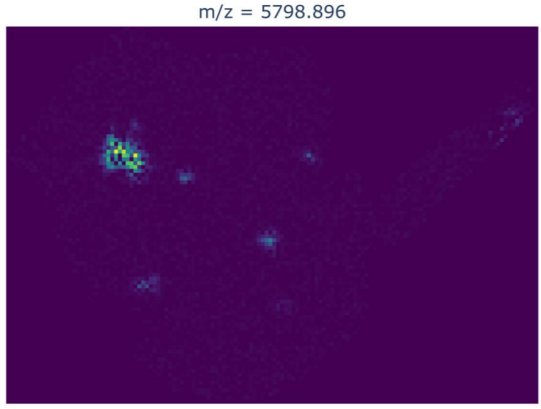


Figure S 9. Ion image of Ins2 for M2.

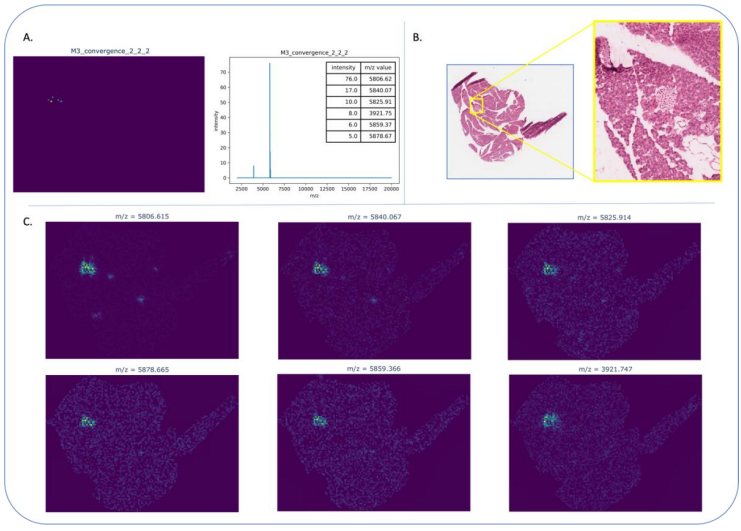


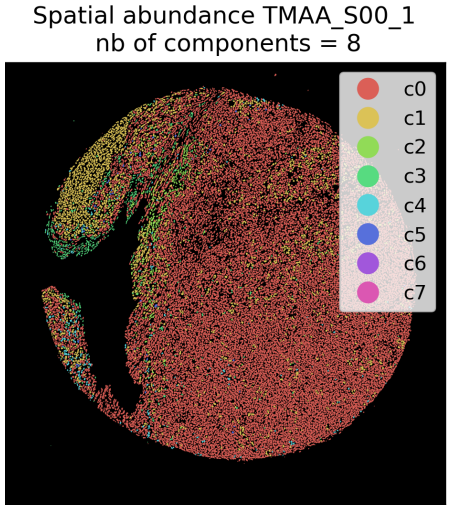
Figure S 10. Spatial correspondence of cluster "M3_convergence_2_2" and the ion images of the corresponding m/z-peaks. The cluster shows relation between Ins1 (m/z = 5808.5 Da and m/z = 2906.59 Da) and the outer periphery of the islets of Langerhans in mice. (A) Result cluster obtained with rr2-NMF. (B) H&E staining of the M3 sample. (C) Ion images of the peak values corresponding to this cluster.

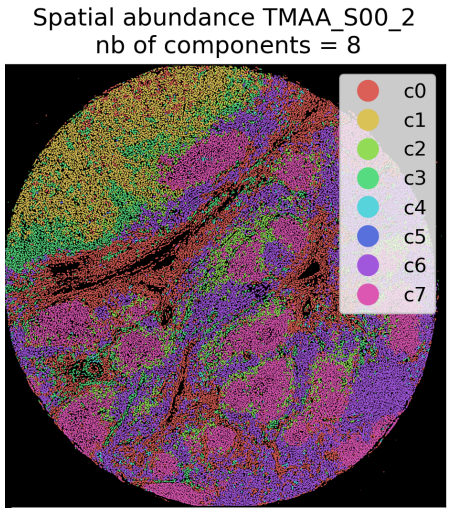
Appendix D

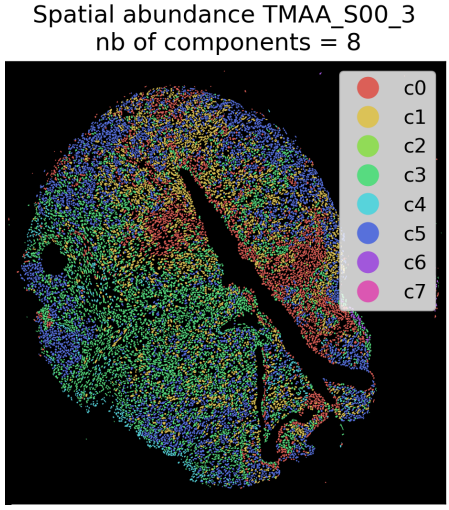
Supporting Information for An NMF-MLP pipeline for fast exploration and annotation of single cell MILAN data

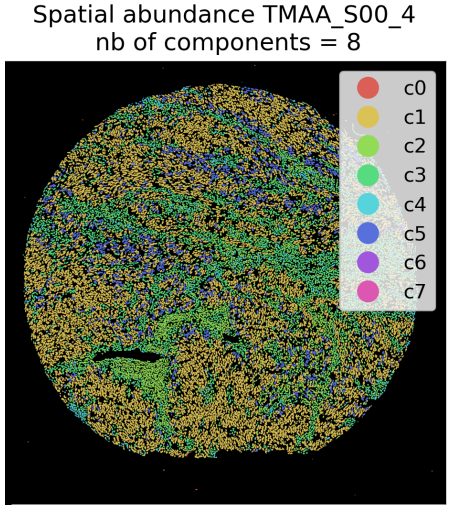
D.1 Multi-sample melanoma analysis

The following figures belong to the multi-sample analysis of the melanoma dataset, as described in Chapter 5.

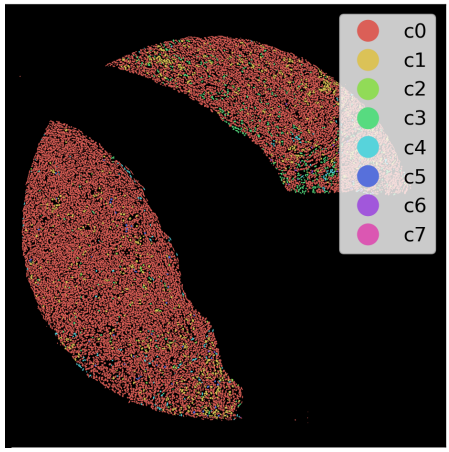




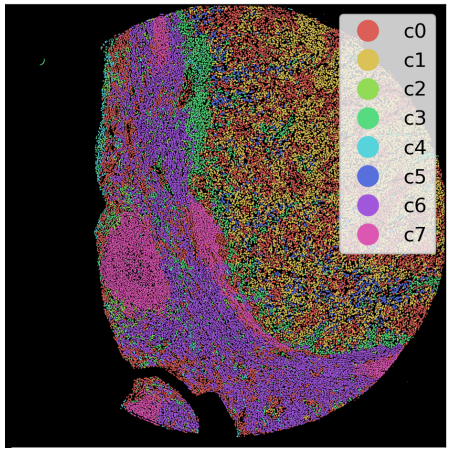

Spatial abundance TMAA_S00_3
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

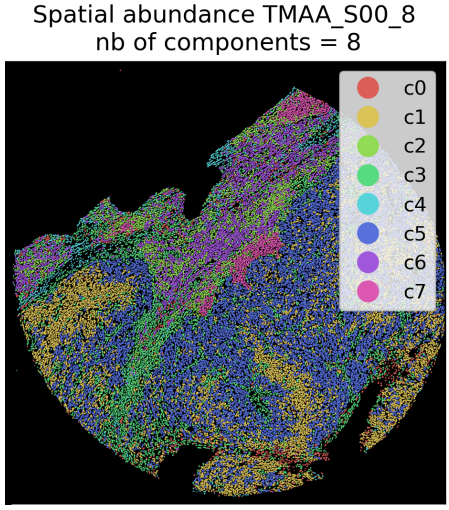

Spatial abundance TMAA_S00_4
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

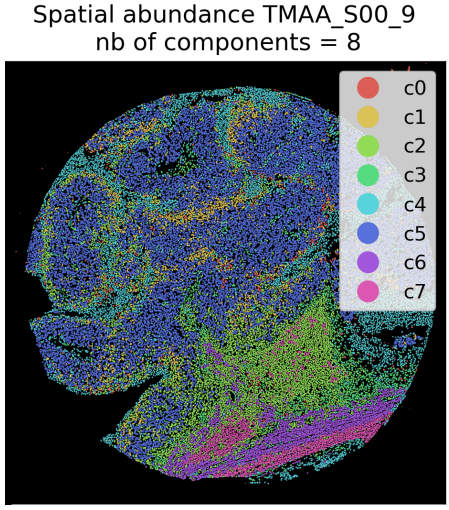
Spatial abundance TMAA_S00_6
nb of components = 8

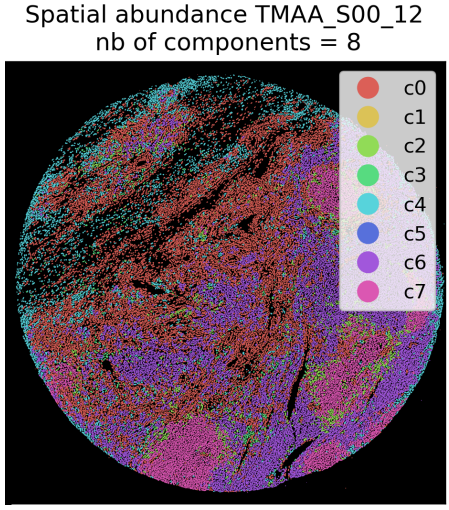


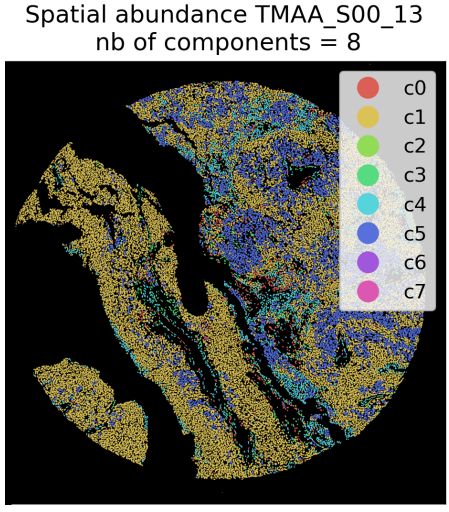
Spatial abundance TMAA_S00_7
nb of components = 8

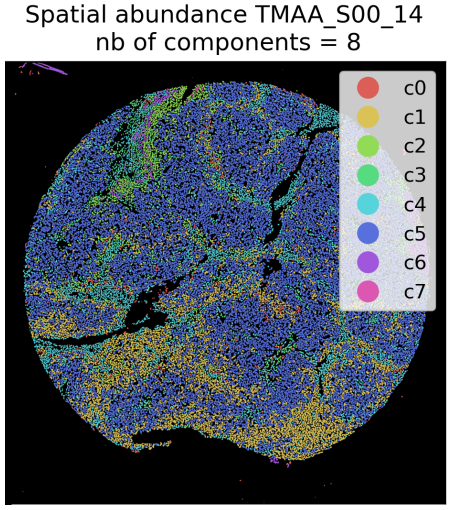


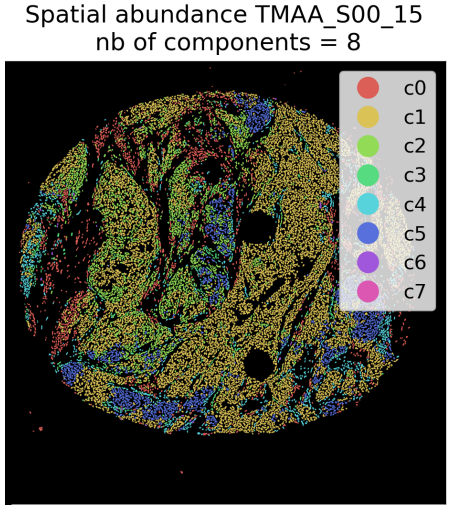

Spatial abundance TMAA_S00_8
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7


Spatial abundance TMAA_S00_9
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

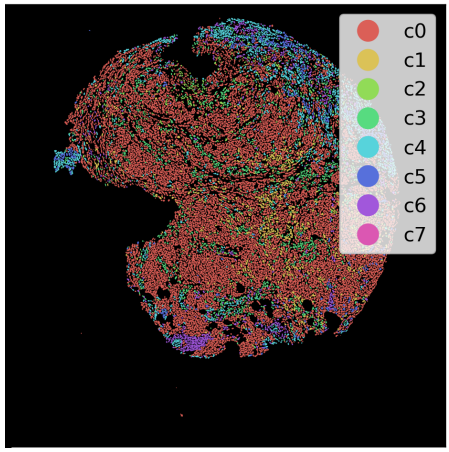

Spatial abundance TMAA_S00_12
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7


Spatial abundance TMAA_S00_13
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

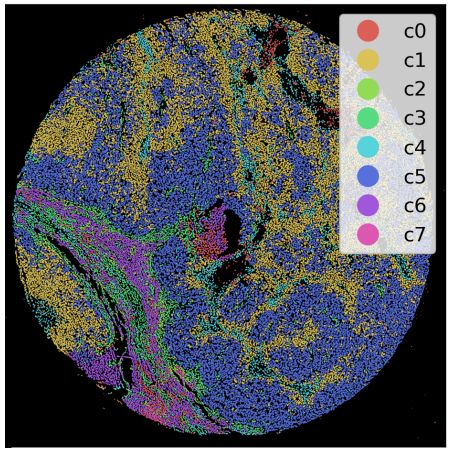

Spatial abundance TMAA_S00_14
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

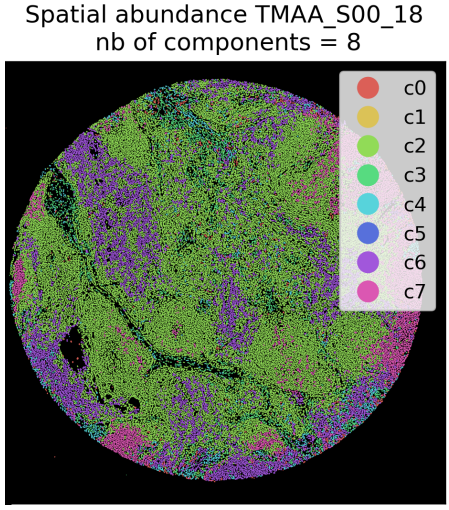

Spatial abundance TMAA_S00_15
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

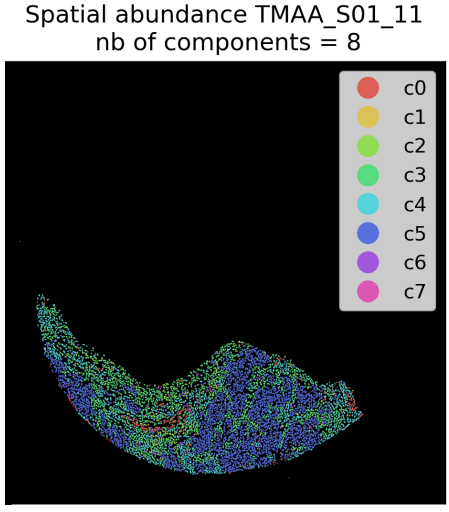
Spatial abundance TMAA_S00_16
nb of components = 8

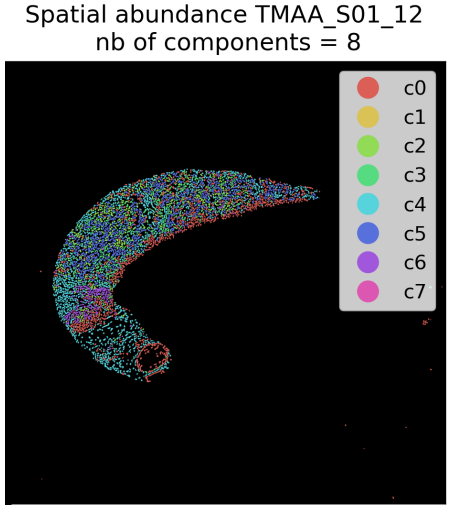


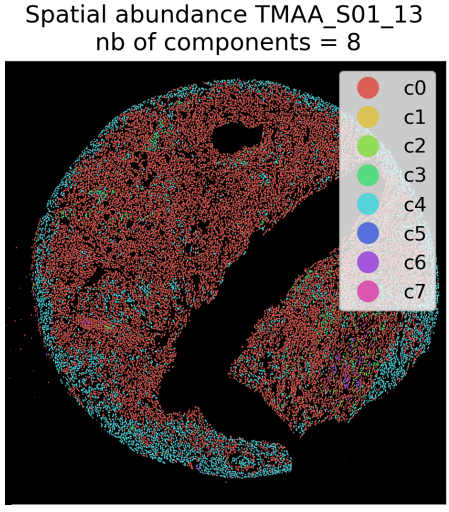
Spatial abundance TMAA_S00_17
nb of components = 8

