KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Manifold learning for visualization, prioritization, and data fusion of Mass Spectrometry Imaging data

Tina Smets

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

October 2021

Manifold learning for visualization, prioritization, and data fusion of Mass Spectrometry Imaging data

Tina SMETS

Examination committee:
Prof. dr. ir. Patrick Wollants, chair
Prof. dr. ir. Bart De Moor, supervisor
Prof. dr . Etienne Waelkens, supervisor
Prof. dr. ir. Johan Suykens
Prof. dr. ir. Frederik De Smet
dr. ir. Anneleen Daemen
(ORIC Pharmaceuticals, U.S.A.)
Prof. dr. ir. Olivier Gevaert
(Stanford University, U.S.A.)

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

October 2021

© 2021 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Tina Smets, 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.

Preface

> There are no facts, only interpretations. All things are subject to interpretation, whichever interpretation prevails at a given time is a function of power and not truth.
>
> *Friedrich Nietzsche*

Al tijdens mijn opleiding in de biomedische wetenschappen werd ik gebeten door de bio-informatica. Gigantische hoeveelheden data en dito potentieel.
De prominente rol die ESAT in dit verhaal speelde werd mij echter al snel duidelijk want Stein Aerts, mijn professor in de bio-informatica, bleek een voormalig doctoraatsstudent van Bart De Moor. Tijdens mijn doortocht bij Cartagenia viel mij vervolgens de eer te beurt een ander significant deel van zijn wetenschappelijk nageslacht te leren kennen. Nu, ettelijke jaren later ben ik zelf klaar om mijn doctoraat onder de vleugels van Bart af te leggen.

Om te beginnen wil ik daarom mijn promotor, professor Bart De Moor, bedanken voor de opportuniteit om mij vier jaar lang in de wondere wereld van de bio-informatica en meer specifiek de spatial omics te verdiepen.
Bart, bedankt voor het vertrouwen en de vrijheid om mijn onderzoek zelf richting te geven en samenwerkingen uit te bouwen, jouw sterke ondernemingszin is bewonderenswaardig!

Ook zou ik graag professor Etienne Waelkens bedanken om ondanks zijn welverdiende pensioen, toch nog steeds met een niet aflatend enthousiasme paraat te staan. Jouw positieve ingesteldheid en passie voor het onderzoek hebben een enorme motivatie gevormd gedurende mijn doctoraatstraject.
Bedankt om altijd je fantastische zelve te zijn, Etienne!

Daarnaast wens ik de verschillende leden van de jury te bedanken voor hun

waardevolle bijdragen aan deze thesis. Professor Johan Suykens wil ik graag bedanken voor zijn feedback omtrent de methodologie en zijn sterke interesse in manifold learning. Professor Frederik De Smet wil ik graag bedanken voor de immer waardevolle praktische en klinische blik op mijn werk en de vele interessante samenwerkingen waarin hij betrokken is. Ik wil zeker ook mijn externe juryleden professor Olivier Gevaert en Anneleen Daemen bedanken. Olivier, heel erg bedankt om mij te hosten in Stanford en voor de vele interessante en waardevolle discussies. Anneleen, hartelijk dank om van buiten de universitaire muren een kritische blik op dit werk te werpen, daar hecht ik veel belang aan. Bedankt om hiervoor tijd vrij te maken! Dank ook aan professor Patrick Wollants om als voorzitter van deze doctoraatsjury op te willen treden.

Uiteraard zou ik ook mijn vroegere en huidige collega's op het vierde verdiep willen bedanken om de voorbije vier jaar op te vrolijken.
Gorana, thank you for all the nice moments! I am really glad to have met you and to see our friendship grow. No words can describe what you have to go through at this moment and I hope with all my heart that better news is just waiting around the corner and all of this is just a bad dream.

Marc en Nico, bedankt voor de introductie op ESAT en de fijne tijden zowel op den bureau als op congres! Jullie hebben mij meermaals de slappe lach bezorgd en Nico zijn onovertreffelijke nepflauwval-acts zal ik nooit vergeten!
Christof, Bram, Oliver bedankt voor de vele inventieve vieruurtjes: choco-crèmebal kent geen geheimen meer voor mij of toch...;-) Giulia door corona hebben we helaas minder tijd op kantoor kunnen doorbrengen, maar de korte tijd samen was wel heel aangenaam. Bedankt daarvoor! Melanie, bedankt om zo een leuk bureaumaatje te zijn met een grote passie voor MSI! Hopelijk kunnen we de leuke afspraken na dit verhaal nog blijven verder zetten.
Mauricio, thanks for the numerous cheerful conversations and for spoiling us with delicious Columbian treats! Willem, bedankt voor de immer goede samenwerking op de verschillende projecten, jij bent een echte teamplayer en fantastische collega! Hannes, Naomi, Katrien, Lola, Wanqiu, Thibaut en Xi bedankt voor de fijne spontane gesprekken op de gang.

Ik wil zeker ook Ida, John, Wim, Elsy en Aldona bedanken voor hun geweldige ondersteuning en fijne babbels.

Ik wil hierbij ook mijn familie en schoonfamilie bedanken voor hun steun en interesse, hopelijk heeft mijn onderzoek na vandaag geen geheimen meer voor jullie!

In het bijzonder wil ik Tom bedanken voor de onvoorwaardelijke steun ondanks de recente moeilijke maanden. Bedankt om altijd mijn grootste fan te zijn en

de kleinste momenten van elke dag waardevol te maken.

Tina Smets
October 2021

Abstract

Mass Spectrometry Imaging (MSI) is a powerful molecular imaging technology that can detect the spatial distribution of molecules in a tissue section. Because MSI does not require any a priori labeling, the technique has become very popular for the explorative comparison of metabolites, lipids, peptides and proteins between various tissue regions. Since it has been shown that tumor heterogeneity plays an important role in tumor biology, it has become clear that we need to elucidate the spatial distribution of molecules. MSI can therefore be of significant relevance in predicting cancer progression and treatment response, which often remains a challenge in today's clinical practice.

A single measurement can however lead to complex and high dimensional data with file sizes in the gigabyte and even the terabyte range. As such manually exploring the data is becoming infeasible and support from computational methods is required. The focus of this work is therefore the development and application of computational methods to MSI data.
Specifically we have concentrated on the topics related to non-linear dimensionality reduction, the prioritization of molecules measured per tissue region, and data fusion of MSI with the corresponding histology or microscopy image.

For the non-linear dimensionality reduction, we make use of Uniform Manifold Approximation and Projection (UMAP) through which we achieve excellent visualizations of MSI data as demonstrated by the corresponding histology or microscopy images. We have conducted an extensive evaluation regarding the performance and results in comparison to other dimensionality reduction methods. To this end, we have used spatial autocorrelation and spectral similarity as a benchmark. In addition, we have empirically evaluated a number of different distance metrics, where we show that the choice of a particular distance metric might impact the visualization outcome.

Building further on the obtained visualizations using UMAP, we proposed a bi-directional dimensionality reduction approach to prioritize the molecules

driving these observations. This approach enables the prioritization of m/z-values in individual tissue samples but also across different tissue samples through the incorporation of both spatial and spectral information.
The approach was demonstrated for tissue samples obtained from healthy mouse pancreas tissue.

Finally, we introduce the correspondence-aware manifold learning paradigm for data fusion of molecular imaging data with the corresponding microscopy images. This enables us to bring the molecular information to a higher spatial resolution. As such these visualizations can play an important role in the digital pathology field for the quick assessment of a complete MSI dataset from a pathologist's perspective. We have shown that using this approach it becomes possible to identify a single infiltrating plasma cell amongst a group of phenotypically different (epithelial) cells. The identification of single aberrant cells is of crucial importance to evaluate a wide range of pathologies, in particular in cancer.

Beknopte samenvatting

Mass Spectrometry Imaging (MSI) is een krachtige moleculaire beeldvormingstechnologie die de spatiale verdeling van moleculen in een weefselcoupe kan detecteren. Omdat MSI geen a priori labeling vereist, is de techniek zeer populair geworden voor de exploratieve vergelijking van metabolieten, lipiden, peptiden en proteïnen tussen de verschillende weefselregio's. Vermits het aangetoond is dat tumorale heterogeniteit een belangrijke rol speelt in tumorbiologie, is het duidelijk geworden dat we de spatiale distributie van moleculen dienen op te helderen. MSI kan daarom van significant belang zijn in het voorspellen van kanker progressie en behandelingsrespons, hetgeen vaak een uitdaging blijft in de huidige klinische praktijk.