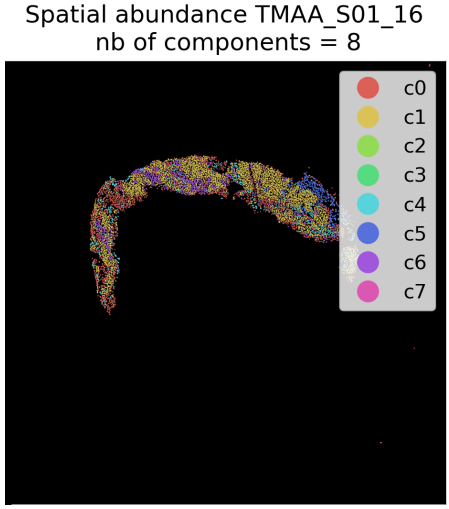


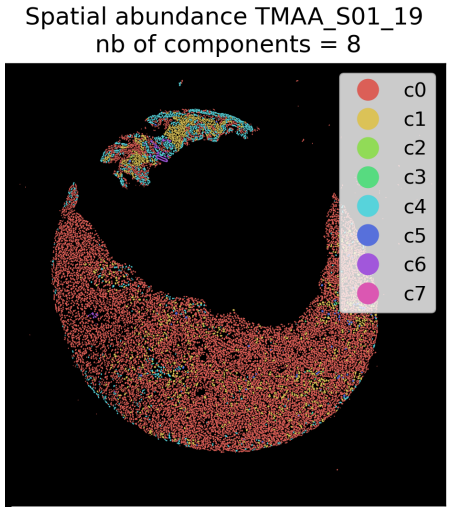

Spatial abundance TMAA_S00_18
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7


Spatial abundance TMAA_S01_11
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

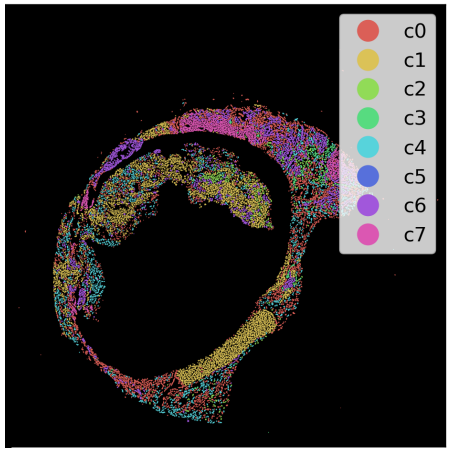

Spatial abundance TMAA_S01_12
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7


Spatial abundance TMAA_S01_13
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

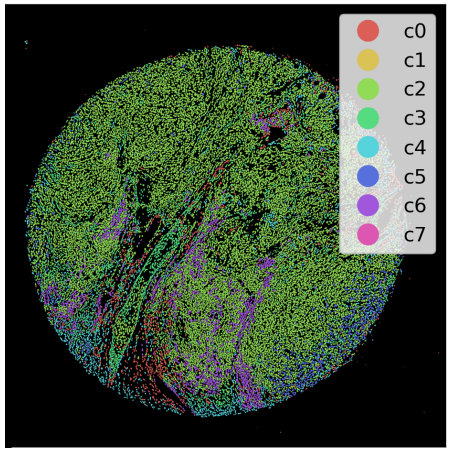

Spatial abundance TMAA_S01_16
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

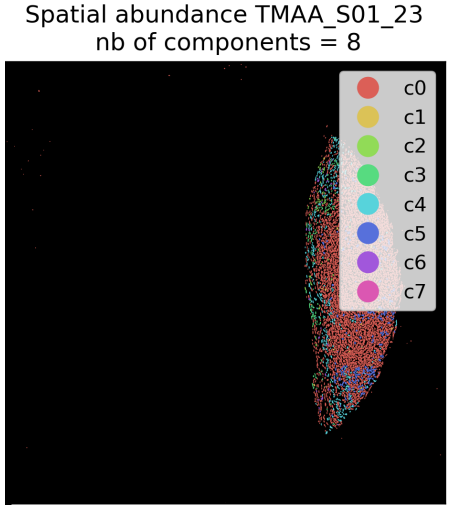

Spatial abundance TMAA_S01_19
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

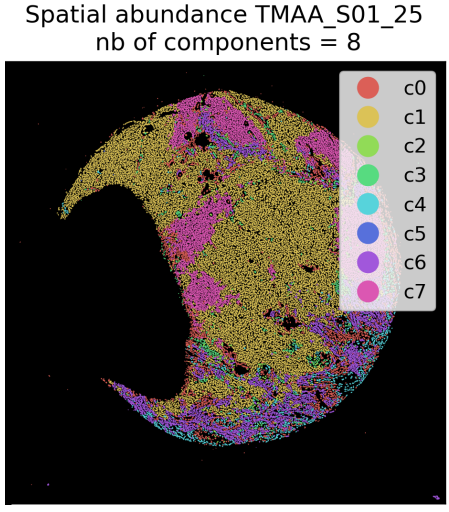
Spatial abundance TMAA_S01_20
nb of components = 8

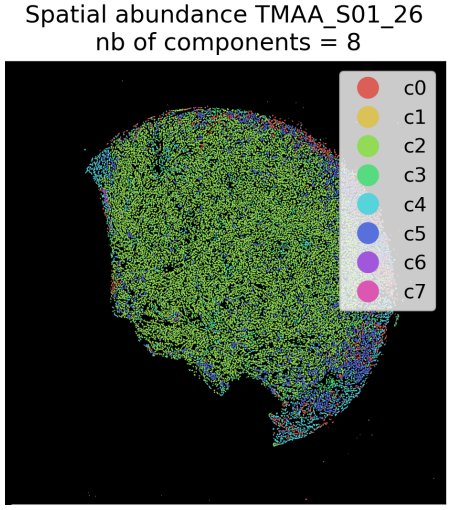


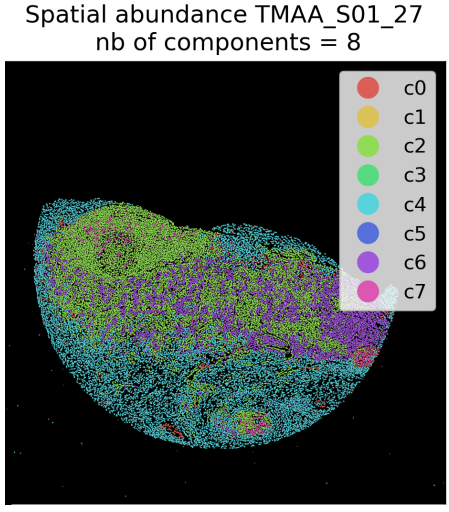
Spatial abundance TMAA_S01_21
nb of components = 8

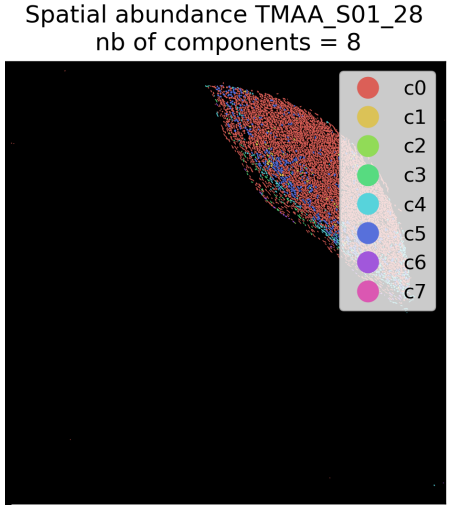


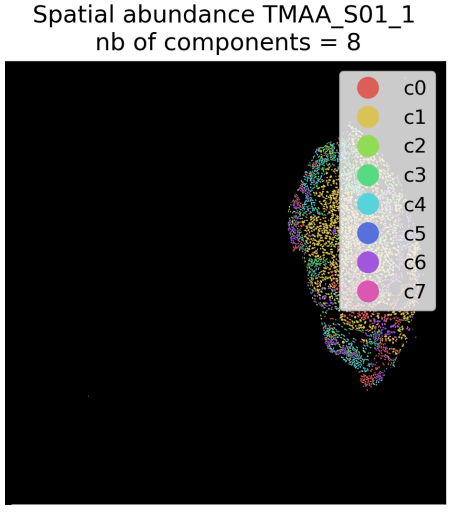

Spatial abundance TMAA_S01_23
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7


Spatial abundance TMAA_S01_25
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

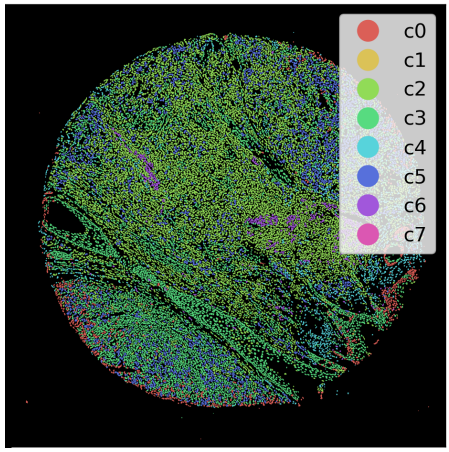

Spatial abundance TMAA_S01_26
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7


Spatial abundance TMAA_S01_27
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

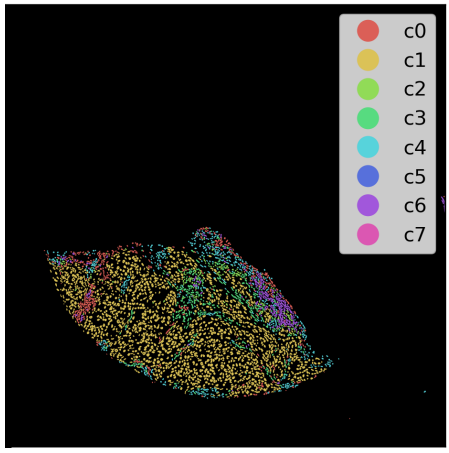

Spatial abundance TMAA_S01_28
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

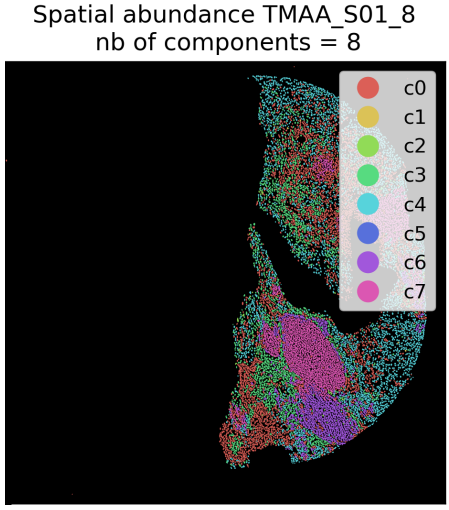

Spatial abundance TMAA_S01_1
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

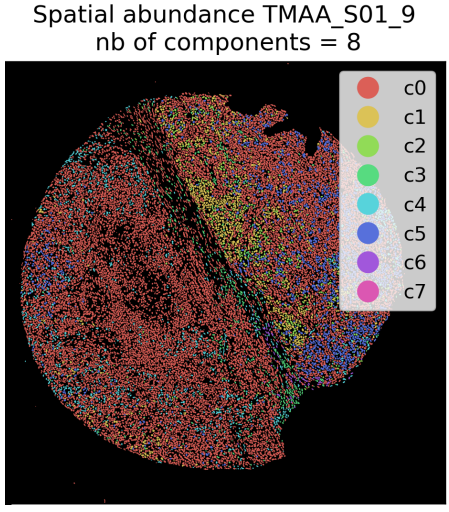
Spatial abundance TMAA_S01_4
nb of components = 8

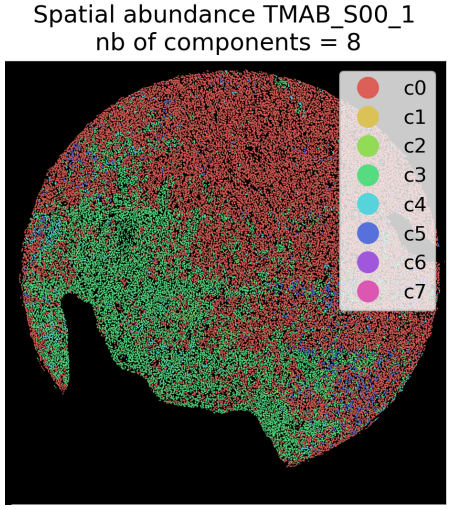


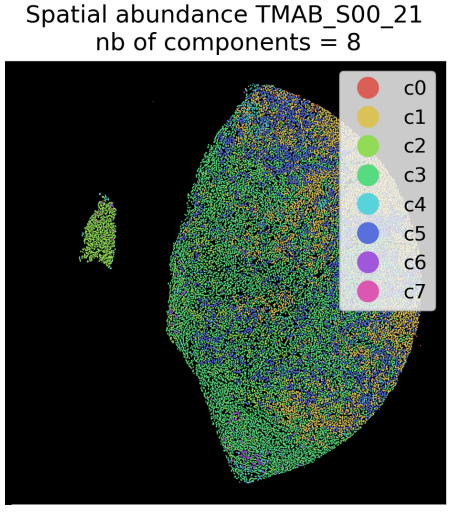
Spatial abundance TMAA_S01_6
nb of components = 8

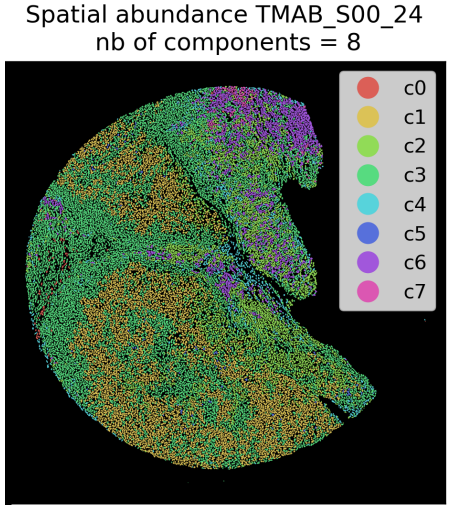


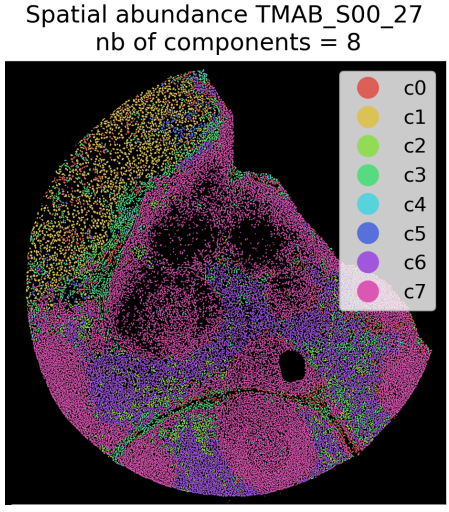

Spatial abundance TMAA_S01_8
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

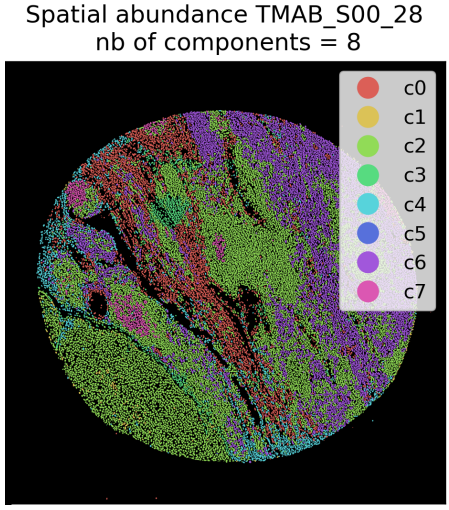

Spatial abundance TMAA_S01_9
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

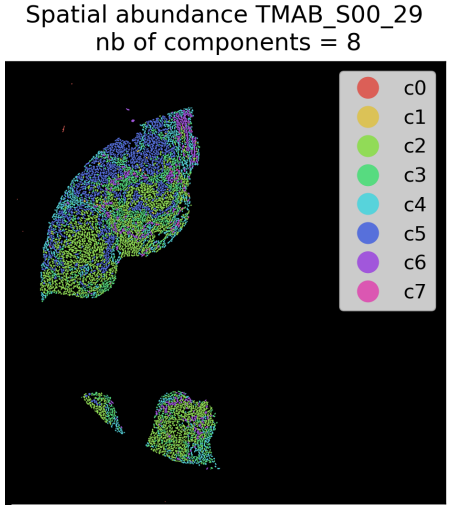

Spatial abundance TMAB_S00_1
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

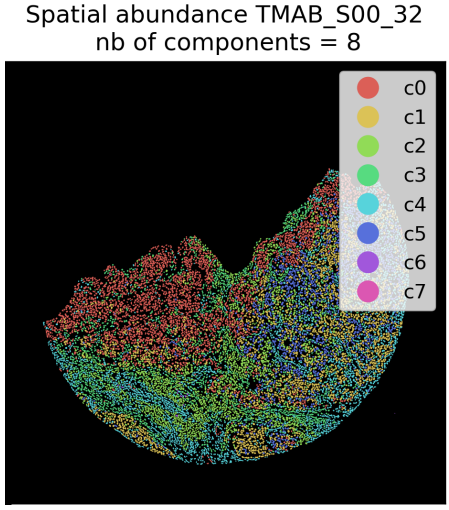

Spatial abundance TMAB_S00_21
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

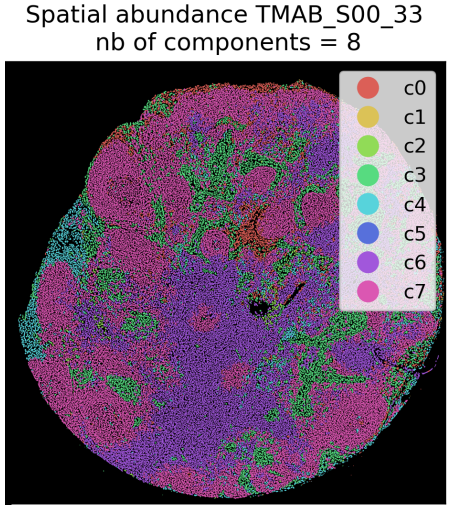

Spatial abundance TMAB_S00_24
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7


Spatial abundance TMAB_S00_27
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

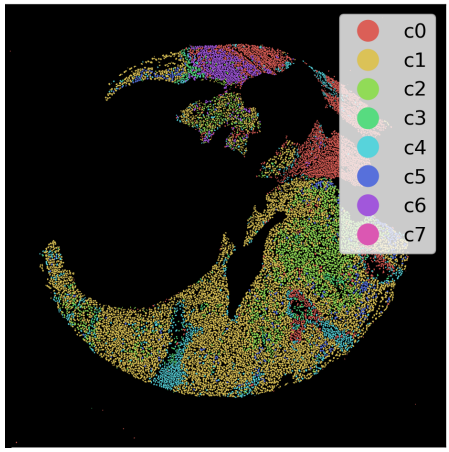

Spatial abundance TMAB_S00_28
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7


Spatial abundance TMAB_S00_29
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

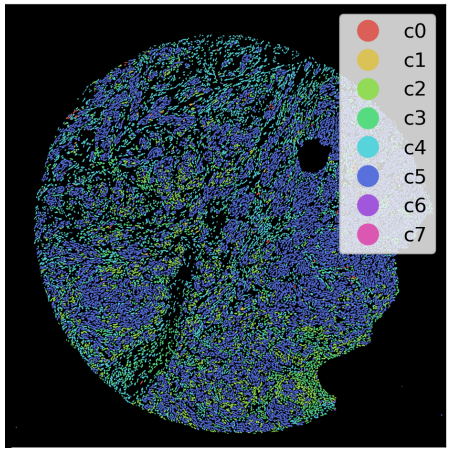

Spatial abundance TMAB_S00_32
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

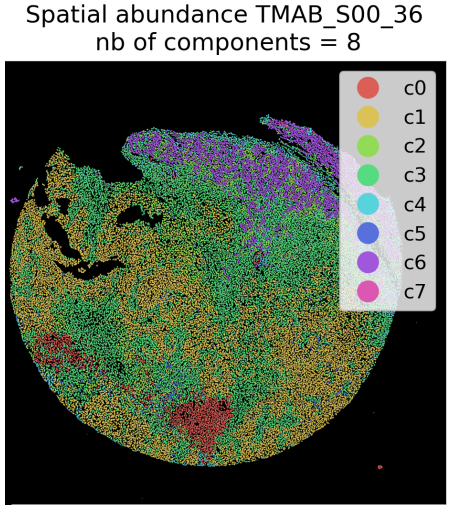

Spatial abundance TMAB_S00_33
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

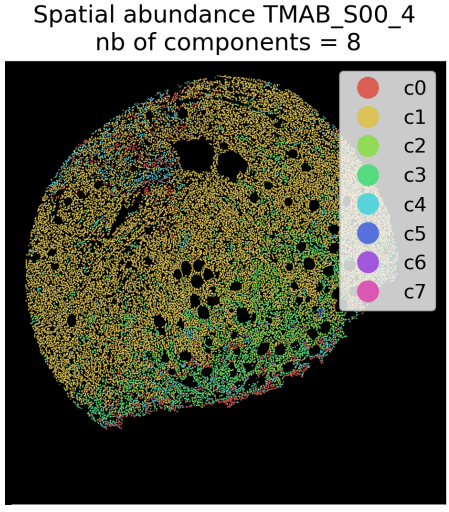
Spatial abundance TMAB_S00_34
nb of components = 8

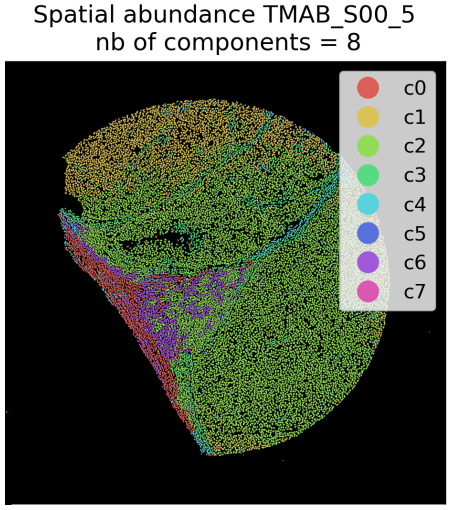


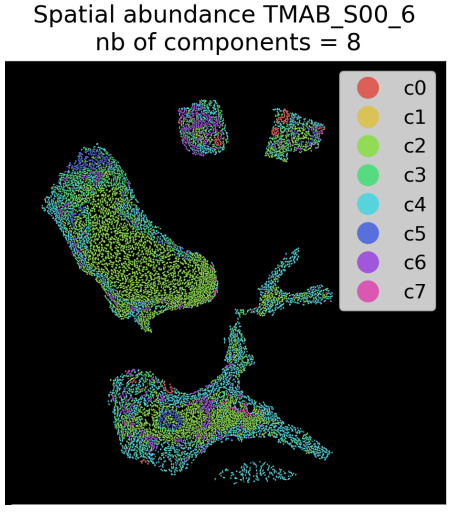
Spatial abundance TMAB_S00_35
nb of components = 8

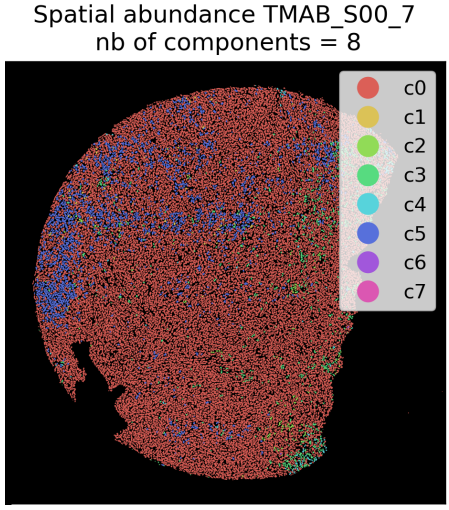


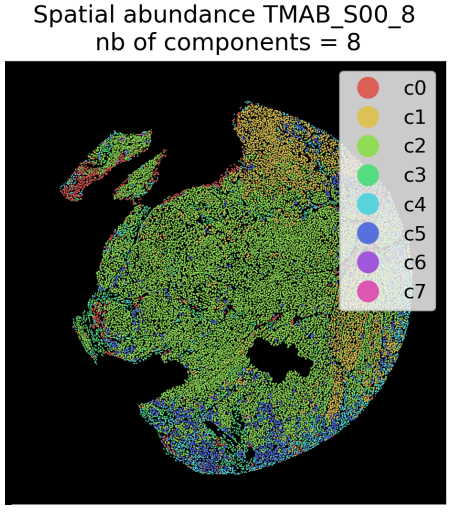

Spatial abundance TMAB_S00_36
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7


Spatial abundance TMAB_S00_4
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

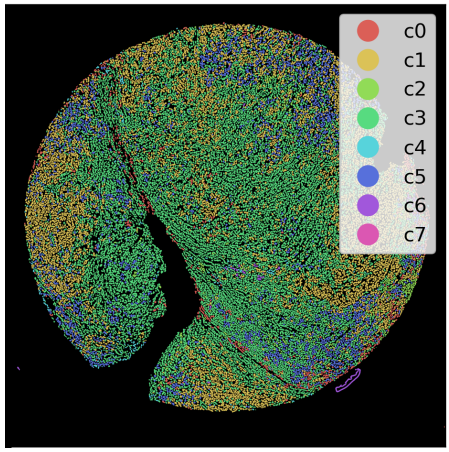

Spatial abundance TMAB_S00_5
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7


Spatial abundance TMAB_S00_6
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

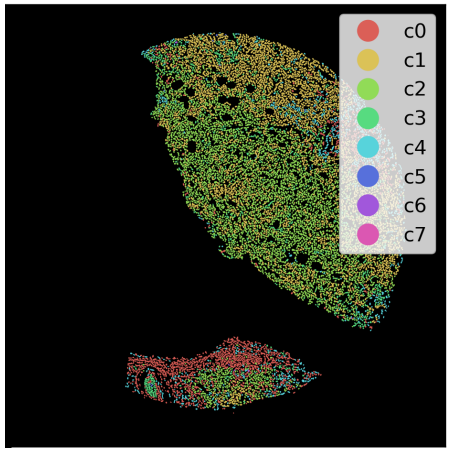

Spatial abundance TMAB_S00_7
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

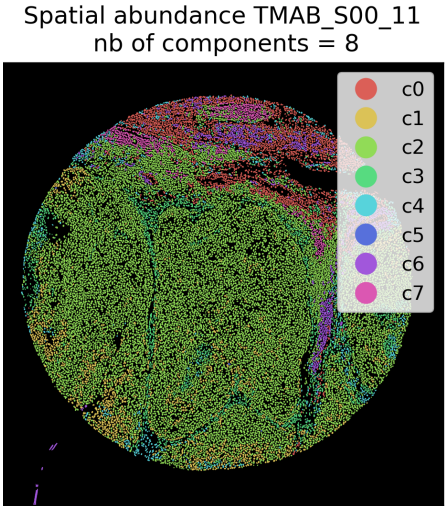

Spatial abundance TMAB_S00_8
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

Spatial abundance TMAB_S00_9
nb of components = 8

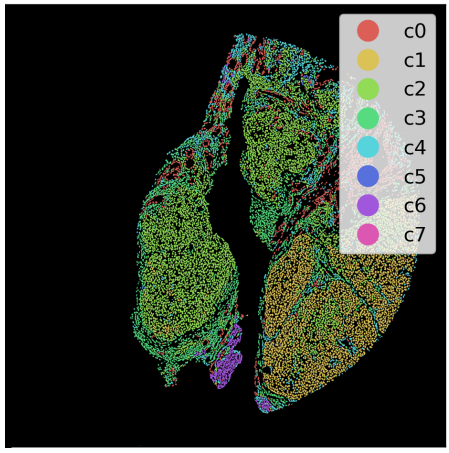


Spatial abundance TMAB_S00_10
nb of components = 8

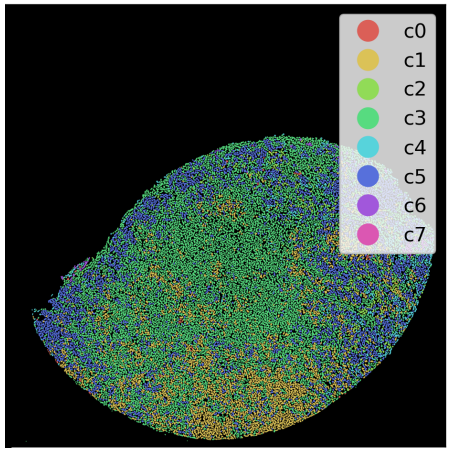


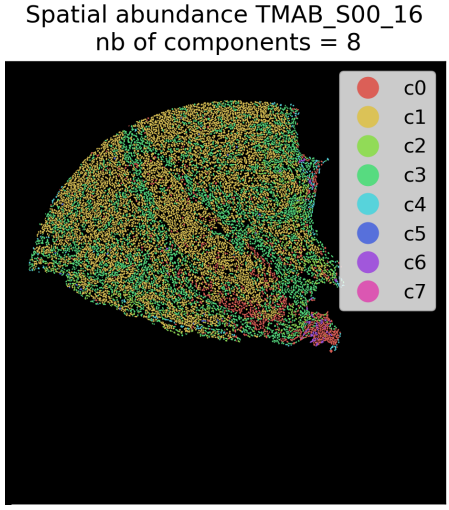

Spatial abundance TMAB_S00_11
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

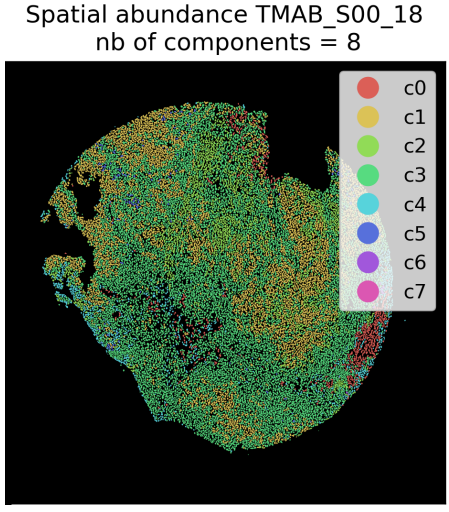
Spatial abundance TMAB_S00_14
nb of components = 8

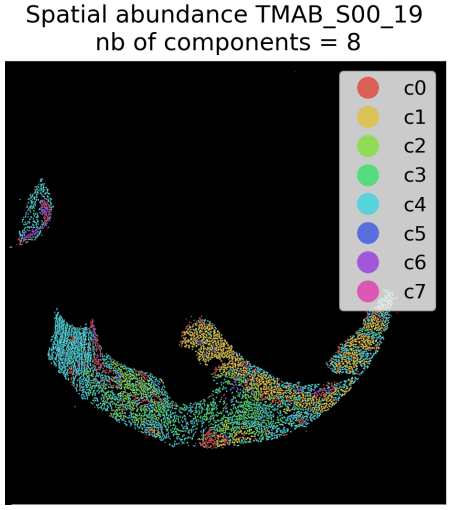


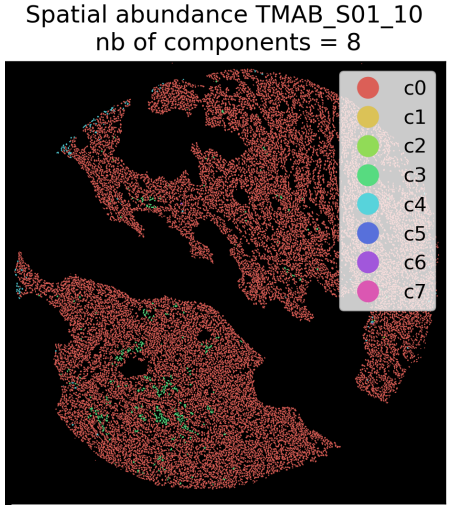
Spatial abundance TMAB_S00_15
nb of components = 8