Een enkele meting kan echter tot complexe en hoog-dimensionale data leiden in de gigabyte en zelfs tot in de terabyte grootteorde. Het manueel exploreren van de data is daarom onmogelijk geworden en ondersteuning door computationele methoden is vereist. De focus van dit werk ligt daarom op de ontwikkeling en toepassing van computationele methodes op MSI data. Specifiek hebben we ons gefocust op de onderwerpen gerelateerd aan niet-lineaire dimensionaliteitsreductie, de ranking van moleculen gemeten per weefselregio en een data fusion aanpak van MSI met de overeenkomstige histologie of microscopie afbeelding.

Voor de niet-lineaire dimensionaliteitsreductie, maken we gebruik van Uniform Manifold Approximation and Projection (UMAP) waardoor we excellente visualisaties bekomen van MSI data, hetgeen we kunnen vergelijken met de corresponderende histologische secties. We hebben bovendien een uitgebreide evaluatie uitgevoerd aangaande de performantie en resultaten in vergelijking met andere dimensionaliteitsreductie aanpakken. Daartoe hebben we gebruik gemaakt van spatiale autocorrelatie met de spectrale similariteit als benchmark. Daarenboven hebben we een empirische evaluatie uitgevoerd met betrekking tot het gebruik van verschillende afstandsmaten waarbij we tonen dat de keuze van

een bepaalde afstandsmaat de uitkomst van visualisatie kan beïnvloeden.

Verder bouwend op de verkregen visualisaties door middel van UMAP stellen we vervolgens een bi-directionele dimensionaliteitsreductie voor om de moleculen te prioriteren die deze observaties drijven. Deze aanpak laat ons toe om de m/z - waarden te rangschikken in individuele stalen maar ook over de verschillende weefsel stalen heen door het in rekening brengen van zowel spatiale als spectrale informatie. Deze aanpak werd aangetoond voor een reeks stalen van wildtype muizen.

Finaal introduceren we het correspondence-aware manifold learning paradigma voor data fusie van moleculaire beeldvorming data met de corresponderende microscopie beelden. Dit laat ons toe om moleculaire informatie naar een hogere spatiale resolutie te brengen. Zo kunnen deze visualisaties namelijk bijdragen aan het digitale pathologie domein voor de snelle interpretatie van een volledige MSI dataset vanuit het perspectief van de patholoog. We hebben aangetoond dat door middel van deze aanpak het mogelijk wordt om een enkele infiltrerende plasma cel te onderscheiden van een groep fenotypisch verschillende (epitheliale) cellen. Het onderscheiden van aparte, afwijkende cellen is echter van cruciaal belang voor de evaluatie van een brede waaier aan pathologieën, met name voor kanker.

List of Abbreviations

CCS collisional cross section. 77

DESI Desorption Electrospray Ionization. 5

DHB 2,5-dihydroxybenzoic acid. 8

FDR False Discovery Rate. 12

FT Fourier Transformation. 11

FTICR Fourier transform ion cyclotron resonance. 7, 8

FTIR Fourier-transform infrared. 3

IMS Ion Mobility Separation. 77

MALDI Matrix Assisted Laser Desorption/Ionization. 5–7

MS Mass Spectrometry. 2, 6, 78

MSI mass spectrometry imaging. 3, 9, 12, 19, 22, 32, 37, 39, 54–56, 58, 64, 68, 73, 74, 77, 78, 80

NIRS Near-infrared spectroscopy. 3

NMF non-negative matrix factorization. 13

PCA principal component analysis. 13, 19

QMS Quadrupole Mass Spectrometer. 7

SIMS Secondary Ion Mass Spectrometry. 5, 7

TOF Time-of-Flight. 7, 8

Contents

Abstract v

Beknopte samenvatting vii

List of Abbreviations ix

List of Symbols xi

Contents xi

List of Figures xv

List of Tables xxi

1 Introduction 1

1.1 Spatial Metabolomics 2

1.2 Spatial Metabolomics: Mass Spectrometry Imaging (MSI) . . . 3

1.2.1 Flavors of MSI 4

1.2.2 Advantages and applications of MSI 9

1.2.3 Computational analysis of MSI data 10

2 Visualization of Mass Spectrometry Imaging data 17

2.1 Evaluation of distance metrics and spatial autocorrelation in Uniform Manifold Approximation and Projection applied to Mass Spectrometry Imaging data 19

2.1.1 Introduction 19

2.1.2 Experimental Section 22

2.1.3 Results and Discussion 27

2.1.4 Towards a relative comparison of the embeddings obtained by UMAP and t-SNE via spatial autocorrelation 29

2.1.5 The underestimated importance of the distance metric . 31

2.2 Conclusion 33

3 Prioritization of m/z-values in mass spectrometry imaging data 35

3.1 Prioritization of m/z-values in mass spectrometry imaging profiles obtained using Uniform Manifold Approximation and Projection for dimensionality reduction 38

3.1.1 Introduction 38

3.1.2 Experimental Section 41

3.1.3 Mass Spectrometry Imaging (MSI) data acquisition and processing 41

3.1.4 Data processing of Mouse Pancreas datasets 41

3.1.5 Results and Discussion 43

3.1.6 Spatial and spectral prioritization based on the hyperspectral visualization obtained using UMAP enables quality control and molecular insights 43

3.1.7 Conclusion 52

4 Data fusion of Mass Spectrometry Imaging and histology data 53

4.1 Correspondence-aware manifold learning for microscopic and spatial omics imaging: a novel data fusion method bringing Mass Spectrometry Imaging to a cellular resolution 55

4.1.1 Introduction 55

4.2 Methods 57
4.2.1 UMAP outline 57
4.2.2 Capturing correspondence 58
4.2.3 Correspondence-aware manifold learning (CAML) . . . 58
4.2.4 Data measurements 61
4.3 Results 61
4.3.1 Correspondence-aware manifold learning for data fusion of molecular and microscopy images 62
4.3.2 Out-of-sample prediction 64
4.3.3 Data fusion for spatial transcriptomics data 64
4.4 Discussion 64

5 Conclusion and future perspectives 72

5.1 Conclusion 73
5.2 Future perspectives 77
5.2.1 Technical improvements leading to novel, more complex or larger datasets 77
5.2.2 The multi-omics and single cell era: integration with other data modalities 78
5.2.3 Molecular identification 80

A Supporting Information for Evaluation of distance metrics and spatial autocorrelation in Uniform Manifold Approximation and Projection applied to Mass Spectrometry Imaging data 83

A.1 Overview of distance metrics 84
A.2 Overview of runtimes associated with t-SNE, Barnes-Hut t-SNE and UMAP 84
A.3 Pancreas M2: relative comparison of obtained embeddings through spatial autocorrelation 84
A.4 Pancreas M2: UMAP and Barnes-Hut t-SNE applied with different distance metrics 84

A.5 Distance metric evaluation on MSI data without TIC normalisation 84

A.6 Custom distance metric - histomatch 84

B Prioritization of m/z values in mass spectrometry imaging profiles obtained using UMAP for dimensionality reduction 104

C Supporting Information for Correspondence-aware manifold learning for microscopic and spatial omics imaging: a novel data fusion method bringing Mass Spectrometry Imaging to a cellular resolution 118

Bibliography 127

List of publications 141

C.1 International journal publications 141

C.2 Posters and presentations at International Conferences 142

List of Figures

1.1 Overview of MSI workflow. 4

1.2 Overview of data processing in a typical MSI workflow. 12

2.1 Conceptual overview 23

2.2 General outline of the UMAP algorithm. 25

2.3 Comparison of t-SNE and UMAP for pancreas samples 28

2.4 Comparison of dimensionality reduction approaches. 28

2.5 Spatial autocorrelation for embeddings obtained using the Euclidean distance metric. 30

2.6 Comparison of the Chebyshev, correlation and cosine distance metrics using BH t-SNE and UMAP for the pancreas samples. . 31

2.7 Comparison of the Chebyshev, correlation and cosine distance metrics using UMAP for the lymphoma samples. 32

2.8 Hyperspectral visualisation of the lymphoma P1 sample with corresponding scatterplot. 33

2.9 Hyperspectral visualisation of lymphoma UMAP embedding using a custom distance metric called 'histomatch' with corresponding scatterplot compared with the results obtained using the cosine distance metric. 34

3.1 Method overview - prioritization of ion images. 40

3.2 Overview of the hyperspectral visualizations for the three healthy pancreas tissue samples. 44

3.3 Overview of endocrine region ranking across different pancreas tissues. 46

3.4 Detail of prioritized endocrine ion images and histological staining. 47

3.5 Overview of exocrine region ranking across different pancreas tissues. 48

3.6 Illustration of ranking and biological association with exocrine pancreas tissue. 49

3.7 Hyperspectral visualization for distance metric comparison. . . 50

3.8 Ranking of ion images for distance metric comparison. 51

4.1 Conceptual overview. Correspondence-aware manifold learning is able to fuse the complete molecular feature space of high-dimensional omics data with their corresponding microscopy image. 56

4.2 Method overview - correspondence-aware manifold learning. . . 67

4.3 Low-dimensional representation of lymphoma MSI dataset and corresponding H&E image. 68

4.4 Fusion of mucosa-associated lymphoid tissue of the tonsil MSI dataset and corresponding H&E image. 69

4.5 A detail of the fused data. Panel A shows the extracellular matrix and stroma comprising glycoproteins and proteoglycans, as well as plasma cells on top and the multi-layered squamous epithelium below. Panel B shows a detailed view of a cluster of plasma cells. 70

4.6 Example of out-of-sample prediction using correspondence-aware manifold learning. 71

5.1 Overview of average dimensionality associated with the demonstrated MSI datasets and associated runtimes. 74

5.2 Experimental comparison of runtimes associated with a number of dimensionality reduction methods in function of increasing dataset sizes. The graphs illustrate the superiority of UMAP in particular for large datasets. 75

5.3 On the left an annotated H&E staining is shown, with in blue morphologically similar glands. On the right, spatial lipidomics uncovers distinct molecular profiles for the same tumor specimen across morphologically comparable glands in prostate cancer tumor tissue but requires stronger discriminatory power to unravel the underlying mechanisms. 79

5.4 Strengths and weaknesses associated with each data modality. Using data fusion we can for example increase the spatial resolution of the molecular imaging methods to the resolution of microscopic H&E images. 80

A.1 Overview of different distance metrics used. 85

A.2 Overview of rumtimes for different dimensionality reduction methods used. 86

A.3 Spatial autocorrelation and spectral similarity benchmark for the pancreas tissue embeddings obtained using the cosine distance metric. 87

A.4 Spatial autocorrelation and spectral similarity benchmark for the pancreas tissue embeddings obtained using the correlation distance metric. 88

A.5 Spatial autocorrelation and spectral similarity benchmark for the pancreas tissue embeddings obtained using the Chebyshev distance metric. 89

A.6 Evaluation of robustness of t-SNE and UMAP embeddings obtained using Euclidean distance metric with spatial autocorrelation and spectral similarity used as a benchmark. 90

A.7 Hyperspectral visualization of pancreas tissue embeddings obtained with different distance metrics - part 1. 91

A.8 Hyperspectral visualization of pancreas tissue embeddings obtained with different distance metrics - part 2. 91

A.9 Impact of total ion current (TIC) normalization on performance of non-linear dimensionality reduction methods - M1 sample, Euclidean and correlation distance metric. 92

A.10 Impact of total ion current (TIC) normalization on performance of non-linear dimensionality reduction methods - M1 sample, cosine and Chebyshev distance metric. 93

A.11 Cumulative variance explained by number of principal components - pancreas data (1) 94

A.12 Cumulative variance explained by number of principal components - pancreas data (2) 95

A.13 Cumulative variance explained by number of principal components - lymphoma data (1) 96

A.14 Cumulative variance explained by number of principal components - lymphoma data (2) 97

A.15 Impact of total ion current (TIC) normalization on performance of non-linear dimensionality reduction methods -M2 sample, Euclidean and correlation distance metric. 98

A.16 Impact of total ion current (TIC) normalization on performance of non-linear dimensionality reduction methods - M2 sample, cosine and Chebyshev distance metric. 99

A.17 Impact of total ion current (TIC) normalization on performance of non-linear dimensionality reduction methods - P1 lymphoma sample, Euclidean, correlation, cosine distance metric. 100

A.18 Impact of total ion current (TIC) normalization on performance of non-linear dimensionality reduction methods - P1 lymphoma sample, Euclidean distance metric 101

A.19 Spatial autocorrelation and spectral similarity benchmark for the pancreas tissue embeddings obtained using the cosine and histomatch distance metric. 102

A.20 Spatial autocorrelation and spectral similarity benchmark for the P2 lymphoma tissue embedding obtained using the histomatch and cosine distance metric. 103

B.1 Hyperspectral visualization for sample 1 obtained from a healthy mouse pancreas with extracted profile corresponding to the light green region. 105

B.2 Spectral and spatial results ranked vertically from top to bottom corresponding to light green region in M1 pancreas sample. . . 106

B.3 Hyperspectral visualization for sample 1 obtained from a healthy mouse pancreas with extracted profile corresponding to the red region. 107

B.4 Spectral and spatial results ranked vertically from top to bottom corresponding to the red region in M1 pancreas sample 108

B.5 Hyperspectral visualization for sample 2 obtained from a healthy mouse pancreas with extracted profile corresponding to the dark blue region. 109

B.6 Spectral and spatial results ranked vertically from top to bottom corresponding to the dark blue region in S2 pancreas sample . . 110

B.7 Hyperspectral visualization for sample 2 obtained from a healthy mouse with the extracted profile corresponding to the light green region. 111

B.8 Spectral and spatial results ranked vertically from top to bottom corresponding to the light green region in S2 pancreas sample . 112

B.9 Hyperspectral visualization for sample 3 obtained from a healthy mouse pancreas with extracted profile corresponding to the yellow regio 113

B.10 Spectral and spatial results ranked vertically from top to bottom corresponding to the yellow region in S3 pancreas sample . . . 114

B.11 Hyperspectral visualization for sample 3 obtained from a healthy mouse pancreas with the extracted profile corresponding to the light green region. 115

B.12 Spectral and spatial results ranked vertically from top to bottom corresponding to the light green region in S3 pancreas sample . 116

B.13 Ion images for m/z value around 3484 Da which could putatively be associated to glucagon 117

C.1 Overview of the behavior of an increasing penalty parameter. . 119

C.2 Correspondence-aware manifold learning data fusion result (left) and histological detail of Germinal Center and Lymphocyte Corona. 120

C.3 Detail of correspondence-aware manifold learning data fusion result, highlighting an arteriole and venule, as confirmed by the corresponding histology staining. 121

C.4 Detail of artefact in lymphoma tissue sample, emphasized by correspondence-aware manifold learning data fusion. 121

C.5 Demonstration of correspondence-ware manifold learning to a public spatial transcriptomics dataset. 122

C.6 Illustration of consistent mapping of cell nuclei by green color upon data fusion of spatial transcriptomics mouse brain data with corresponding H&E staining. 123

C.7 CAML validation and general applicability. We built a multi-modal dataset based on MNIST, the handwritten digits dataset. 124

C.8 Validation of CAML out-of-sample prediction.. 125

C.9 Evaluation of the impact on final result when changing the order of the RGB axes. 126

List of Tables

2.1 Overview of dimensionality associated with the used datasets . 23

Chapter 1

Introduction

In the first chapter, we start by situating the biological and practical context of this work. Rapid technological advances in the medical sciences and artificial intelligence are steering the healthcare field towards precision medicine. The term **precision medicine** can be specified as 'The tailoring of medical treatment to the individual characteristics of each patient to classify individuals into subpopulations that differ in their susceptibility to a particular disease or their response to a specific treatment' [37]. To enable precision medicine a number of individual variabilities need to be taken into account such as environmental and lifestyle parameters, genetics, and the molecular phenotype of an individual.