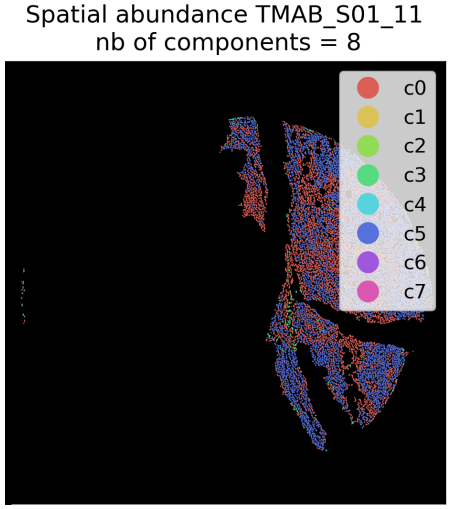


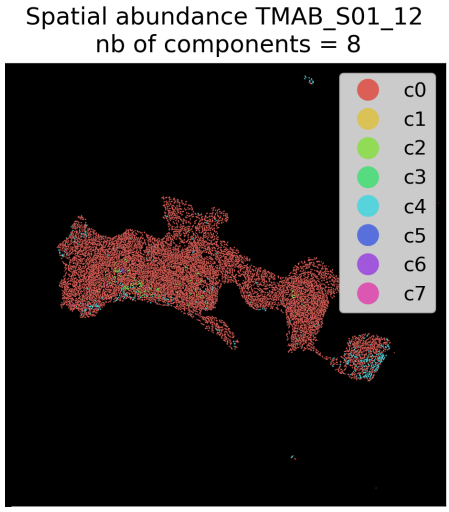

Spatial abundance TMAB_S00_16
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7


Spatial abundance TMAB_S00_18
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

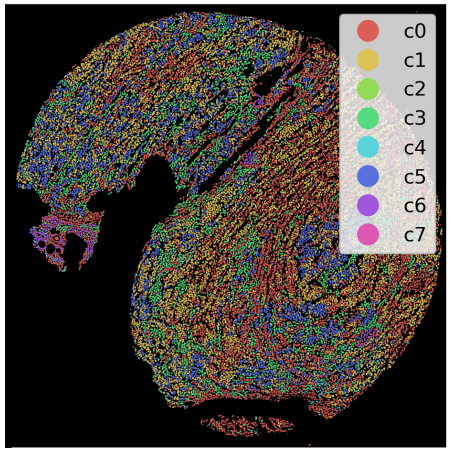

Spatial abundance TMAB_S00_19
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7


Spatial abundance TMAB_S01_10
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

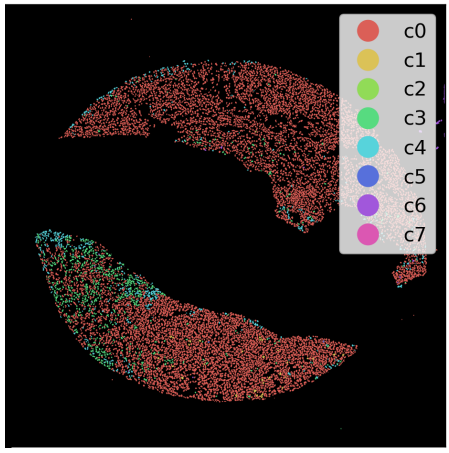

Spatial abundance TMAB_S01_11
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

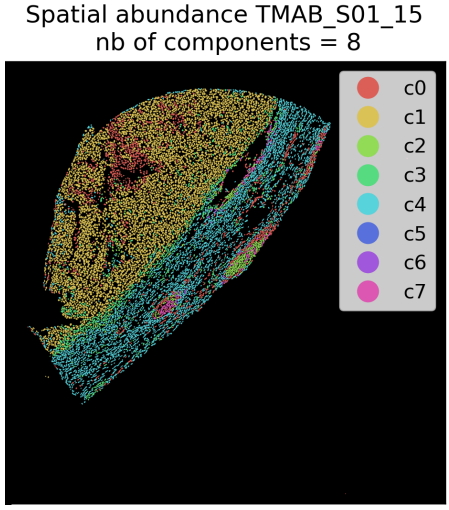

Spatial abundance TMAB_S01_12
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

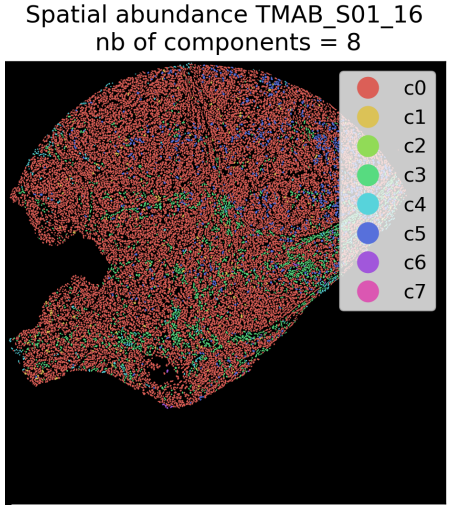
Spatial abundance TMAB_S01_13
nb of components = 8

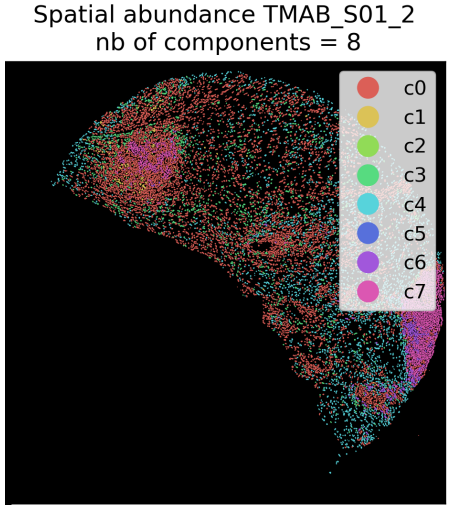


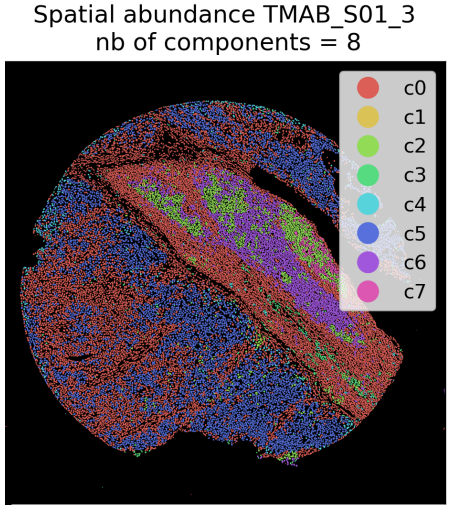
Spatial abundance TMAB_S01_14
nb of components = 8

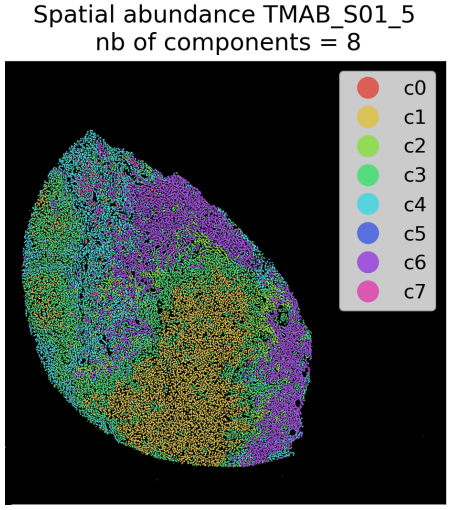


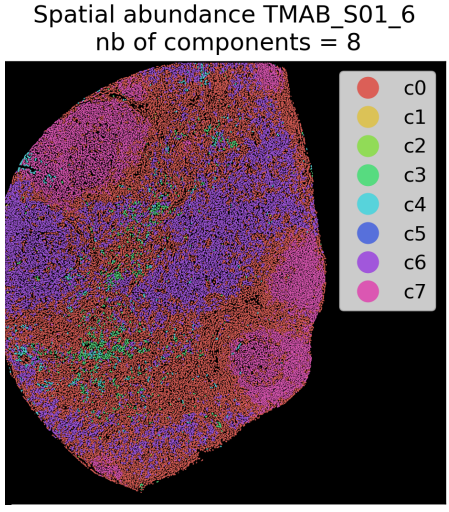

Spatial abundance TMAB_S01_15
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

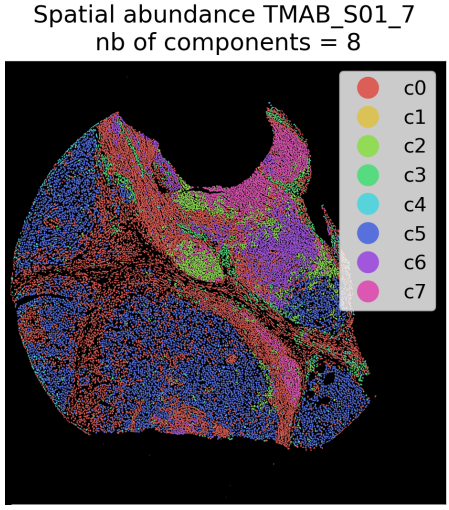

Spatial abundance TMAB_S01_16
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

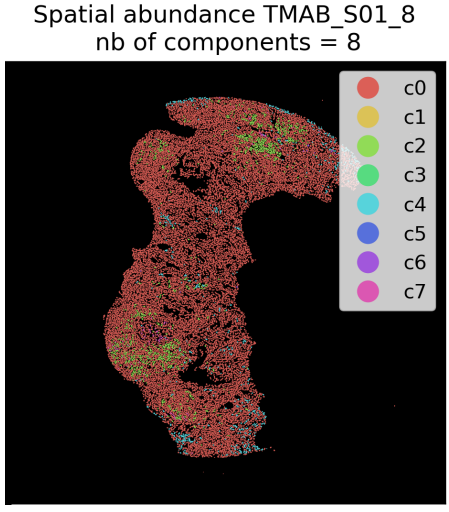

Spatial abundance TMAB_S01_2
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

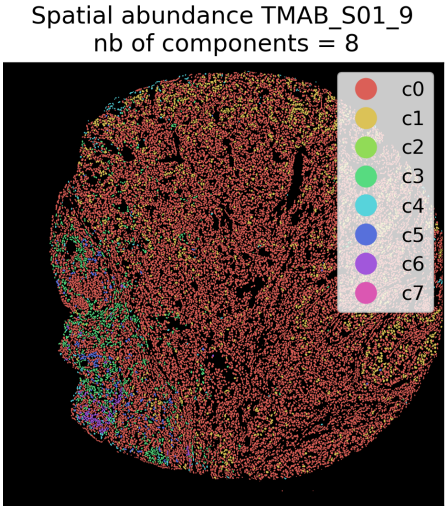

Spatial abundance TMAB_S01_3
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7


Spatial abundance TMAB_S01_5
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7


Spatial abundance TMAB_S01_6
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7


Spatial abundance TMAB_S01_7
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

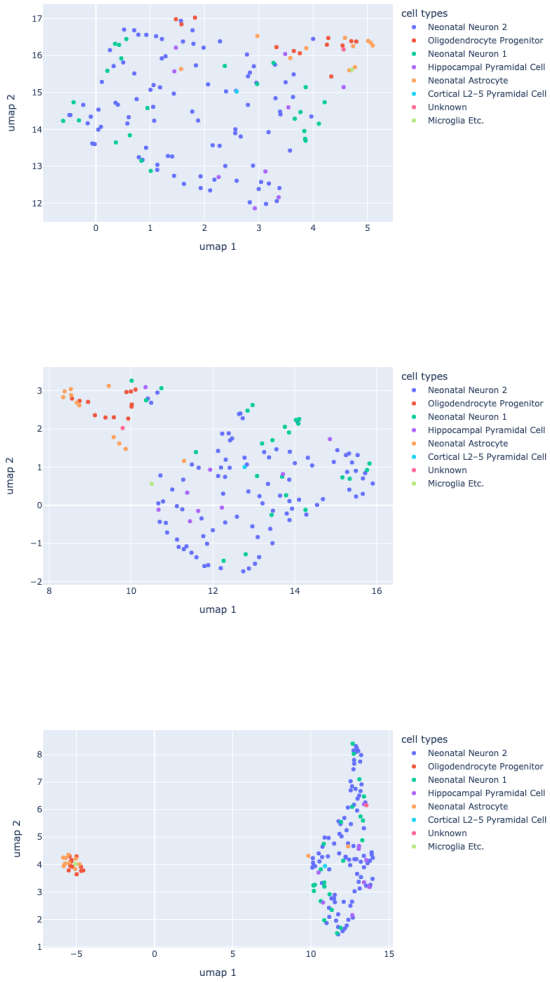

Spatial abundance TMAB_S01_8
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7


Spatial abundance TMAB_S01_9
nb of components = 8
c0
c1
c2
c3
c4
c5
c6
c7

Appendix E

Supporting Information for Matrix and tensor factorization of single cell Hi-C data

E.1 UMAP embeddings for standard NMF

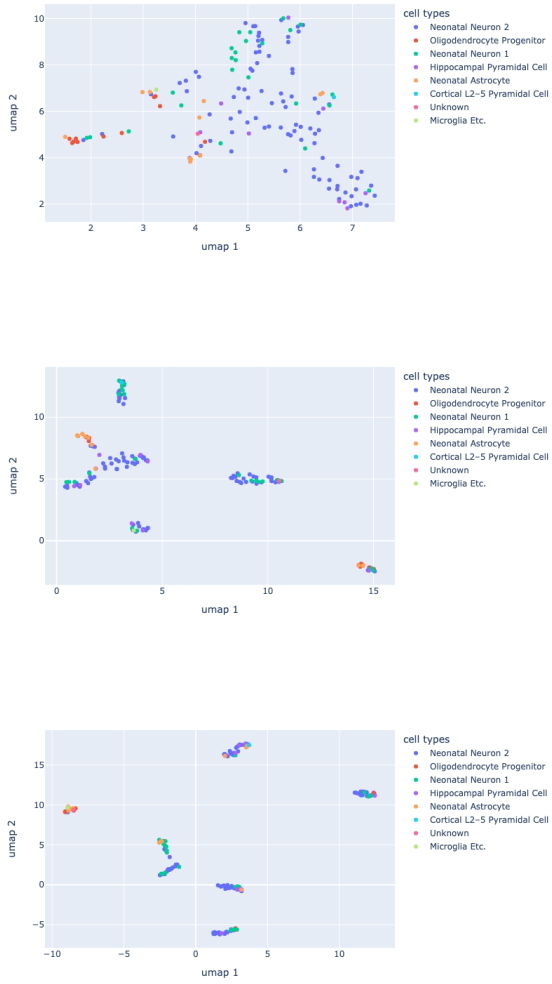


UMAP embeddings of NMF results for different resolutions. (a) 1Mb resolution. (b) 500kb resolution. (c) 100kb resolution.



UMAP embeddings of NMF results for different preprocessing techniques for bin resolution of 1Mb. (a) BandNorm. (b) Band correction for decay function. (c) Imputed data using polymer chain model.

E.2 UMAP embeddings for KL-NMF

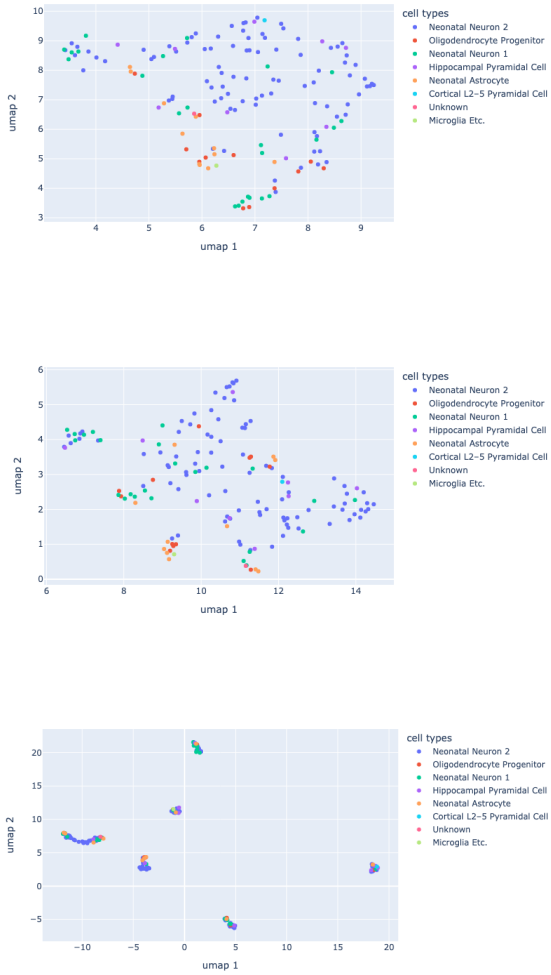


UMAP embeddings of KL-NMF results for different resolutions. (a) 1Mb resolution. (b) 500kb resolution. (c) 100kb resolution.



UMAP embeddings of KL-NMF results for different preprocessing techniques for bin resolution of 1Mb. (a) BandNorm. (b) Band correction for decay function. (c) Imputed data using polymer chain model.