Progress in the field of human genetics has allowed the genome-wide interrogation of individual and population-wide sequence variation, enabling the identification of new variants for Mendelian disorders with important contributions to clinical care. However, still a lot of work has to be done to support the precision medicine use case of more complex diseases, in which multiple genetic and environmental effects come together with an increased disease risk [137].

The application of clinical genomics in cancer for instance constitutes a powerful tool to assess individual variability, but it does not provide information regarding other biomolecules, such as metabolites, which also play a crucial role in **defining molecular phenotypes**. Metabolomics research is therefore directed at the detection and interpretation of metabolites, lipids, drugs and other small molecules in the context of cells, tissues, organs and organisms [35, 117]. Typically these measurements are conducted by **extracting metabolites** in bulk from cells and tissues or bodily fluids for further detection and quantification using **Mass Spectrometry** (MS) [64, 95, 74] which provides a good molecular coverage, sensitivity and specificity.

A downside of this approach is the infeasability of localizing the metabolites in specific cells or anatomically defined parts of organs or tissues. **The ability to detect metabolites in situ is of particular importance for studying the tumor microenvironment** and immunotherapy but also for example to dissect the interaction between cells [90], the functioning of immune cells [24], and their contribution to inflammation [124, 47, 42] and the metabolic regulation of epigenetics [73, 5].

1.1 Spatial Metabolomics

Given the importance of spatially correlating metabolites to their anatomical structures, many technological efforts have been devoted to the **development**

of spatial metabolomics approaches such as Raman imaging spectroscopy, Fourier-transform infrared (FTIR) spectroscopy, and Mass Spectrometry Imaging.

Raman imaging spectroscopy combines a microscope with a Raman spectrometer and relies on the inelastic scattering of light, the so-called Raman scattering, to determine vibrational modes of molecules [119, 33].

Fourier-transform infrared spectroscopy quantifies the absorption of infrared light by molecules such as lipids, lipopolysaccharides, nucleic acids, carbohydrates, and proteins, resulting in a specific FTIR spectrum that reflects the overall composition of the sample [20, 132].

Near-infrared spectroscopy NIRS measures the absorption of electromagnetic radiation including wavelengths from 750 to 2500 nm. Through the electromagnetic radiation absorbed from molecular bonds in these NIR wavelengths produces spectra which are unique to a sample acting as a 'fingerprint' [100, 104].

In this work we will focus on MSI data which can be considered as one of the most popular and a key enabling technology for spatial metabolomics.

1.2 Spatial Metabolomics: Mass Spectrometry Imaging (MSI)

Mass Spectrometry Imaging (MSI). MSI makes it possible to detect metabolites, lipids, drugs and other small molecules in their spatial context [96]. A typical workflow of a MSI experiment is shown in Figure 1.1. After sample preparation, an (x, y) grid is placed over the surface of the sample such that the molecules on the sample surface are subjected to ionisation by the mass spectrometer through the application of a laser or another ablation method. A mass spectrum is collected at each pixel location where the spatial resolution will depend on the chosen pixel size.

The output of an MSI experiment is a mass spectrum corresponding to each pixel location in the tissue sample which can be visualized as the intensity per mass-to-charge (m/z) value measured. These intensities can then be combined into a heat map showing the relative distribution of a particular m/z-value throughout a sample. To reveal the identity of a measured m/z-value tandem MS (MS/MS) fragmentation can be performed such that the fragments can be used to puzzle together the structure of the unknown molecule. Alternatively, identification of molecules can be conducted based on its intact

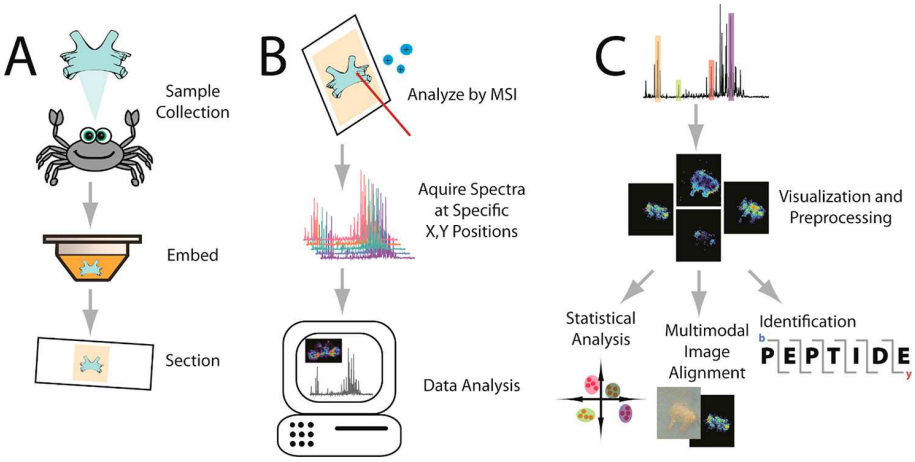


Figure 1.1: **Overview of MSI workflow.**
In this example, a sample is obtained from a crustacean's brain.
(A) Sample preparation. The sample is embedded in a supporting medium for sectioning onto slides. It should be noted that cryosectioning is also a common sample preparation step. Depending on the molecular species of interest or the instrument used, alternative sample processing may be performed (e.g. applications of enzymes, matrix, or derivatization agents.
(B) Sample analysis. A spectrum is acquired at each (x,y) grid point on the sample, computational methods are used to process and visualize the data. A laser or antoher ablation method is used to ionize molecules.
(C) Data processing. After preprocessing (e.g., baseline correction, normalization,..), heatmaps of selected molecules can be visualized. From there, downstream analysis such as identification of the m/z-values and statistical analysis between different images or image coregistration with other image modalities can take place. Figure reprinted with permission from [22]. Copyright (2018) American Chemical Society.

mass by accurate mass matching to databases of known molecules within a certain mass error range [22, 19, 138].

1.2.1 Flavors of MSI

Different acquisition approaches can be used to obtain MSI data. The different flavors of MSI depend on the properties of the mass spectrometer and the most important elements having an impact on the data are the ion source, the mass analyzer type and sample preparation.

These parameters have an effect on the molecular content of the data since they determine which molecular classes will be desorbed and ionized [5, 22].

The analyzer type however will also affect the spectral properties of the data and this has some consequences for the computational analysis because it can affect the size of the data and thus signal processing as well [5].

Ion Source

The most used ionization methods are **DESI**, **SIMS**, and **MALDI**. The ion source is essential to ionize the molecules present in a tissue sample, after which the ions are guided towards the mass analyzer. The choice of the ionization technique contributes to the desorption and ionization efficiency and therefore also has an influence on which molecular species can be detected [106].

- **Desorption Electrospray Ionization (DESI)**

  DESI was introduced in 2004 by Takats et al.[120] and emerged as a powerful ambient ionization surface analysis technique capable of detecting a wide range of molecules including polymers, pharmaceutical compounds, lipids, glycans, peptides, and proteins [45, 31, 79, 18, 122, 121, 34].

  The method consists of spraying the sample with a nebulizing gas carrying charged solvent droplets. The impact of these charged droplets on the surface causes desorption and ionization of molecules present on the surface that are then transported toward the atmospheric pressure inlet of the mass spectrometer. DESI has the advantage that it can operate at ambient pressures and does not need any sample preparation which differentiates DESI from other mainstream MSI modalities.

  The sprayer is however a critical factor in the DESI imaging system shown to cause poor repeatability and reproducibilty in several studies for imaging of tissue sections [65, 53, 34, 106].

- **Secondary ion mass spectrometry imaging (SIMS)**

  The idea of employing mass spectrometry for imaging purposes was originally introduced in the early sixties by Castaing and Slodzian [71] based on the mass spectrometry of secondary ions. In SIMS, ions are generated by bombarding a sample with high-energy 'primary ions' causing an ejection of 'secondary ions' from the surface. These secondary ions are collected and separated according to their mass ratios. SIMS offers the highest lateral resolution of the MSI methods available, going from the micrometer to nanometer scale spatial resolution [10, 52, 106].

  The application of SIMS to characterize biologically related samples has increased in recent years as new instruments are being developed that

combine capabilities to balance strengths and weaknesses. For example, coupling TOF-SIMS with tandem MS capabilities would allow for rapid imaging of a sample followed by verification of the structural confirmation of ions of interest. Moreover, the introduction of cluster ion sources over the past decade enable a softer ionization and has therefore improved the mass analysis range and sensitivity of SIMS considerably. An advantage of SIMS is that it requires little to no sample preparation [10, 12, 57, 106].

- **Matrix Assisted Laser Desorption/Ionization (MALDI)**
  MALDI as the name suggests involves the deposition of a chemical matrix onto the tissue sample to extract molecules from the complex sample. The formation of analyte-doped matrix crystals segregates the analytes from detrimental co-factors such as salts, and disentangles and isolates the large molecules in the matrix. Ablation of these crystals with UV laser pulses causes ionization and desorption. As such it becomes possible to analyze variety of biomolecules ranging from metabolites and lipids to peptides and proteins with a mass range going over 100 kDa. While MALDI is one of the most popular configurations in MSI and the preferred method to study intact proteins, this approach is is also associated with some disadvantages [82].
  Firstly, the matrix deposition is of great importance for the ionization efficiency of the analyte molecules which means that an uneven matrix deposition may result in misleading measurements ie. differing ion intensities for tissue that actually holds the same biomolecular content. Secondly, matrix deposition may result in some spatial migration of the molecular analytes while the matrix is still liquid. An impeccable sample preparation is therefore indispensable for robust MALDI MSI experiments. Innovative sublimation techniques can support researchers in applying an even matrix deposition and reducing the risk of analyte migration because it causes the matrix to crystallize immediately [136, 82].

Mass analyzer

The mass resolution, accuracy and sensitivity all depend on the choice of mass analyzer. It is the mass analyzer that separates the ions according to their mass-to-charge ratio (m/z). Because packets of ions of the same mass and charge are separated in space/time from ions with a different mass and charge, ions can be recognized individually. This also means that two ions are indistinguishable when they have the same mass-to-charge ratio (eg. stereoisomers). To distinguish these molecules, the ions need to be broken apart and the fragmentation investigated by an additional mass analyzer. This multiplex approach is called tandem mass spectrometry or MS/MS. When a

variety of isobaric molecules are present, having knowledge of the exact mass is not sufficient for identification of the analytes.

Tandem MS plays an important role in this regard, because via fragmentation the tandem mass spectrum may provide a unique profile of the product ions which represent a molecular fingerprint. Although further fragmentation steps might be necessary for example in case of some structurally similar lipid species [106, 60]

We can categorize mass analyzers in **TOF**, **quadrupole**, **ion cyclotron resonance**, **sector**, **ion trap**, and **orbitrap** instruments.

- **Time-of-flight (TOF)**
  The TOF procedure relies on the determination of an ion's mass-to-charge ratio via a time of flight measurement. Ions are accelerated by an electric field of known strength, resulting in an ion having the same kinetic energy as any other ion that has the same charge [36]. The velocity of the ion will be correlated to the mass-to-charge ratio ie. heavier ions of the same charge reach lower speeds, although ions with higher charge will also increase in velocity. The time that it takes for the ion to reach a detector is measured, which will depend on the velocity of the ion, and therefore is a measure of its mass-to-charge ratio that can be used to identify the ion [36].
  The majority of SIMS and MALDI imaging measurements make use of TOF analyzers for their strong detention efficiency and parallel detection of ions resulting in a high sensitivity. Moreover, TOF analyzers are known for their operational speed and broad m/z range, enabling the analysis of large singly charged molecular ions produced by MALDI [106, 57].

- **Quadrupole**
  Quadrupole mass analyzers are also known as transmission quadrupole mass spectrometer (QMS). As the name suggests the system consists of four parallel metal rods to which an oscillating electric field is applied separating the ions based on the stability of their trajectories. Ions with unstable trajectories will collide with the rods enabling the operator to scan for a range of m/z-values by continuously varying the applied voltage [59].

- **Ion cyclotron resonance**
  This approach involves the application of a strong magnetic field in combination with a weak electric field to the ions. A cyclotron is used for accelerating the ions to measure the masses of an ionized analyte. A prominent example is Fourier transform ion cyclotron resonance (FTICR) [101].

FTICR is based on the circular (cyclotron) movement of charged particles in a strong magnetic field. The cyclotron frequency depends on the mass-to-charge ratio of the ions. The fourier transform is used to convert the periodic movement of ion packets to a frequency spectrum, subsequently converted to a mass spectrum after calibration.
FTICR instruments have a superior mass resolution making it possible to discern ions with closely located mass-to-charge ratios, e.g., isotopic peaks of high-mass ions. Furthermore combining FTICR with fragmentation methods such as electron capture dissociation or infrared laser activation, enables the fragmentation of very large molecules such as for example whole proteins [56].

- **Sector**
  Sector mass analyzers refer to a class of mass spectrometers using a static electric (E) or magnetic (B) sector or some combination of the two (separately in space) as a mass analyzer. Most modern sector instruments are double-focusing instruments meaning they focus the ion beams both in direction and velocity [86, 26].

- **Ion trap**
  Ion trapping refers to the combination of electric or magnetic fields to capture charged ions. In the mass spectrometry field a prominent example is the Orbitrap, which was introduced to the field in 1999.
  The orbitrap borrows several elements from the other approaches such as the principle of image current detection from FTICR, the use of ion trapping in precisely defined electrode structures from the radiofrequency ion trap, pulsed injection and the use of electrostatic fields from the TOF analyzers. In combination, these features resulted in a powerful collection of analytical features reaching ultrahigh mass-resolutions [61, 141].

Sample preparation The sample preparation has a potential impact on the data quality through its effect on desorption, ionization and sensitivity. Part of the sample preparation is the choice of a particular solvent. In case of MALDI MSI, the matrix plays a crucial role because this will determine the detectable molecular classes eg. 2,5-dihydroxybenzoic acid DHB, is a matrix suitable for peptides and glycopeptides, Norharmane for the detecton of lipids etc.

Taking all information together the 'ideal' flavor of MSI depends on the specific aim of a MSI experiment in particular the target molecular class, but also the required spatial and mass resolution needs to be taken into account. Different approaches can be complementary and as such it can be advantageous to combine different flavors of MSI to maximize the molecular readout through so-called multi-modal studies [32, 94].

In this work we will focus on **MALDI-TOF MSI**.

1.2.2 Advantages and applications of MSI

The multiplex and label-free approach of MSI constitute a major advantage since it enables the investigation of proteomic, peptidomic and metabolomic content in one and the same measurement and study their distribution and the correlation thereof with tissue morphology.

MSI is therefore a suitable technique for experimental purposes because there is no need to identify and label any target molecules beforehand. In contrast to methods such as immunohistochemistry that enable the analysis of just a few labeled markers, limiting its use in this respect to a more confirmatory role. The spatio-molecular information collected can then be linked to biological pathways or functions with a wide variety of biological applications [4].

- **Spatio-molecular registration**

  MSI has often been used as a discovery tool for proteomic signatures and biomarkers by evaluating the spatial distribution of single proteins, protein isoforms, posttranslational modifications, cleavage, and degradation. However increasingly MSI is being used to analyze lipids in situ because of the important role played by lipids in cancer and cell metabolism in general. Besides being a major structural component of cell membranes, lipids such as phospholipid derivatives are important for proper signal transduction and are thus very relevant in a number of pathological problems. [4]

- **Clinico-pathological applications**

  With the ability to detect peptides, proteins and metabolites in situ MSI holds enormous potential to support a variety of medical applications. For example mass spectrometric signatures based on a combination of biomolecules can be used as a biomarker for overall survival or patient response to therapy. Such signatures have for instance been instrumental in predicting unfavorable overall survival after surgical resection of gastric cancer and the response to a neoadjuvant therapy with Paclitaxel in breast cancer [4, 11, 14].

  Moreover, MSI can support the differentiation among tumor subtypes, stages and degree of metastasis. Traditionally, clinical diagnosis in the pathology field is based on histological and immunohistochemical assessment of tumor tissue. Pathologists are highly specialized and trained to identify morphological aberrations in a cellular context but

it has been shown that cancer cell populations can be be histologically indistinguishable at the microscopical level, while having unique molecular phenotypes that drive tumor progression and patient outcomes [4, 135, 50].

A better understanding of these molecular differences will therefore be crucial to gain insight in cancer development and the impact of intratumor heterogeneity in the management of cancer patients[4, 126]. MSI can thus be used to offer support in the subclassification of tumor samples.