E.3 UMAP embeddings for LDA

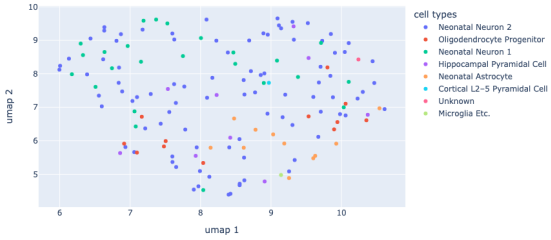


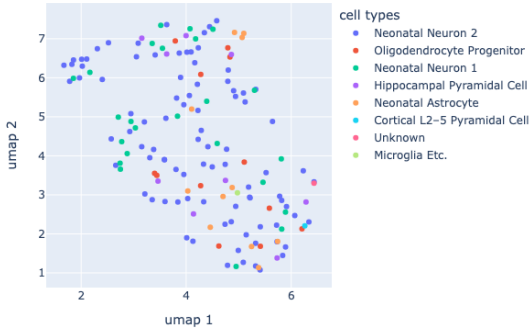
UMAP embeddings of LDA results for different resolutions. (a) 1Mb resolution. (b) 500kb resolution. (c) 100kb resolution.

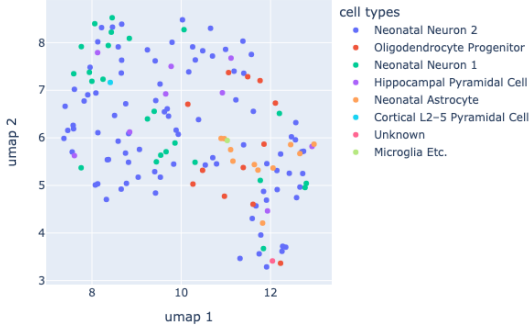


UMAP embeddings of LDA results for different preprocessing techniques for bin resolution of 1Mb. (a) BandNorm. (b) Band correction for decay function. (c) Imputed data using polymer chain model.

E.4 UMAP embeddings for Tucker





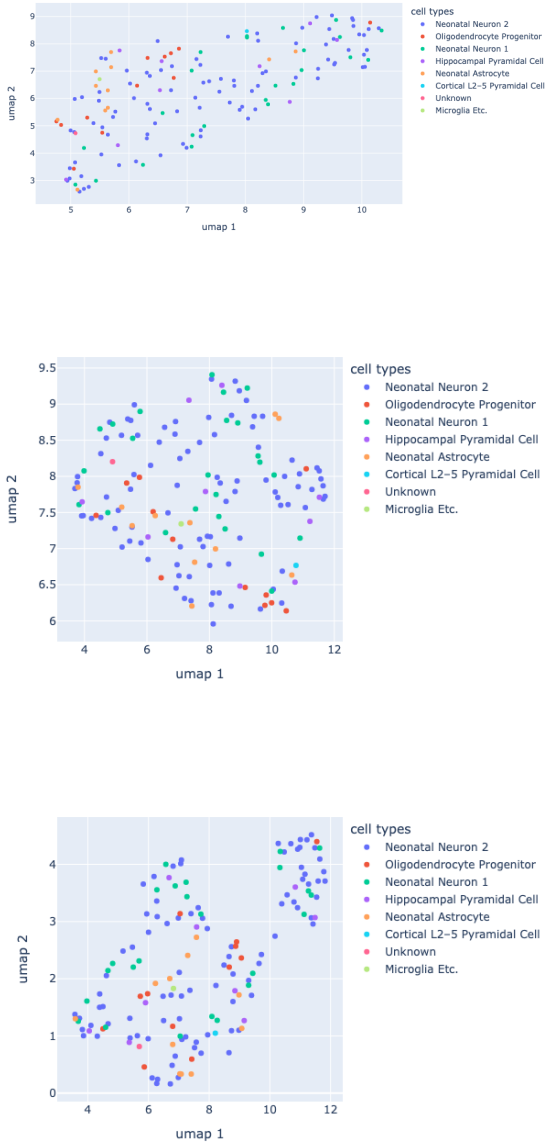


UMAP embeddings of Tucker results for different resolutions. (a) 1Mb resolution. (b) 500kb resolution. (c) 100kb resolution.



UMAP embeddings of Tucker results for different preprocessing techniques for bin resolution of 1Mb. (a) BandNorm. (b) Band correction for decay function. (c) Imputed data using polymer chain model.

E.5 UMAP embeddings for CPD



UMAP embeddings of CPD results for different resolutions. (a) 1Mb resolution. (b) 500kb resolution. (c) 100kb resolution.



UMAP embeddings of CPD results for different preprocessing techniques for bin resolution of 1Mb. (a) BandNorm. (b) Band correction for decay function. (c) Imputed data using polymer chain model.

Bibliography

[1] GitHub - lmcinnes/enstop: Ensemble topic modelling with pLSA.

[2] Multiplexed immunohistochemistry @KU Leuven.

[3] Abdelmoula, W. M., Balluff, B., Englert, S., Dijkstra, J., Reinders, M. J. T., Walch, A., McDonnell, L. A., and Lelieveldt, B. P. F. Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data. *Proceedings of the National Academy of Sciences 113*, 43 (Oct. 2016), 12244–12249. Publisher: Proceedings of the National Academy of Sciences.

[4] Abdelmoula, W. M., Lopez, B. G.-C., Randall, E. C., Kapur, T., Sarkaria, J. N., White, F. M., Agar, J. N., Wells, W. M., and Agar, N. Y. R. Peak learning of mass spectrometry imaging data using artificial neural networks. *Nature Communications 12*, 1 (Sept. 2021), 5544. Publisher: Nature Publishing Group.

[5] Abdelmoula, W. M., Pezzotti, N., Hölt, T., Dijkstra, J., Vilanova, A., McDonnell, L. A., and 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. *Journal of Proteome Research 17*, 3 (Mar. 2018), 1054–1064. Publisher: American Chemical Society.

[6] Abdelmoula, W. M., Stopka, S. A., Randall, E. C., Regan, M., Agar, J. N., Sarkaria, J. N., Wells, W. M., Kapur, T., and Agar, N. Y. R. massNet: integrated processing and classification of spatially resolved mass spectrometry data using deep learning for rapid tumor delineation. *Bioinformatics 38*, 7 (Mar. 2022), 2015–2021.

[7] Aebersold, R., and Mann, M. Mass spectrometry-based proteomics. *Nature 422*, 6928 (Mar. 2003), 198–207. Publisher: Nature Publishing Group.

[8] Al-Amrani, S., Al-Jabri, Z., Al-Zaabi, A., Alshekaili, J., and Al-Khabori, M. Proteomics: Concepts and applications in human medicine. *World Journal of Biological Chemistry 12*, 5 (Sept. 2021), 57–69.

[9] Alexandrov, T. MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC Bioinformatics 13*, 16 (Nov. 2012), S11.

[10] Alexandrov, T. Spatial Metabolomics and Imaging Mass Spectrometry in the Age of Artificial Intelligence. *Annual review of biomedical data science 3*, 1 (July 2020), 61–87. Publisher: Annual Reviews.

[11] Alexandrov, T., Becker, M., Deininger, S.-O., Ernst, G., Wehder, L., Grasmair, M., Von Eggeling, F., Thiele, H., and Maass, P. Spatial Segmentation of Imaging Mass Spectrometry Data with Edge-Preserving Image Denoising and Clustering. *Journal of Proteome Research 9*, 12 (Dec. 2010), 6535–6546.

[12] Aljarah, I., Faris, H., and Mirjalili, S., Eds. *Evolutionary Data Clustering: Algorithms and Applications*. Algorithms for Intelligent Systems. Springer Singapore, Singapore, 2021.

[13] Amézquita, E. J., Quigley, M. Y., Ophelders, T., Munch, E., and Chitwood, D. H. The shape of things to come: Topological data analysis and biology, from molecules to organisms. *Developmental Dynamics 249*, 7 (2020), 816–833. _eprint: https://anatomypubs.onlinelibrary.wiley.com/doi/pdf/10.1002/dvdy.175.

[14] Aslam, B., Basit, M., Nisar, M. A., Khurshid, M., and Rasool, M. H. Proteomics: Technologies and Their Applications. *Journal of Chromatographic Science 55*, 2 (Feb. 2017), 182–196.

[15] Baquer, G., Sementé, L., Mahamdi, T., Correig, X., Ràfols, P., and 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 42*, 5 (2023), 1927–1964. _eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/mas.21794.

[16] Barh, D., Tiwari, S., Weener, M. E., Azevedo, V., Góes-Neto, A., Gromiha, M. M., and Ghosh, P. Multi-omics-based identification

of SARS-CoV-2 infection biology and candidate drugs against COVID-19. *Computers in Biology and Medicine 126* (Nov. 2020), 104051.

[17] Barh, D., Zambare, V., and Azevedo, V. *OMICS: Applications in Biomedical, Agricultural, and Environmental Sciences*. CRC Press, Mar. 2013. Google-Books-ID: 8fiHoD7bTJMC.

[18] Barré, F. P., Claes, B. S., Dewez, F., Peutz-Kootstra, C., Munch-Petersen, H. F., Grønbæk, K., Lund, A. H., Heeren, R. M., Côme, C., and Cillero-Pastor, B. Specific Lipid and Metabolic Profiles of R-CHOP-Resistant Diffuse Large B-Cell Lymphoma Elucidated by Matrix-Assisted Laser Desorption Ionization Mass Spectrometry Imaging and in Vivo Imaging. *Analytical Chemistry 90*, 24 (Dec. 2018), 14198–14206. Publisher: American Chemical Society.

[19] Bayani, J., and Squire, J. A. Fluorescence In Situ Hybridization (FISH). *Current Protocols in Cell Biology 23*, 1 (2004), 22.4.1–22.4.52.

[20] Behrmann, J., Etmann, C., Boskamp, T., Casadonte, R., Kriegsmann, J., and Maaβ, P. Deep learning for tumor classification in imaging mass spectrometry. *Bioinformatics 34*, 7 (Apr. 2018), 1215–1223.

[21] Blei, D. M., Ng, A. Y., and Edu, J. B. Latent Dirichlet Allocation Michael I. Jordan. Tech. rep., 2003. Publication Title: Journal of Machine Learning Research Volume: 3.

[22] Bosisio, F. M., Antoranz, A., van Herck, Y., Bolognesi, M. M., Marcelis, L., Chinello, C., Wouters, J., Magni, F., Alexopoulos, L., Stas, M., Boecxstaens, V., Bechter, O., Cattoretti, G., and van den Oord, J. Functional heterogeneity of lymphocytic patterns in primary melanoma dissected through single-cell multiplexing. *eLife 9* (Feb. 2020), e53008. Publisher: eLife Sciences Publications, Ltd.

[23] Bosisio, F. M., Van Herck, Y., Messiaen, J., Bolognesi, M. M., Marcelis, L., Van Haele, M., Cattoretti, G., Antoranz, A., and De Smet, F. Next-Generation Pathology Using Multiplexed Immunohistochemistry: Mapping Tissue Architecture at Single-Cell Level. *Frontiers in Oncology 12* (July 2022), 3444. Publisher: Frontiers Media S.A.

[24] Boskamp, T., Casadonte, R., Hauberg-Lotte, L., Deininger, S., Kriegsmann, J., and Maass, P. Cross-Normalization of MALDI Mass Spectrometry Imaging Data Improves Site-to-Site Reproducibility. *Analytical Chemistry 93*, 30 (Aug. 2021), 10584–10592. Publisher: American Chemical Society.

[25] Boskamp, T., Lachmund, D., Oetjen, J., Cordero Hernandez, Y., Trede, D., Maass, P., Casadonte, R., Kriegsmann, J., Warth, A., Dienemann, H., Weichert, W., and Kriegsmann, M. A new classification method for MALDI imaging mass spectrometry data acquired on formalin-fixed paraffin-embedded tissue samples, 2017. Pages: 916-926 Volume: 1865 Issue: 7 ISSN: 18781454.

[26] Boutsidis, C., and Gallopoulos, E. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition 41*, 4 (Apr. 2008), 1350–1362. Publisher: Pergamon.

[27] Brissova, M., Fowler, M. J., Nicholson, W. E., Chu, A., Hirshberg, B., Harlan, D. M., and Powers, A. C. Assessment of Human Pancreatic Islet Architecture and Composition by Laser Scanning Confocal Microscopy. *Journal of Histochemistry & Cytochemistry 53*, 9 (Sept. 2005), 1087–1097. Publisher: Journal of Histochemistry & Cytochemistry.

[28] Bubenik, P. Statistical Topological Data Analysis using Persistence Landscapes.

[29] Buchberger, A. R., DeLaney, K., Johnson, J., and Li, L. Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights. *Analytical Chemistry 90*, 1 (Jan. 2018), 240–265. Publisher: American Chemical Society.

[30] Bukkuri, A., Andor, N., and Darcy, I. K. Applications of Topological Data Analysis in Oncology. *Frontiers in Artificial Intelligence 4* (Apr. 2021). Publisher: Frontiers.

[31] Cabrera, O., Berman, D. M., Kenyon, N. S., Ricordi, C., Berggren, P.-O., and Caicedo, A. The unique cytoarchitecture of human pancreatic islets has implications for islet cell function. *Proceedings of the National Academy of Sciences 103*, 7 (Feb. 2006), 2334–2339. Publisher: Proceedings of the National Academy of Sciences.

[32] Caliński, T., and Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics 3*, 1 (Jan. 1974), 1–27. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/03610927408827101.

[33] Campbell, J. E., and Newgard, C. B. Mechanisms controlling pancreatic islet cell function in insulin secretion. *Nature Reviews Molecular Cell Biology 22*, 2 (Feb. 2021), 142–158. Publisher: Nature Publishing Group.

[34] Campi, G., Nicora, G., Fiorentino, G., Smith, A., Magni, F., Garagna, S., Zuccotti, M., and Bellazzi, R. A Topological Data Analysis Mapper of the Ovarian Folliculogenesis Based on MALDI Mass Spectrometry Imaging Proteomics. In *Artificial Intelligence in Medicine* (Cham, 2021), A. Tucker, P. Henriques Abreu, J. Cardoso, P. Pereira Rodrigues, and D. Riaño, Eds., Lecture Notes in Computer Science, Springer International Publishing, pp. 43–47.

[35] Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., and Baudot, A. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications 12*, 1 (Dec. 2021). Publisher: Nature Research.

[36] Caprioli, R. M., Farmer, T. B., and Gile, J. Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. *Analytical Chemistry 69*, 23 (Dec. 1997), 4751–4760. Publisher: American Chemical Society.

[37] Carass, A., Roy, S., Gherman, A., Reinhold, J. C., Jesson, A., Arbel, T., Maier, O., Handels, H., Ghafoorian, M., Platel, B., Birenbaum, A., Greenspan, H., Pham, D. L., Crainiceanu, C. M., Calabresi, P. A., Prince, J. L., Roncal, W. R. G., Shinohara, R. T., and Oguz, I. Evaluating White Matter Lesion Segmentations with Refined Sørensen-Dice Analysis. *Scientific Reports 10*, 1 (May 2020), 8242. Publisher: Nature Publishing Group.

[38] Cardiff, R. D., Miller, C. H., and Munn, R. J. Manual Hematoxylin and Eosin Staining of Mouse Tissue Sections. *Cold Spring Harbor Protocols 2014*, 6 (Jan. 2014), pdb.prot073411. Publisher: Cold Spring Harbor Laboratory Press.

[39] Chazal, F., and Michel, B. An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. *Frontiers in Artificial Intelligence 4* (Sept. 2021). Publisher: Frontiers.

[40] Chernyavsky, I., Alexandrov, T., Maass, P., and Nikolenko, S. I. A two-step soft segmentation procedure for MALDI imaging mass spectrometry data. *German Conference on Bioinformatics 2012, GCB 2012 12* (2012), 39–48. Publisher: ACM ISBN: 9783939897446.

[41] Cichocki, A., Mandic, D., De Lathauwer, L., Zhou, G., Zhao, Q., Caiafa, C., and Phan, H. A. Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE Signal Processing Magazine 32*, 2 (2015), 145–163. arXiv: 1403.4462.

[42] Clish, C. B. Metabolomics: an emerging but powerful tool for precision medicine. *Molecular Case Studies 1*, 1 (Oct. 2015), a000588.

[43] Cobb, M. 60 years ago, Francis Crick changed the logic of biology. *PLoS Biology 15*, 9 (Sept. 2017). Publisher: PLOS.

[44] Da Silva Xavier, G. The Cells of the Islets of Langerhans. *Journal of Clinical Medicine 7*, 3 (Mar. 2018), 54. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[45] Deepaisarn, S., Tar, P. D., Thacker, N. A., Seepujak, A., and McMahon, A. W. Quantifying biological samples using Linear Poisson Independent Component Analysis for MALDI-ToF mass spectra. *Bioinformatics 34*, 6 (Mar. 2018), 1001–1008. Publisher: Oxford University Press.

[46] Deininger, S.-O., Cornett, D. S., Paape, R., Becker, M., Pineau, C., Rauser, S., Walch, A., and Wolski, E. Normalization in MALDI-TOF imaging datasets of proteins: practical considerations. *Analytical and Bioanalytical Chemistry 401*, 1 (July 2011), 167–181.

[47] Deininger, S.-O., Ebert, M. P., Fütterer, A., Gerhard, M., and Röcken, C. MALDI Imaging Combined with Hierarchical Clustering as a New Tool for the Interpretation of Complex Human Cancers. *Journal of Proteome Research 7*, 12 (Dec. 2008), 5230–5236. Publisher: American Chemical Society.

[48] Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. Capturing Chromosome Conformation. *Science 295*, 5558 (Feb. 2002), 1306–1311. Publisher: American Association for the Advancement of Science.

[49] Deltour, L., Leduque, P., Blume, N., Madsen, O., Dubois, P., Jami, J., and Bucchini, D. Differential expression of the two nonallelic proinsulin genes in the developing mouse embryo. *Proceedings of the National Academy of Sciences 90*, 2 (Jan. 1993), 527–531. Publisher: Proceedings of the National Academy of Sciences.

[50] Derwae, H., Nijs, M., Geysels, A., Waelkens, E., and De Moor, B. Spatiochemical Characterization of the Pancreas Using Mass Spectrometry Imaging and Topological Data Analysis. *Analytical Chemistry 95*, 28 (July 2023), 10550–10556. Publisher: American Chemical Society.

[51] Dettmer, K., Aronov, P. A., and Hammock, B. D. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews 26*, 1 (2007), 51–78.

[52] Dexter, A., Race, A. M., Steven, R. T., Barnes, J. R., Hulme, H., Goodwin, R. J. A., Styles, I. B., and Bunch, J. Two-Phase and Graph-Based Clustering Methods for Accurate and Efficient Segmentation of Large Mass Spectrometry Images. *Analytical Chemistry 89*, 21 (Nov. 2017), 11293–11300. Publisher: American Chemical Society.

[53] Dill, A. L., Eberlin, L. S., Zheng, C., Costa, A. B., Ifa, D. R., Cheng, L., Masterson, T. A., Koch, M. O., Vitek, O., and Cooks, R. G. Multivariate statistical differentiation of renal cell carcinomas based on lipidomic analysis by ambient ionization imaging mass spectrometry. *Analytical and Bioanalytical Chemistry 398*, 7 (Dec. 2010), 2969–2978.

[54] Dilmetz, B. A., Lee, Y.-R., Condina, M. R., Briggs, M., Young, C., Desire, C. T., Klingler-Hoffmann, M., and Hoffmann, P. Novel technical developments in mass spectrometry imaging in 2020: A mini review. *Analytical Science Advances 2*, 3-4 (2021), 225–237. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ansa.202000176.

[55] Ding, C., Li, T., and Peng, W. On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing. *Computational Statistics and Data Analysis 52*, 8 (2008), 3913–3927.

[56] Disha, M. Post by Disha Mukherjee | LinkedIn.

[57] Doerr, A. Mass spectrometry imaging takes off. *Nature Methods 15*, 1 (Jan. 2018), 32–32. Publisher: Nature Publishing Group.

[58] Dolenšek, J., Rupnik, M. S., and Stožer, A. Structural similarities and differences between the human and the mouse pancreas. *Islets 7*, 1 (Jan. 2015), e1024405. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19382014.2015.1024405.

[59] Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D., and Dekker, J. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research 16*, 10 (Jan. 2006), 1299–1309. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[60] Duponchel, L. When remote sensing meets topological data analysis. *Journal of Spectral Imaging 7* (Feb. 2018). Publisher: IM Publications Open.

[61] Duraiyan, J., Govindarajan, R., Kaliyappan, K., and Palanisamy, M. Applications of immunohistochemistry. *Journal of Pharmacy and Bioallied Sciences 4*, Suppl 2 (Aug. 2012), S307.

[62] Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., and Aiden, E. L. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems 3*, 1 (July 2016), 95–98.

[63] Eberlin, L. S., Ferreira, C. R., Dill, A. L., Ifa, D. R., Cheng, L., and Cooks, R. G. Non-Destructive, Histologically Compatible Tissue Imaging by Desorption Electrospray Ionization Mass Spectrometry. *Chembiochem : a European journal of chemical biology 12*, 14 (Sept. 2011), 2129–2132.

[64] Eberlin, L. S., Margulis, K., Planell-Mendez, I., Zare, R. N., Tibshirani, R., Longacre, T. A., Jalali, M., Norton, J. A., and Poultsides, G. A. Pancreatic Cancer Surgical Resection Margins: Molecular Assessment by Mass Spectrometry Imaging. *PLOS Medicine 13*, 8 (Aug. 2016), e1002108. Publisher: Public Library of Science.

[65] Eisenstein, M. Seven technologies to watch in 2022. *Nature 601*, 7894 (Jan. 2022), 658–661. Bandiera_abtest: a Cg_type: Technology Feature Number: 7894 Publisher: Nature Publishing Group Subject_term: Biological techniques, Genomics, Genetics, Gene therapy, Quantum information.

[66] Eisenstein, M. Seven technologies to watch in 2023. *Nature 613*, 7945 (Jan. 2023), 794–797. Bandiera_abtest: a Cg_type: Technology Feature Number: 7945 Publisher: Nature Publishing Group Subject_term: Biological techniques, Astronomy and astrophysics, Archaeology.

[67] Esposito, F., Gillis, N., Nicoletta, · ., Buono, D., and Buono, N. D. Mathematical Biology Orthogonal joint sparse NMF for microarray data analysis. *Journal of Mathematical Biology 79* (2019), 223–247.

[68] Ester, M., Kriegel, H.-P., and Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.

[69] Faleiros, T. D. P., and Lopes, A. D. A. On the equivalence between algorithms for Non-negative Matrix Factorization and Latent Dirichlet Allocation. Tech. rep., 2016. Publication Title: ESANN 2016 - 24th European Symposium on Artificial Neural Networks ISBN: 9782875870278.

[70] Févotte, C., and Idier, J. Algorithms for nonnegative matrix factorization with the β-divergence. Tech. rep., 2011. arXiv: 1010.1763

Publication Title: Neural Computation Volume: 23 Issue: 9 ISSN: 1530888X.

[71] Fischer, A. H., Jacobson, K. A., Rose, J., and Zeller, R. Hematoxylin and Eosin Staining of Tissue and Cell Sections. *Cold Spring Harbor Protocols 2008*, 5 (Jan. 2008), pdb.prot4986. Publisher: Cold Spring Harbor Laboratory Press.

[72] Frosini, P. A distance for similarity classes of submanifolds of a Euclidean space. *Bulletin of the Australian Mathematical Society 42*, 3 (Dec. 1990), 407–415.

[73] Frosini, P. Discrete computation of size functions. *Journal of Combinatorics, Information & System Sciences 17*, 3-4 (1992), 232–250.

[74] Frosini, P. Measuring shapes by size functions. In *Intelligent Robots and Computer Vision X: Algorithms and Techniques* (Feb. 1992), vol. 1607, SPIE, pp. 122–133.

[75] Galli, M., Zoppis, I., De Sio, G., Chinello, C., Pagni, F., Magni, F., and Mauri, G. A Support Vector Machine Classification of Thyroid Bioptic Specimens Using MALDI-MSI Data. *Advances in Bioinformatics 2016* (2016), 3791214.

[76] Gardner, W., Maliki, R., Cutts, S. M., Muir, B. W., Ballabio, D., Winkler, D. A., and Pigram, P. J. Self-Organizing Map and Relational Perspective Mapping for the Accurate Visualization of High-Dimensional Hyperspectral Data. *Analytical Chemistry 92*, 15 (Aug. 2020), 10450–10459. Publisher: American Chemical Society.

[77] Gaussier, E., and Goutte, C. Relation between PLSA and NMF and implications. *SIGIR 2005 - Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, August (2005), 601–602. ISBN: 1595930345.

[78] Gavish, M., and Donoho, D. L. The optimal hard threshold for singular values is 4/?3. *IEEE Transactions on Information Theory 60*, 8 (2014), 5040–5053. arXiv: 1305.5870.

[79] Gibbs, R. A. The Human Genome Project changed everything. *Nature Reviews Genetics 21*, 10 (Oct. 2020), 575–576. Publisher: Nature Publishing Group.

[80] Gillis, N. The Why and How of Nonnegative Matrix Factorization. *Regularization, Optimization, Kernels, and Support Vector Machines*, June (2014), 257–291. arXiv: 1401.5226 ISBN: 9781482241402.

[81] Gillis, N. *Nonnegative Matrix Factorization*. Society for Industrial and Applied Mathematics, Jan. 2020. Publication Title: Nonnegative Matrix Factorization.

[82] Gonçalves, J. P. L., Bollwein, C., and Schwamborn, K. Mass Spectrometry Imaging Spatial Tissue Analysis toward Personalized Medicine. *Life (Basel, Switzerland) 12*, 7 (July 2022), 1037.

[83] González-Fernández, A., Dexter, A., Nikula, C. J., and Bunch, J. NECTAR: A New Algorithm for Characterizing and Correcting Noise in QToF-Mass Spectrometry Imaging Data. *Journal of the American Society for Mass Spectrometry 34*, 11 (Nov. 2023), 2443–2453. Publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved.

[84] 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., and Siveke, J. T. MALDI Imaging Mass Spectrometry for In Situ Proteomic Analysis of Preneoplastic Lesions in Pancreatic Cancer. *PLOS ONE 7*, 6 (June 2012), e39424. Publisher: Public Library of Science.

[85] Guo, L., Xie, C., Miao, R., Xu, J., Xu, X., Fang, J., Wang, X., Liu, W., Liao, X., Wang, J., Dong, J., and Cai, Z. DeepION: A Deep Learning-Based Low-Dimensional Representation Model of Ion Images for Mass Spectrometry Imaging. *Analytical Chemistry* (Feb. 2024). Publisher: American Chemical Society.

[86] Gürsoy, G., Xu, Y., Kenter, A. L., and Liang, J. Computational construction of 3D chromatin ensembles and prediction of functional interactions of alpha-globin locus from 5C data. *Nucleic Acids Research 45*, 20 (Nov. 2017), 11547–11558. Publisher: Oxford Academic.

[87] Gut, Y., Boiret, M., Bultel, L., Renaud, T., Chetouani, A., Hafiane, A., Ginot, Y.-M., and Jennane, R. Application of chemometric algorithms to MALDI mass spectrometry imaging of pharmaceutical tablets. *Journal of Pharmaceutical and Biomedical Analysis 105* (Feb. 2015), 91–100.

[88] Hanselmann, M., Kirchner, M., Renard, B. Y., Amstalden, E. R., Glunde, K., Heeren, R. M., and Hamprecht, F. A. Concise representation of mass spectrometry images by probabilistic latent semantic analysis. *Analytical Chemistry 80*, 24 (Dec. 2008), 9649–9658. Publisher: American Chemical Society.

[89] Hanselmann, M., Köthe, U., Kirchner, M., Renard, B. Y., Amstalden, E. R., Glunde, K., Heeren, R. M. A., and Hamprecht, F. A. Toward Digital Staining using Imaging Mass Spectrometry and Random Forests. *Journal of Proteome Research 8*, 7 (July 2009), 3558–3567. Publisher: American Chemical Society.

[90] Harn, Y. C., Powers, M. J., Shank, E. A., and Jojic, V. Deconvolving molecular signatures of interactions between microbial colonies. In *Bioinformatics* (June 2015), vol. 31, Oxford University Press, pp. i142–i150. Issue: 12 ISSN: 14602059.

[91] Heiles, S. Advanced tandem mass spectrometry in metabolomics and lipidomics—methods and applications. *Analytical and Bioanalytical Chemistry 413*, 24 (Oct. 2021), 5927–5948.

[92] Hien, L. T. K., and Gillis, N. Algorithms for Nonnegative Matrix Factorization with the Kullback–Leibler Divergence. *Journal of Scientific Computing 87*, 3 (2021), 1–30. arXiv: 2010.01935.

[93] Hoef, L. V., Adams, H., King, E. J., and Ebert-Uphoff, I. A Primer on Topological Data Analysis to Support Image Analysis Tasks in Environmental Science. *Artificial Intelligence for the Earth Systems 2*, 1 (Feb. 2023). Publisher: American Meteorological Society Section: Artificial Intelligence for the Earth Systems.

[94] Hoffman, M. D., Blei, D. M., and Bach, F. Online learning for Latent Dirichlet Allocation. Tech. rep., 2010. Publication Title: Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010 ISBN: 9781617823800.

[95] Hofmann, T. Probabilistic Latent Semantic Analysis. Tech. rep., 2013. arXiv: 1301.6705.

[96] Horgan, R. P., and Kenny, L. C. 'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist 13*, 3 (2011), 189–195. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1576/toag.13.3.189.27672.

[97] Hu, H., Bindu, J. P., and Laskin, J. Self-supervised clustering of mass spectrometry imaging data using contrastive learning. *Chemical Science 13*, 1 (Jan. 2022), 90–98. Publisher: Royal Society of Chemistry.

[98] Hu, H., and Laskin, J. Emerging Computational Methods in Mass Spectrometry Imaging. *Advanced Science* (2022), 2203339. Publisher: John Wiley & Sons, Ltd.

[99] Hu, H., Yin, R., Brown, H. M., and Laskin, J. Spatial Segmentation of Mass Spectrometry Imaging Data by Combining Multivariate Clustering and Univariate Thresholding. *Analytical Chemistry 93*, 7 (Feb. 2021), 3477–3485. Publisher: American Chemical Society.

[100] Hubert, L., and Arabie, P. Comparing partitions. *Journal of Classification 2*, 1 (Dec. 1985), 193–218.

[101] Ilin, Y., and Kraft, M. L. Secondary ion mass spectrometry and Raman spectroscopy for tissue engineering applications. *Current Opinion in Biotechnology 31* (Feb. 2015), 108–116.

[102] Jetybayeva, A., Borodinov, N., Ievlev, A. V., Haque, M. I. U., Hinkle, J., Lamberti, W. A., Meredith, J. C., Abmayr, D., and Ovchinnikova, O. S. A review on recent machine learning applications for imaging mass spectrometry studies. *Journal of Applied Physics 133*, 2 (Jan. 2023), 020702.

[103] Jones, E. A., van Remoortere, A., van Zeijl, R. J., Hogendoorn, P. C., Bovée, J. V., Deelder, A. M., and McDonnell, L. A. Multiple statistical analysis techniques corroborate intratumor heterogeneity in imaging mass spectrometry datasets of myxofibrosarcoma. *PLoS ONE 6*, 9 (Sept. 2011).

[104] Jørgensen, M. C., Ahnfelt-Rønne, J., Hald, J., Madsen, O. D., Serup, P., and Hecksher-Sørensen, J. An Illustrated Review of Early Pancreas Development in the Mouse. *Endocrine Reviews 28*, 6 (Oct. 2007), 685–705.

[105] Keenan, M. R., and Kotula, P. G. Accounting for Poisson noise in the multivariate analysis of ToF-SIMS spectrum images. *Surface and Interface Analysis 36*, 3 (Mar. 2004), 203–212.

[106] Kim, H.-J., Yardimci, G. G., Bonora, G., Ramani, V., Liu, J., Qiu, R., Lee, C., Hesson, J., Ware, C. B., Shendure, J., Duan, Z., and Noble, W. S. Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell Hi-C data. *PLOS Computational Biology 16*, 9 (Sept. 2020), e1008173.

[107] Klaila, G., Vutov, V., and Stefanou, A. Supervised topological data analysis for MALDI mass spectrometry imaging applications. *BMC Bioinformatics 24*, 1 (July 2023), 279.

[108] Kolda, T. G., and Bader, B. W. Tensor Decompositions and Applications. *SIAM Review 51*, 3 (Aug. 2009), 455–500.

[109] Krassowski, M., Das, V., Sahu, S. K., and Misra, B. B. State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. *Frontiers in Genetics 11* (2020).