- **Pharmaceutical applications**

  Pharmaceutical studies focusing on drug discovery require absorption, distribution, metabolism, and excretion data to fully understand the efficacy, safety, and thus viability of a compound following administration. Gaining rapid knowledge about these elements is crucial in order to save time and costs in the drug discovery pipeline [99, 4].

  Drug distribution in tissue sections has typically been evaluated using autoradiography which means compounds of interest need a radioactive chemical label. As such only one compound could be assessed per experiment. MSI enables multiplex and label-free measurements but also holds a lot of potential to support pharmacokinetic studies directly in the tissue through quantitation.

  Moreover, using MSI it is for instance possible to study the distribution of a drug in relation to vascularization when coregistrating the MSI measurements to the immunohistochemical-stained section with eg. CD31. Combining the information obtained using MSI concerning the distribution of the drug and its metabolites with its underlying cell metabolic, proteomic, lipidomic signatures and related pathways will improve insights regarding mechanisms of action and will therefore be an important element for the pharmaceutical industry [4].

1.2.3 Computational analysis of MSI data

A single MSI experiment returns thousands of molecules at each pixel across a complete tissue sample simultaneously. The wide range of molecules that can be measured (proteins, metabolites and lipids) makes MSI an excellent method for explorative studies. Nevertheless using MSI leads to the generation of gigabytes and even terabytes of data depending on the acquisition method and associated parameters. MSI data can be described as a three dimensional data cube with two spatial dimensions (x and y coordinates) and one molecular dimension (mass-to-charge or m/z- values).

The number of pixels is typically situated between 10 000 and 100 000 pixels. With techological advancements of novel instruments (eg. laser focus stability) and increasing acquisition speed, dataset sizes are further increasing making the routine collection of experiments with more than one million pixels possible.

The size of the spectral dimension depends on the analyzer ranging from 10^3 for TOF analyzers, 10^4 for quadrupole TOF analyzers and over 10^4 for FT analyzers. Because the data collected using FT analyzers tends to become very large due to the high mass resolution of these instruments, centroiding is typically applied along the spectral dimension.

Centroiding involves the selection of one central m/z-value (ie. the 'centroid') for every spectral peak to reduce the dimension and size of multigigabyte data by an order of magnitude [5]. The mass resolving power predetermines how detailed the spectra are and whether two molecules with similar $m/z - values$ can be distinguished. The analyzer type therefore has a large impact on the spectral properties and size of the data thereby affecting not only data analysis in itself but also data management.

Mass accuracy on the other hand determines how the measured centroids of the spectral peaks for a molecule represent its theoretical m/z value with a large impact on data quality and overall interpretability in particular with regard to molecular ambiguity. The mass accuracy is greatly influenced by the calibration and stability of the mass spectrometer but also the thickness of the sample in case of a TOF approach. Other parameters such as sample preparation, source type and analyzer polarity do not directly affect the spectral properties of the data and hence have no downstream impact on the data analysis workflow in general, however they have an influence on which molecular classes will be desorbed and ionized and thus strongly affect the molecular content of the data [5, 69, 111].

Taking everything together, the experimental set-up has important consequences in terms of data size and properties but also in terms of molecular content being assessed. Moreover a number of parameters might impact data quality and might inject noise into the data. Retrieving the relevant biomolecular information from the millions of ion images generated to discover correlations between the measured molecules manually has become infeasible, making computational methods that can reduce the size of the data and prioritize the ion images present in the data indispensable. Figure 1.2 shows a comprehensible overview of data characteristics and the associated computational requirements.

- **Data management and signal processing** Data management and interoperability are becoming increasingly important due to growing

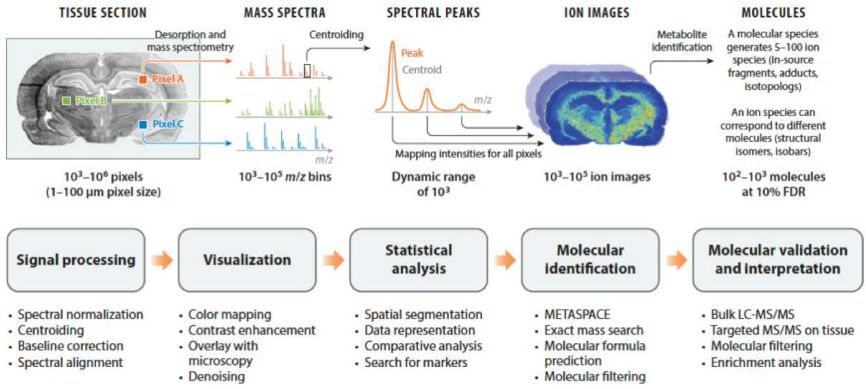


Figure 1.2: **Overview of data aspects and associated processing workflow**.
MSI data represents a data matrix with a collection of spectra obtained from a raster of pixels according to the surface of a tissue section. The number of pixels depends on the spatial resolution (typically ranging from 1 - 100 μm). The peaks present in the acquired mass spectra are often reduced to centroids, in particular when using FTICR and Orbitrap instruments with a high-resolving power analyzer. MSI data can then represent up to 10^3 molecules with a false discovery rate (FDR) of 10 %.
The image below shows the different data processing steps taken to process a MSI dataset. The focus of computational research has shifted towards downstream processing since low-level signal processing is increasingly taken care of by the vendor software.
Figure reprinted with permission from [5] Copyright (2020) Annual Reviews.

dataset sizes. The most important format used for storing and interchanging MSI data is currently the .imzML format. The format is not the most efficient in terms of data access hence with dataset sizes becoming larger, improvements are desirable at this level [109]. Alternatives based on the HDF5 format have been proposed which rely on HDF5's built-in gzip support and are said to have excellent storage efficiency as well as support for both parallel read and write operations [107].

Signal processing refers to elements of the data analysis pipeline such as baseline correction, spectral normalization and spectral centroiding.

While many approaches are available, some methods addressing spectral normalization were shown to also produce artifacts. No universal or standardized approach is available leaving the choice of applying normalization and which approach taken still open and the lack of ground truth data containing known distributions of biomolecules hampers the benchmarking of existing methods [22, 5].

- **Visualization of ion images**

  An important advantage of MSI lies in the availability of spatial information, good visualization methods can therefore provide an intuitive perspective on the detected metabolites, their localization and spatial correlation. In addition to visualizing the molecular information, an overlay with the microscopy image from the same or a consecutive tissue section can support in identifying metabolites with histopathologically defined regions of a particular type or phenotype.
  Overlays can also be generated with images obtained from the expression of a particular protein or transcript using eg. immunohistochemistry or immunofluorescence assays. The choice of a colormap has been an important topic in the MSI field due to the need for visually comparing a number of ion images. Rainbow and jet colormaps are not recommended and preference is given to perceptually uniform colormaps like viridis [133, 22, 92].

- **Data analysis**

  Visual examination of each ion image individually is becoming infeasible for the growing amount of ion channels being measured ($10^3 - 10^5$) in a common MSI dataset. Some compressed forms of visual representation are therefore needed to represent a complete MSI dataset to support expert assessment and hypothesis generation. Methods to represent a complete MSI dataset in a single or a few images can be broadly categorized in dimensionality reduction methods, spatial segmentation (eg. clustering of pixels) and clustering of ion images.
  Dimensionality reduction has mainly been conducted using methods such as PCA, and NMF. PCA and NMF are important approaches in the field of MSI because they can visualize both distinct spatial regions and the associated spectral patterns. Spatial segmentation refers to the partitioning of a dataset into distinct regions with molecularly similar pixels which can be obtained in a number of different ways such as unsupervised clustering of spectra. To reduce pixel-to-pixel variability specific methods for MSI have been developed such as spatially aware clustering and spatially shrunken centroids. In addition to clustering spectra, methods for clustering ion images are important in MSI to group spatially similar ion images [5, 7, 9, 17, 70, 8].

In this thesis, we will focus on three elements in particular, with each having its own chapter dedicated to it:

- **In chapter 2** we focus on dimensionality reduction of MSI data using (uniform manifold approximation and projection (UMAP)). More specifically, we show the added value of non-linear dimensionality reduction for visualizing MSI data and we evaluate the impact of the distance metric chosen in the dimensionality reduction process.
- **In chapter 3** we build further on chapter 2. Starting from the visualizations obtained for a tissue sample using UMAP, we use a bi-directional dimensionality reduction approach to prioritize the molecules driving the obtained visualizations. This approach enables the prioritization of m/z-values in individual tissue samples but also across different tissue samples. The approach is demonstrated for tissue samples obtained from healthy mouse pancreas tissue.
- **Chapter 4** deals with the integration or data fusion of MSI data with the associated microscopy image. Fusing MSI data (high molecular resolution, low spatial resolution) with the associated microscopy image (low molecular resolution, high spatial resolution) brings MSI to a higher spatial resolution, supporting the pathologist in interpreting the obtained molecular information. Instead of using UMAP for dimensionality reduction, we exploite the algorithm for data fusion through a correspondence-aware manifold learning approach.
- **Chapter 5** is the concluding chapter with future perspectives.

This work was published in the following papers:

Smets, Tina; Verbeeck, Nico; Claesen, Marc; Asperger, Arndt; Griffioen, Gerard; Tousseyn, Thomas; Waelput, Wim; Waelkens, Etienne; De Moor, Bart, "Evaluation of distance metrics and spatial autocorrelation in uniform manifold approximation and projection applied to mass spectrometry imaging data", *Analytical chemistry*, 91, 9, 5706-5714, 2019.

Smets, Tina; Waelkens, Etienne; De Moor, Bart, "Prioritization of m/z-values in mass spectrometry imaging profiles obtained using uniform manifold approximation and projection for dimensionality reduction", *Analytical chemistry*, 92, 7, 5240-5248, 2020.

Smets, Tina; De Keyser, Tom; Tousseyn, Thomas; Waelkens, Etienne; De Moor, Bart, "Correspondence-aware manifold learning for microscopic and spatial omics imaging: a novel data fusion method bringing MSI to a cellular resolution", *Analytical chemistry*, 93, 7, 3452-3460,2021.

Chapter 2

Visualization of Mass Spectrometry Imaging data

This chapter is published and reprinted with permission from:

Smets, Tina; Verbeeck, Nico; Claesen, Marc; Asperger, Arndt; Griffioen, Gerard; Tousseyn, Thomas; Waelput, Wim; Waelkens, Etienne; De Moor, Bart, "Evaluation of distance metrics and spatial autocorrelation in uniform manifold approximation and projection applied to mass spectrometry imaging data", *Analytical chemistry*, 91, 9, 5706-5714, 2019.
Copyright 2019, American Chemical Society.

We focus on non-linear dimensionality reduction of MSI data using Uniform Manifold Approximation and Projection (UMAP). More specifically, we show the added value of non-linear dimensionality reduction for visualizing MSI data. To this end we have conducted an extensive evaluation regarding the performance in comparison to other dimensionality reduction methods using spatial autocorrelation and spectral similarity as a benchmark. In addition, we have empirically evaluated a number of different distance metrics, where we show that the choice of a particular distance metric might impact the visualization outcome.

2.1 Evaluation of distance metrics and spatial autocorrelation in Uniform Manifold Approximation and Projection applied to Mass Spectrometry Imaging data

In this work, UMAP is applied for nonlinear dimensionality reduction and visualization of MSI data. We evaluate the performance of the UMAP algorithm on MSI data sets acquired in mouse pancreas and human lymphoma samples and compare it to those of PCA, t-distributed stochastic neighbor embedding (t-SNE), and the Barnes-Hut (BH) approximation of t-SNE. Furthermore, we compare different distance metrics in t-SNE and UMAP and propose the use of spatial autocorrelation as a means of comparing the resulting low-dimensional embeddings. The results indicate that UMAP is competitive with t-SNE in terms of visualization and is well-suited for the dimensionality reduction of large ($>$ 100 000 pixels) MSI data sets. With an almost fourfold decrease in runtime, it is more scalable in comparison with the current state-of-the-art: t-SNE or the Barnes-Hut approximation of t-SNE. In what seems to be the first application of UMAP to MSI data, we assess the value of applying alternative distance metrics, such as the correlation, cosine, and the Chebyshev metric, in contrast to the traditionally used Euclidean distance metric. Furthermore, we propose 'histomatch' as an additional custom distance metric for the analysis of MSI data.

2.1.1 Introduction

Mass Spectrometry Imaging (MSI) is a molecular imaging technology that enables the direct study of the spatial distribution of biomolecular species in a tissue section[28, 105]. MSI provides a very rich biochemical characterisation of a sample, however, a single MSI experiment can lead to gigabytes of complex data. Furthermore, the number of pixels collected in an experiment, as well as the number of m/z bins measured is ever-increasing with improving technology.

Due to the sheer volume and complexity of MSI data, there is a growing need for scalable dimensionality reduction techniques to extract the underlying trends from these data, in order to (i) facilitate human interpretation, (ii) reduce data size and complexity for additional data analysis steps, such as clustering, and (iii) allow for visualisation of these data. Dimensionality reduction techniques such as principal component analysis (PCA), non-negative matrix factorization (NMF) and probabilistic latent semantic analysis (pLSA) [6, 103, 113, 125, 54] have been successfully used to this end, with PCA probably being the

most widely used technique. PCA relies on the determination of orthogonal eigenvectors along which the largest variances in the data are to be found. PCA works very well to approximate data by a low-dimensional subspace, which is equivalent to the existence of many linear relations between the projected data points [75]. In NMF a data matrix, X, is factored into two matrices, W and H, such that by iteratively minimising the residual squared distance between the product of these two factorized matrices, W and H, the resulting product resembles the original data matrix, X, as well as possible ($X \approx WH$)[63]. pLSA, on the other hand, relies on a statistical mixture model and aims to decompose the original data into the underlying latent variables via the iterative Expectation Maximization (EM) procedure, resulting in a set of probability distributions for these discovered latent variables [1].

A common issue in these methods, is the selection of the number of components to be retained for dimensionality reduction. Each of the resulting components contains a part of the total information in the data, and the inclusion of an insufficient number of components will result in a loss of information and data features. This is especially relevant when trying to visualize the data in 2 or 3 dimensions. Another limitation specific to PCA is related to the fact that the interpretation is hampered by the possible presence of negative peaks in the pseudo-spectra, particularly in the context of MSI data where negative values are not expected in the measurements. This disadvantage is alleviated in NMF and pLSA through the enforcement of non-negative constraints.

While linear dimensionality reduction techniques, of which PCA, NMF and pLSA are prominent examples, have shown satisfactory results in the past and have been instrumental in extracting information from MSI data, these models assume that there exist linear relationships between the variables. This is however a very strong assumption which most often can not be imposed as many biological models are inherently non-linear.

These linear methods thus face a limitation with regards to non-linear pattern recognition, resulting in an incomplete capture of the underlying structure. As a result non-linear dimensionality reduction (NLDR) techniques have gained increasing popularity for the analysis of biological data, including MSI, with t-distributed Stochastic Neighbor Embedding (t-SNE) being the leading example [44, 128].

t-SNE has been shown to be very valuable for MSI data analysis, not only for dimensionality reduction in itself but also for visualisation purposes [44, 128]. This is due to the fact that t-SNE is strong at maintaining the local distance between data points and the number of features that can be captured in the reduced space is not restricted by the number of dimensions selected. Because t-SNE is able to embed all features into two or three dimensions, even if more

than two or three features are present, the visualisation of the data is greatly facilitated. This is an important advantage of t-SNE and similar non-linear methods that is absent from methods such as PCA, NMF, pLSA and clustering in general.

A major drawback of most NLDR methods on the other hand is their high computational cost. In t-SNE the pairwise distance matrix needs to be calculated between pixels, which makes the method computationally hard and not suitable for large (>100 000 pixels) MSI datasets. The Barnes-Hut approach therefore seeks to approximate these calculations by using a nearest neighbour's approach but is nevertheless confronted with long runtimes for larger datasets [127]. To mitigate these computational constraints Thomas et al. [123] have presented the application of autoencoders for unsupervised NLDR. This approach treats every pixel as a training example such that the complete dataset is not loaded in memory at once and as long as the number of hidden neurons is smaller than the number of channels in the mass spectrum, this method will be computationally less expensive than t-SNE. While greatly improving on standard t-SNE, the computation time of the weight matrix is still very large (20h for a dataset containing 409 x 404 pixels and 7036 m/z bins or 8Gb in size) [123].