[110] Kriegsmann, M., Casadonte, R., Kriegsmann, J., Dienemann, H., Schirmacher, P., Kobarg, J. H., Schwamborn, K., Stenzinger, A., Warth, A., and Weichert, W. Reliable Entity Subtyping in Non-small Cell Lung Cancer by Matrix-assisted Laser Desorption/Ionization Imaging Mass Spectrometry on Formalin-fixed Paraffin-embedded Tissue Specimens *. *Molecular & Cellular Proteomics 15*, 10 (Oct. 2016), 3081–3089. Publisher: Elsevier.

[111] Kukurba, K. R., and Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harbor Protocols 2015*, 11 (Jan. 2015), pdb.top084970. Publisher: Cold Spring Harbor Laboratory Press.

[112] Lajoie, B. R., Dekker, J., and Kaplan, N. The Hitchhiker's Guide to Hi-C Analysis: Practical guidelines. *Methods (San Diego, Calif.) 72*, C (2015), 65. Publisher: NIH Public Access.

[113] Lang, Y., Zhou, L., and Imamura, Y. Development of Machine-Learning Techniques for Time-of-Flight Secondary Ion Mass Spectrometry Spectral Analysis: Application for the Identification of Silane Coupling Agents in Multicomponent Films. *Analytical Chemistry 94*, 5 (Feb. 2022), 2546–2553. Publisher: American Chemical Society.

[114] Lathauwer, L. D., Moor, B. D., Member, S., and Vandewalle, J. Independent Component Analysis and (Simultaneous). *IEEE Transactions on Signal Processing 49*, 10 (2001), 2262–2271.

[115] Ledford, H. The race to map the human body — one cell at a time. *Nature 542*, 7642 (Feb. 2017), 404–405. Publisher: Nature Publishing Group.

[116] Lee, D. D., and Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature 401*, 6755 (Oct. 1999), 788–791. Publisher: Nature Publishing Group.

[117] Leopold, J., Popkova, Y., Engel, K. M., and Schiller, J. Recent Developments of Useful MALDI Matrices for the Mass Spectrometric Characterization of Lipids. *Biomolecules 8*, 4 (Dec. 2018), 173.

[118] Leuschner, J., Schmidt, M., Fernsel, P., Lachmund, D., Boskamp, T., and Maass, P. Supervised non-negative matrix factorization methods for MALDI imaging applications. *Bioinformatics 35*, 11 (2019), 1940–1947.

[119] Li, C., Chu, S., Tan, S., Yin, X., Jiang, Y., Dai, X., Gong, X., Fang, X., and Tian, D. Towards Higher Sensitivity of Mass Spectrometry: A Perspective From the Mass Analyzers. *Frontiers in Chemistry 9* (2021), 813359.

[120] Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science 326*, 5950 (Oct. 2009), 289–293. Publisher: American Association for the Advancement of Science.

[121] Liebl, H. Ion Microprobe Mass Analyzer. *Journal of Applied Physics 38*, 13 (Dec. 1967), 5277–5283.

[122] Lin, Y.-W. E., Shnitzer, T., Talmon, R., Villarroel-Espindola, F., Desai, S., Schalper, K., and Kluger, Y. Graph of graphs analysis for multiplexed data with application to imaging mass cytometry. *PLOS Computational Biology 17*, 3 (2021), e1008741. Publisher: Public Library of Science.

[123] Liu, J. X., Wang, D., Gao, Y. L., Zheng, C. H., Xu, Y., and Yu, J. Regularized Non-Negative Matrix Factorization for Identifying Differentially Expressed Genes and Clustering Samples: A Survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 15*, 3 (May 2018), 974–987. Publisher: Institute of Electrical and Electronics Engineers Inc.

[124] Liu, Q., Zeng, W., Zhang, W., Wang, S., Chen, H., Jiang, R., Zhou, M., and Zhang, S. Deep generative modeling and clustering of single cell Hi-C data. *Briefings in Bioinformatics 24*, 1 (Jan. 2023), bbac494.

[125] Lloyd, S. P. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory 28*, 2 (1982), 129–137.

[126] López-Fernández, H., Reboiro-Jato, M., Madeira, S. C., López-Cortés, R., Nunes-Miranda, J. D., Santos, H. M., Fdez-Riverola, F., and 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* (Heidelberg, 2013), M. S. Mohamad, L. Nanni, M. P. Rocha, and F. Fdez-Riverola, Eds., Advances in Intelligent Systems and Computing, Springer International Publishing, pp. 145–153.

[127] López-Fernández, O., Domínguez, R., Pateiro, M., Munekata, P. E. S., Rocchetti, G., and Lorenzo, J. M. Determination of Polyphenols Using Liquid Chromatography–Tandem Mass Spectrometry Technique (LC–MS/MS): A Review. *Antioxidants 9*, 6 (June 2020), 479. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[128] Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. Transcriptomics technologies. *PLoS Computational Biology 13*, 5 (May 2017), e1005457.

[129] Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., and Carlsson, G. Extracting insights from the shape of complex data using topology. *Scientific Reports 3*, 1 (Feb. 2013), 1236. Publisher: Nature Publishing Group.

[130] Lyu, H., Liu, E., and Wu, Z. Comparison of normalization methods for Hi-C data. *BioTechniques 68*, 2 (Oct. 2020), 56–64. Publisher: Future Science Ltd.

[131] Ma, S., and Dai, Y. Principal component analysis based Methods in bioinformatics studies. *Briefings in Bioinformatics 12*, 6 (2011), 714–722.

[132] Madeira, S., and Oliveira, A. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics 1*, 1 (Jan. 2004), 24–45. Conference Name: IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[133] Mann, M., Hendrickson, R. C., and Pandey, A. Analysis of proteins and proteomes by mass spectrometry. *Annual Review of Biochemistry 70* (2001), 437–473.

[134] Mas, S., Torro, A., Bec, N., Fernández, L., Erschov, G., Gongora, C., Larroque, C., Martineau, P., de Juan, A., and Marco, S. Use of physiological information based on grayscale images to improve mass spectrometry imaging data analysis from biological tissues. *Analytica Chimica Acta 1074* (Oct. 2019), 69–79.

[135] McInnes, L., Healy, J., and Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv: 1802.03426.

[136] Meding, S., Nitsche, U., Balluff, B., Elsner, M., Rauser, S., Schöne, C., Nipp, M., Maak, M., Feith, M., Ebert, M. P., Friess, H., Langer, R., Höfler, H., Zitzelsberger, H., Rosenberg, R., and Walch, A. Tumor Classification of Six Common Cancer Types

Based on Proteomic Profiling by MALDI Imaging. *Journal of Proteome Research 11*, 3 (Mar. 2012), 1996–2003. Publisher: American Chemical Society.

[137] Mirkin, B. *Mathematical Classification and Clustering*, vol. 11 of *Nonconvex Optimization and Its Applications*. Springer US, Boston, MA, 1996.

[138] Miura, D., Fujimura, Y., and Wariishi, H. *In situ* metabolomic mass spectrometry imaging: Recent advances and difficulties. *Journal of Proteomics 75*, 16 (Aug. 2012), 5052–5060.

[139] Moore, J. L., and Charkoftaki, G. A Guide to MALDI Imaging Mass Spectrometry for Tissues. *Journal of Proteome Research 22*, 11 (Nov. 2023), 3401–3417. Publisher: American Chemical Society.

[140] Moroni, D., and Pascali, M. A. Learning Topology: Bridging Computational Topology and Machine Learning. *Pattern Recognition and Image Analysis 31*, 3 (July 2021), 443–453.

[141] Muhlich, J. L., Chen, Y.-A., Yapp, C., Russell, D., Santagata, S., and Sorger, P. K. Stitching and registering highly multiplexed whole-slide images of tissues and tumors using ASHLAR. *Bioinformatics 38*, 19 (Sept. 2022), 4613–4621.

[142] Munch, E. A User's Guide to Topological Data Analysis. *Journal of Learning Analytics 4*, 2 (July 2017), 47–61. Number: 2.

[143] Murayama, C., Kimura, Y., and Setou, M. Imaging mass spectrometry: principle and application. *Biophysical Reviews 1*, 3 (Sept. 2009), 131–139.

[144] Murta, T., Steven, R. T., Nikula, C. J., Thomas, S. A., Zeiger, L. B., Dexter, A., Elia, E. A., Yan, B., Campbell, A. D., Goodwin, R. J. A., Takáts, Z., Sansom, O. J., and Bunch, J. Implications of Peak Selection in the Interpretation of Unsupervised Mass Spectrometry Imaging Data Analyses. *Analytical Chemistry 93*, 4 (Feb. 2021), 2309–2316. Publisher: American Chemical Society.

[145] Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., and Fraser, P. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature 502*, 7469 (Oct. 2013), 59–64. Number: 7469 Publisher: Nature Publishing Group.

[146] Nathaniel Saul, C. T. Scikit-TDA: Topological Data Analysis for Python, 2019.

[147] Negi, A., Shukla, A., Jaiswar, A., Shrinet, J., and Jasrotia, R. S. Chapter 6 - Applications and challenges of microarray and RNA-sequencing. In *Bioinformatics*, D. B. Singh and R. K. Pathak, Eds. Academic Press, Jan. 2022, pp. 91–103.

[148] Nijs, M., Smets, T., Waelkens, E., and 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 35*, 21 (Nov. 2021), e9181. Publisher: John Wiley & Sons, Ltd.

[149] Nijs, M., Waelkens, E., and De Moor, B. Hierarchical bi-clustering of mouse pancreas mass spectrometry imaging data using recursive rank-2 non-negative matrix factorization. Submitted to Journal of the American Society for Mass Spectrometry.

[150] Offroy, M., and Duponchel, L. Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry. *Analytica Chimica Acta 910* (Mar. 2016), 1–11.

[151] Ovchinnikova, K., Kovalev, V., Stuart, L., and Alexandrov, T. OffsampleAI: artificial intelligence approach to recognize off-sample mass spectrometry images. *BMC Bioinformatics 21*, 1 (Apr. 2020), 129.

[152] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., and Cournapeau, D. Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*.

[153] Piehowski, P. D., Davey, A. M., Kurczy, M. E., Sheets, E. D., Winograd, N., Ewing, A. G., and Heien, M. L. Time-of-flight secondary ion mass spectrometry imaging of subcellular lipid heterogeneity: Poisson counting and spatial resolution. *Analytical Chemistry 81*, 14 (July 2009), 5593–5602. Publisher: American Chemical Society.

[154] Prasad, M., Postma, G., Franceschi, P., Buydens, L. M. C., and Jansen, J. J. Evaluation and comparison of unsupervised methods for the extraction of spatial patterns from mass spectrometry imaging data (MSI). *Scientific Reports 12*, 1 (Sept. 2022), 15687. Publisher: Nature Publishing Group.

[155] 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., and Caprioli, R. M. Imaging mass spectrometry enables molecular profiling of mouse and human pancreatic tissue. *Diabetologia 62*, 6 (June 2019), 1036–1047.

[156] Rabadán, R., and Blumberg, A. J. *Topological Data Analysis for Genomics and Evolution: Topology in Biology*. Cambridge University Press, Oct. 2019. Google-Books-ID: BADDDwAAQBAJ.

[157] Rabilloud, T., Chevallet, M., Luche, S., and Lelong, C. Two-dimensional gel electrophoresis in proteomics: Past, present and future. *Journal of Proteomics 73*, 11 (Oct. 2010), 2064–2077.

[158] Ràfols, P., Vilalta, D., Brezmes, J., Cañellas, N., Del Castillo, E., Yanes, O., Ramírez, N., and Correig, X. Signal preprocessing, multivariate analysis and software tools for MA(LDI)-TOF mass spectrometry imaging for biological applications. *Mass Spectrometry Reviews 37*, 3 (May 2018), 281–306.

[159] Ramos-Vara, J. A., and Miller, M. A. When tissue antigens and antibodies get along: revisiting the technical aspects of immunohistochemistry–the red, brown, and blue technique. *Veterinary Pathology 51*, 1 (Jan. 2014), 42–87.

[160] Reder, G. K., Young, A., Altosaar, J., Rajniak, J., Elhadad, N., Fischbach, M., and Holmes, S. Supervised topic modeling for predicting molecular substructure from mass spectrometry. *F1000Research 10* (May 2021), 403. Publisher: F1000 Research Limited.

[161] Roderigo-Milne, H., Hauge-Evans, A. C., Persaud, S. J., and Jones, P. M. Differential expression of insulin genes 1 and 2 in MIN6 cells and pseudoislets. *Biochemical and Biophysical Research Communications 296*, 3 (Aug. 2002), 589–595.

[162] Rowley, M. J., and Corces, V. G. Organizational principles of 3D genome architecture. *Nature Reviews Genetics 19*, 12 (Dec. 2018), 789–800. Number: 12 Publisher: Nature Publishing Group.

[163] Rozanova, S., Barkovits, K., Nikolov, M., Schmidt, C., Urlaub, H., and Marcus, K. Quantitative Mass Spectrometry-Based Proteomics: An Overview. In *Quantitative Methods in Proteomics*, K. Marcus, M. Eisenacher, and B. Sitek, Eds., Methods in Molecular Biology. Springer US, New York, NY, 2021, pp. 85–116.

[164] See, Y. X., Wang, B. Z., and Fullwood, M. J. Chromatin Interactions and Regulatory Elements in Cancer: From Bench to Bedside. *Trends in Genetics 35*, 2 (Feb. 2019), 145–158.

[165] Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., and Waterston, R. H. DNA sequencing at 40: past, present and future. *Nature 550*, 7676 (Oct. 2017), 345–353. Publisher: Nature Publishing Group.

[166] Shi, Y., Xiang, R., Horváth, C., and Wilkins, J. A. The role of liquid chromatography in proteomics. *Journal of Chromatography A 1053*, 1 (Oct. 2004), 27–36.

[167] Shiao, M.-S., Liao, B.-Y., Long, M., and Yu, H.-T. Adaptive Evolution of the Insulin Two-Gene System in Mouse. *Genetics 178*, 3 (Mar. 2008), 1683–1691.

[168] Sidiropoulos, N. D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E. E., and Faloutsos, C. Tensor Decomposition for Signal Processing and Machine Learning. *IEEE Transactions on Signal Processing 65*, 13 (July 2017), 3551–3582. Conference Name: IEEE Transactions on Signal Processing.

[169] Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nature Genetics 38*, 11 (Nov. 2006), 1348–1354. Number: 11 Publisher: Nature Publishing Group.

[170] Singh, G., Mémoli, F., and Carlsson, G. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition.

[171] Singh, Y., Farrelly, C. M., Hathaway, Q. A., Leiner, T., Jagtap, J., Carlsson, G. E., and Erickson, B. J. Topological data analysis in medical imaging: current state of the art. *Insights into Imaging 14*, 1 (Apr. 2023), 58.

[172] Siy, P. W., Moffitt, R. A., Parry, R. M., Chen, Y., Liu, Y., Sullards, M. C., Merrill, A. H., and Wang, M. D. Matrix factorization techniques for analysis of imaging mass spectrometry data. *8th IEEE International Conference on BioInformatics and BioEngineering, BIBE 2008* (2008), 1–6. ISBN: 9781424428458.

[173] Skaf, Y., and Laubenbacher, R. Topological data analysis in biomedicine: A review. *Journal of Biomedical Informatics 130* (June 2022), 104082.

[174] Smets, T., Verbeeck, N., Claesen, M., Asperger, A., Griffioen, G., Tousseyn, T., Waelput, W., Waelkens, E., and De Moor, B.

Evaluation of Distance Metrics and Spatial Autocorrelation in Uniform Manifold Approximation and Projection Applied to Mass Spectrometry Imaging Data. *Analytical Chemistry 91*, 9 (2019), 5706–5714.

[175] Smets, T., Waelkens, E., and 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 92*, 7 (Apr. 2020), 5240–5248. Publisher: American Chemical Society.

[176] Song, J.-W., Lam, S. M., Fan, X., Cao, W.-J., Wang, S.-Y., Tian, H., Chua, G. H., Zhang, C., Meng, F.-P., Xu, Z., Fu, J.-L., Huang, L., Xia, P., Yang, T., Zhang, S., Li, B., Jiang, T.-J., Wang, R., Wang, Z., Shi, M., Zhang, J.-Y., Wang, F.-S., and Shui, G. Omics-Driven Systems Interrogation of Metabolic Dysregulation in COVID-19 Pathogenesis. *Cell Metabolism 32*, 2 (Aug. 2020), 188–202.e5.

[177] Sorber, L., Van Barel, M., and De Lathauwer, L. Structured Data Fusion. *IEEE Journal of Selected Topics in Signal Processing 9*, 4 (June 2015), 586–600. Conference Name: IEEE Journal of Selected Topics in Signal Processing.

[178] Stone, G., Clifford, D., Gustafsson, J. O., McColl, S. R., and Hoffmann, P. Visualisation in imaging mass spectrometry using the minimum noise fraction transform. *BMC Research Notes 5*, 1 (Aug. 2012), 419.

[179] Sznurkowska, M. K., Hannezo, E., Azzarelli, R., Chatzeli, L., Ikeda, T., Yoshida, S., Philpott, A., and Simons, B. D. Tracing the cellular basis of islet specification in mouse pancreas. *Nature Communications 11*, 1 (Oct. 2020), 5037. Number: 1 Publisher: Nature Publishing Group.

[180] 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., and Miyazono, K. An analysis modality for vascular structures combining tissue-clearing technology and topological data analysis. *Nature Communications 13*, 1 (Sept. 2022), 5239. Publisher: Nature Publishing Group.

[181] Tamara, S., den Boer, M. A., and Heck, A. J. R. High-Resolution Native Mass Spectrometry. *Chemical Reviews 122*, 8 (Apr. 2022), 7269–7326. Publisher: American Chemical Society.

[182] Tan, L., Ma, W., Wu, H., Zheng, Y., Xing, D., Chen, R., Li, X., Daley, N., Deisseroth, K., and Xie, X. S. Changes in genome

architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development. *Cell 184*, 3 (Feb. 2021), 741–758.e17. Publisher: Cell Press.

[183] Tan, L., Xing, D., Chang, C. H., Li, H., and Xie, X. S. Three-dimensional genome structures of single diploid human cells. *Science 361*, 6405 (Aug. 2018), 924–928. Publisher: American Association for the Advancement of Science.

[184] Tan, L., Xing, D., Daley, N., and Xie, X. S. Three-dimensional genome structures of single sensory neurons in mouse visual and olfactory systems. *Nature Structural & Molecular Biology 2019 26:4 26*, 4 (Apr. 2019), 297–307. Publisher: Nature Publishing Group.

[185] Tang, L. Spatially resolved multiomics. *Nature Methods 20*, 12 (Dec. 2023), 1871–1871. Number: 12 Publisher: Nature Publishing Group.

[186] Tauzin, G., Lupo, U., Tunstall, L., Pérez, J. B., Caorsi, M., Medina-Mardones, A. M., Dassatti, A., and Hess, K. giotto-tda: : A Topological Data Analysis Toolkit for Machine Learning and Data Exploration. *Journal of Machine Learning Research 22*, 39 (2021), 1–6.

[187] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research 51*, D1 (Jan. 2023), D523–D531.

[188] Thomas, S. A., Jin, Y., Bunch, J., and Gilmore, I. S. Enhancing classification of mass spectrometry imaging data with deep neural networks. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)* (Nov. 2017), pp. 1–8.

[189] Thomas, S. A., Race, A. M., Steven, R. T., Gilmore, I. S., and Bunch, J. Dimensionality reduction of mass spectrometry imaging data using autoencoders. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)* (Dec. 2016), pp. 1–7.

[190] Tideman, L. E. M., Migas, L. G., Djambazova, K. V., Patterson, N. H., Caprioli, R. M., Spraggins, J. M., and Van de Plas, R. Automated biomarker candidate discovery in imaging mass spectrometry data through spatially localized Shapley additive explanations. *Analytica Chimica Acta 1177* (Sept. 2021), 338522.

[191] Tillner, J., Wu, V., Jones, E. A., Pringle, S. D., Karancsi, T., Dannhorn, A., Veselkov, K., McKenzie, J. S., and Takats, Z. Faster, More Reproducible DESI-MS for Biological Tissue Imaging. *Journal of the American Society for Mass Spectrometry 28*, 10 (Oct. 2017),

2090–2098. Publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved.

[192] Treu, A., and Römpp, A. Matrix ions as internal standard for high mass accuracy matrix-assisted laser desorption/ionization mass spectrometry imaging. *Rapid Communications in Mass Spectrometry 35*, 16 (2021), e9110. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rcm.9110.

[193] Ubaru, S., and Saad, Y. Fast methods for estimating the numerical rank of large matrices. Tech. rep., 2016. Publication Title: 33rd International Conference on Machine Learning, ICML 2016 Volume: 1 ISBN: 9781510829008.

[194] Ucal, Y., Coskun, A., and Ozpinar, A. Quality will determine the future of mass spectrometry imaging in clinical laboratories: the need for standardization. *Expert Review of Proteomics 16*, 6 (June 2019), 521–532. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14789450.2019.1624165.

[195] Van De Plas, R., Ojeda, F., Dewil, M., Van Den Bosch, L., De Moor, B., and Waelkens, E. Prospective exploration of biochemical tissue composition via imaging mass spectrometry guided by principal component analysis. *Pacific Symposium on Biocomputing 2007, PSB 2007* (2007), 458–469. ISBN: 9812704175.

[196] Van Der Hooft, J. J., Wandy, J., Young, F., Padmanabhan, S., Gerasimidis, K., Burgess, K. E., Barrett, M. P., and Rogers, S. Unsupervised Discovery and Comparison of Structural Families Across Multiple Samples in Untargeted Metabolomics. *Analytical Chemistry 89*, 14 (July 2017), 7569–7577. Publisher: American Chemical Society.

[197] Van Der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E., and Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences of the United States of America 113*, 48 (Nov. 2016), 13738–13743. Publisher: National Academy of Sciences.

[198] Van Der Maaten, L., and Hinton, G. Visualizing data using t-SNE. Tech. rep., 2008. Publication Title: Journal of Machine Learning Research Volume: 9 ISSN: 15324435.

[199] Vandereyken, K., Sifrim, A., Thienpont, B., and Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics 24*, 8 (Aug. 2023), 494–515. Publisher: Nature Publishing Group.

[200] Vejdemo-Johansson, M., and Skraba, P. Topology, Big Data and Optimization. In *Big Data Optimization: Recent Developments and Challenges*, A. Emrouznejad, Ed., Studies in Big Data. Springer International Publishing, Cham, 2016, pp. 147–176.

[201] Verbeeck, N., Caprioli, R. M., and Van de Plas, R. Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. *Mass Spectrometry Reviews 39*, 3 (2020), 245–291.

[202] Verri, A., Uras, C., Frosini, P., and Ferri, M. On the use of size functions for shape analysis. *Biological Cybernetics 70*, 2 (Dec. 1993), 99–107.

[203] Vervliet, N., Debals, O., Sorber, L., Van Barel, M., and De Lathauwer, L. Tensorlab 3.0, Mar. 2016.

[204] Veselkov, K. A., Vingara, L. K., Masson, P., Robinette, S. L., Want, E., Li, J. V., Barton, R. H., Boursier-Neyret, C., Walther, B., Ebbels, T. M., Pelczer, I., Holmes, E., Lindon, J. C., and Nicholson, J. K. Optimized Preprocessing of Ultra-Performance Liquid Chromatography/Mass Spectrometry Urinary Metabolic Profiles for Improved Information Recovery. *Analytical Chemistry 83*, 15 (Aug. 2011), 5864–5872. Publisher: American Chemical Society.

[205] Vijayalakshmi, K., Shankar, V., Bain, R. M., Nolley, R., Sonn, G. A., Kao, C.-S., Zhao, H., Tibshirani, R., Zare, R. N., and Brooks, J. D. Identification of diagnostic metabolic signatures in clear cell renal cell carcinoma using mass spectrometry imaging. *International Journal of Cancer 147*, 1 (2020), 256–265. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ijc.32843.

[206] Von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing 17*, 4 (Dec. 2007), 395–416.

[207] Wadie, B., Stuart, L., Rath, C. M., and Alexandrov, T. METASPACE-ML: Metabolite annotation for imaging mass spectrometry using machine learning. preprint, Bioinformatics, May 2023.

[208] Wall, M. E., Rechtsteiner, A., and Rocha, L. M. Singular Value Decomposition and Principal Component Analysis. In *A Practical Approach to Microarray Data Analysis*, D. P. Berrar, W. Dubitzky, and M. Granzow, Eds. Springer US, Boston, MA, 2003, pp. 91–109.

[209] Wang, Z., Zhang, Y., Tian, R., Luo, Z., Zhang, R., Li, X., and Abliz, Z. Data-Driven Deciphering of Latent Lesions in Heterogeneous

Tissue Using Function-Directed t-SNE of Mass Spectrometry Imaging Data. *Analytical Chemistry 94*, 40 (Oct. 2022), 13927–13935. Publisher: American Chemical Society.

[210] Watrous, J. D., Alexandrov, T., and Dorrestein, P. C. The evolving field of imaging mass spectrometry and its impact on future biological research. *Journal of Mass Spectrometry 46*, 2 (Feb. 2011), 209–222.

[211] Xi, Y., Sohn, A. L., Joignant, A. N., Cologna, S. M., Prentice, B. M., and Muddiman, D. C. SMART: A data reporting standard for mass spectrometry imaging. *Journal of Mass Spectrometry 58*, 2 (Feb. 2023), e4904.

[212] Xie, Q., Han, C., Jin, V., and Lin, S. HiCImpute: A Bayesian hierarchical model for identifying structural zeros and enhancing single cell Hi-C data. *PLoS Computational Biology 18*, 6 (June 2022), e1010129.

[213] Xing, L., Zhao, C.-L., Mou, H.-Z., Pan, J., Kang, B., Chen, H.-Y., and Xu, J.-J. Next Generation of Mass Spectrometry Imaging: from Micrometer to Subcellular Resolution. *Chemical & Biomedical Imaging 1*, 8 (Nov. 2023), 670–682. Publisher: American Chemical Society.

[214] Xiong, X. C., Fang, X., Ouyang, Z., Jiang, Y., Huang, Z. J., and Zhang, Y. K. Feature extraction approach for mass spectrometry imaging data using non-negative matrix factorization. *Fenxi Huaxue/ Chinese Journal of Analytical Chemistry 40*, 5 (2012), 663–669. Publisher: Changchun Institute of Applied Chemistry, Chinese Academy of Sciences.

[215] Xu, G., Gan, S., Guo, B., and Yang, L. Application of clustering strategy for automatic segmentation of tissue regions in mass spectrometry imaging. *Rapid Communications in Mass Spectrometry 38*, 8 (2024), e9717. _eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/rcm.9717.

[216] Yadav, S. K., Sinha, R., and Bora, P. K. An Efficient SVD Shrinkage for Rank Estimation. *IEEE Signal Processing Letters 22*, 12 (Dec. 2015), 2406–2410. Publisher: Institute of Electrical and Electronics Engineers Inc.

[217] Yang, C., Wei, M., Zhao, Y., Yang, Z., Song, M., Mi, J., Yang, X., and Tian, G. Regulation of insulin secretion by the post-translational modifications. *Frontiers in Cell and Developmental Biology 11* (2023).

[218] Zhang, R., Zhou, T., and Ma, J. Ultrafast and interpretable single-cell 3D genome analysis with Fast-Higashi. *Cell Systems 13*, 10 (Oct. 2022), 798–807.e6. Publisher: Elsevier.

[219] Zhang, S., Li, Q., Liu, J., and Zhou, X. J. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules. *Bioinformatics 27*, 13 (2011), 401–409.

[220] Zhang, W., Claesen, M., Moerman, T., Groseclose, M. R., Waelkens, E., De Moor, B., and Verbeeck, N. Spatially aware clustering of ion images in mass spectrometry imaging data using deep learning. *Analytical and Bioanalytical Chemistry 413*, 10 (Apr. 2021), 2803–2819.

[221] Zhang, W., Spotbeen, X., Vanuytven, S., Kint, S., Sarretto, T., Socciarelli, F., Vandereyken, K., Dehairs, J., Idkowiak, J., Wouters, D., Larizgoitia, J. I. A., Partel, G., Ly, A., Laat, V. d., Mantas, M. J. Q., Gevaert, T., Devlies, W., Mah, C. Y., Butler, L. M., Loda, M., Joniau, S., Moor, B. D., Sifrim, A., Ellis, S. R., Voet, T., Claesen, M., Verbeeck, N., and Swinnen, J. V. Integration of Multiple Spatial Omics Modalities Reveals Unique Insights into Molecular Heterogeneity of Prostate Cancer, Aug. 2023. Pages: 2023.08.28.555056 Section: New Results.

[222] Zhang, Z. Y. NMF-based Models for Tumor Clustering: A Systematic Comparison. Tech. rep., 2009.

[223] Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Singh Sandhu, K., Singh, U., Pant, V., Tiwari, V., Kurukuti, S., and Ohlsson, R. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics 38*, 11 (Nov. 2006), 1341–1347. Number: 11 Publisher: Nature Publishing Group.

[224] Zheng, Y., Shen, S., and Keleş, S. Normalization and de-noising of single-cell Hi-C data with BandNorm and scVI-3D. *Genome Biology 23*, 1 (Oct. 2022), 222.

[225] Zhou, B., Xiao, J. F., Tuli, L., and Ressom, H. W. LC-MS-based metabolomics. *Molecular bioSystems 8*, 2 (Feb. 2012), 470–481.

[226] Zhou, J., Ma, J., Chen, Y., Cheng, C., Bao, B., Peng, J., Sejnowski, T. J., Dixon, J. R., and Ecker, J. R. Robust single-cell Hi-C clustering by convolution- and random-walk–based imputation. *Proceedings of the National Academy of Sciences 116*, 28 (July 2019), 14011–14018. Publisher: Proceedings of the National Academy of Sciences.

[227] Zhou, Z., and Zare, R. N. Personal Information from Latent Fingerprints Using Desorption Electrospray Ionization Mass Spectrometry and Machine Learning. *Analytical Chemistry 89*, 2 (Jan. 2017), 1369–1372. Publisher: American Chemical Society.

[228] Zhu, H., and Wang, Z. SCL: a lattice-based approach to infer 3D chromosome structures from single-cell Hi-C data. *Bioinformatics (Oxford, England) 35*, 20 (Oct. 2019), 3981–3988.

List of publications

International journal papers [published]

Nijs, M., Smets, T., Waelkens, E., De Moor, B. (2021). *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, 35, e9181-e9181. doi: 10.1002/rcm.9181.

Derwae, H., Nijs, M., Geysels, A., Waelkens, E., De Moor, B. (2023). *Spatiochemical characterization of the pancreas using mass spectrometry imaging and topological data analysis.* Analytical Chemistry, 95 (28), 10550-10556. doi: 10.1021/acs.analchem.2c05606.

International journal papers [submitted]

Nijs, M., Waelkens, E., De Moor, B. (2024). *Hierarchical bi-clustering of mass spectrometry imaging data using recursive rank-2 non-negative matrix factorization.* Submitted to Journal of the American Society for Mass Spectrometry.

Talks at international conferences

Nijs, M., Smets, T., Waelkens, E., De Moor, B. (2023). *KL-NMF as best suited NMF variant for the analysis of mass spectrometry imaging data.* Presented at the Regional Mass Spectrometry Imaging Spring Workshop, Uppsala, Sweden, 27-29 March, 2023.

Poster presentations at international conferences

Nijs, M., Smets, T., Waelkens, E., De Moor, B. (2023). *KL-NMF as best suited NMF variant for the analysis of mass spectrometry imaging data.* Presented at the Regional Mass Spectrometry Imaging Spring Workshop, Uppsala, Sweden, 27-29 March, 2023.

Nijs, M., Waelkens, E., De Moor, B. (2023). *Hierarchical bi-clustering of mass spectrometry imaging data using recursive rank-2 non-negative matrix factorization.* Presented at the 1st Annual Conference on Mass Spectrometry Imaging and Integrated Topics (IMSIS 2023), Montreal, Canada, 23-25 Oct, 2023. **Awarded best poster abstract.**

Vanhemel, T., Nijs, M., Cuypers, E., De Moor, B. (2024). *Predicting REIMS from MALDI for glioblastoma sections using deep learning.* Presented at the VIB conference: Applied Bioinformatics in Life Sciences (5th edition), Leuven, Belgium, 7-8 March, 2024.

Nijs, M., Waelkens, E., De Moor, B. (2024). *Analysis of MSI data using rr2-NMF reveals detailed spectral and spatial visualization of co-localized molecules, both highly and lowly abundant.* Accepted for presentation at the 72nd ASMS Conference on Mass Spectrometry and Allied Topics, Anaheim, CA, USA, 2-6 June, 2024.

Vanhemel, T., Nijs, M., Birmpili, A., Hendriks, T., Cuypers, E., De Moor, B. *Predicting REIMS from MALDI: to prevent destruction of glioblastoma sections used for model improvement.* Presented at the 72nd ASMS Conference on Mass Spectrometry and Allied Topics, Anaheim, CA, USA, 2-6 June, 2024.



FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
STADIUS
Kasteelpark Arenberg 10 box 2446
B-3001 Leuven
melanienijs@gmail.com
https://www.esat.kuleuven.be/stadius/