In this paper we show the utility of Uniform Manifold Approximation (UMAP) for the analysis of MSI data. The UMAP algorithm seeks to find an embedding by searching for a low dimensional projection of the data that has the closest possible equivalent fuzzy topological structure (see also UMAP outline, Experimental Methods). The method can therefore be used in a similar fashion as t-SNE ie. for visualisation purposes but also for general NLDR. In order to make a comparison to the current state-of-the art used for non-linear dimensionality reduction and visualisation, t-SNE, and linear dimensionality reduction using standard PCA, possible, we propose the use of spatial autocorrelation. Spatial autocorrelation has previously been studied in the context of MSI by Cassese et al. [29], and refers to the observation that neighbouring pixels in an MSI experiment generally more closely resemble each other than pixels that are far away from each other, and are thus more likely to have a higher correlation between them. This is due to the fact that neighbouring pixels are often located in the same tissue type, and have a higher chance of having a similar function, and thus chemical content.

One of the first steps in the application of NLDR algorithms is the calculation of distances or similarities between the pixels in the original high-dimensional space. This distance metric aims to describe how similar or dissimilar the spectra of each pixel, and thus its chemical content, is to the other pixels. In MSI research, and many other fields, the Euclidean distance is very often used as the de facto distance metric, however, the choice of this distance measure

can have a considerable impact on the results, as meaningful patterns might be lost due to the high-dimensionality of the data, and noise or outliers can become amplified [81]. In this work we therefore compare the effect of different distance metrics, namely the traditional Euclidean distance, cosine distance, Chebyshev distance and the Pearson correlation coefficient between pixels. Furthermore, we highlight 'histomatch' as an additional custom distance metric for the analysis of MSI data [30, 48].

To the best of our knowledge, this is the first time that spatial autocorrelation is used for the relative comparison of the embeddings obtained through different dimensionality reduction methods, using a variety of distance metrics instead of relying on the widely applied Euclidean distance metric.

We demonstrate our method on MSI data collected from mouse pancreas and human lymph node tissue (table 2.1) comparing the resulting embeddings through spatial autocorrelation. Moreover, we evaluate the application of different distance metrics, including the proposed 'histomatch'.

2.1.2 Experimental Section

Mass Spectrometry Imaging (MSI) data acquisition and processing

MSI was performed on mouse pancreatic tissue and on human lymph node samples. For all samples, MSI was done on a Bruker rapiflex MALDI-TOF mass spectrometer. For the mouse pancreatic tissue, cryosections of 7 μm thickness were prepared and mounted on ITO glass slides. Sinapinic acid (SA) was used as matrix and applied using the Bruker ImagePrep. Pixel size was set to 50 μm and the recorded m/z range was 2-20 kDa in positive linear mode. The acquisition speed was 9 pixels/sec with 1000 lasershots/pixel and a laser repetition rate of 10kHz.
For the human lymph node samples, cryosections of 5 μm thickness were prepared and mounted on ITO glass slides. 2,5-Dihydroxybenzoic acid (2,5-DHB) was used as matrix and applied using sublimation. Pixel size was set to 10 μm and the recorded m/z range was 620-1200 Da in positive reflector mode. The acquisition was performed with 200 lasershots/pixel and a laser repetition rate of 10kHz resulting in an acquisition speed of 32 pixels/sec.

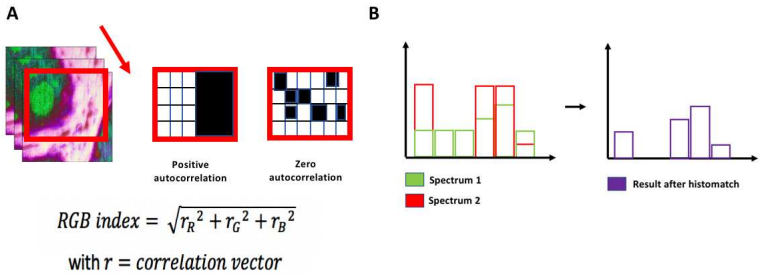


Figure 2.1: A) Spatial Autocorrelation: the resulting 3D embeddings, wherein each dimension corresponds to a color channel, obtained via dimensionality reduction are shifted diagonally according to a number of pixels in order to spatially correlate each pixel given its neighbourhood. An example is given of a perfect positive autocorrelation consistent with a clear binary pattern, in the absence of such pattern as shown in the other example, zero autocorrelation is observed. Given the nature of our embeddings, the correlation is calculated according to each color channel using the RGB index, as shown by the given formula. B) The process of histogram matching illustrated. The high-dimensional data are modified such that its histogram matches the histogram of the reference data.

Sample	Pixels	m/z bins
Lymphoma P1	572 173	8000
Lymphoma P2	552 701	8000
Pancreas M1	10 606	14 000
Pancreas M2	14 791	14 000

Table 2.1: Overview of dimensionality associated with the used datasets

Data processing of Mouse Pancreas and Human Lymphoma datasets

Data Modeling. All data was normalized using Total Ion Count (TIC). Manifold learning approaches were used to embed high-dimensional data in a low-dimensional space for data visualisation and investigation of non-linear relations in the data, in addition PCA was used for comparative reasons. The following three methods were used to map the data to three dimensions: (1) PCA; (2) t-SNE, and the Barnes-Hut (BH) approximation thereof, (3) UMAP. The t-SNE mapping to three dimensions was done using the default settings as discussed by van der Maaten et al [128], apart from

the different distance metrics that were evaluated. The UMAP mapping to three dimensions was performed using the Python implementation as provided by the paper's author, L. McInnes, according to the default settings n_neighbors=15, gamma=1.0, n_epochs=None, alpha=1.0, init='spectral', spread=1.0, min_dist=0.1, a=None, b=None, random_state=None, metric_kwds=, verbose=True, see https://github.com/lmcinnes/umap), except for the different distance metrics that were evaluated. The distance metrics used here are Euclidean, correlation, cosine, and Chebyshev for which the detailed formulas are shown in figure A.1. For UMAP one additional distance metric was evaluated: 'histomatch'. The cosine and histomatch distance metric rely on the assumption that MSI data contains no negative values. Due to the high computational burden associated with the larger lymphoma datasets, the comparison of distance metrics was mainly carried out using the UMAP algorithm for this data. For the analyses where a t-SNE implementation was used on these large datasets an initial dimensionality reduction step (n=100) with PCA was required to make these analyses feasible. A more extensive comparison with regard to the distance metrics in t-SNE implementations has therefore been done using the smaller pancreas samples.

To run standard t-SNE, we relied on the Python code as provided by Laurens Van Der Maaten (https://lvdmaaten.github.io/software/) and for the Barnes-Hut approximation of t-SNE, we relied on the Scikit-learn implementation which uses the Barnes-Hut approximation by default. This implementation, as provided in Scikit-learn, was used for all analyses except for the results shown in figure 2.3. Here a comparison was made between the standard t-SNE implementation, the Barnes-Hut approximation and UMAP. All experiments were run on an Intel(R) Xeon(R) CPU E5-2660 v2 2.20GHz machine with 10 cores and 128 Gb RAM.

UMAP outline. From a general outline UMAP uses local manifold approximations and assembles together their local fuzzy simplicial set representations to form a topological representation of the high dimensional data. Given some low dimensional representation of the data, a similar process can be used to form an equivalent topological representation. The layout of the data representation in the low dimensional space is then optimized through the minimization of the cross-entropy between the two topological representations [83].

An important component of the algorithm is the cost function for the optimization of the embedding through minimization of the fuzzy set cross entropy:

$$C_{UMAP} = \sum_{i \neq j} v_{ij} log\left(\frac{v_{ij}}{w_{ij}}\right) + (1 - v_{ij}) log\left(\frac{1 - v_{ij}}{1 - w_{ij}}\right) \tag{2.1}$$

Algorithm 1 UMAP algorithm

function UMAP(X, n, d, min-dist, n-epochs)
  for all $x \in X$ **do**
    fs-set[x] $\leftarrow$ LocalFuzzySimplicialSet(X, x, n)
  top-rep $\leftarrow \bigcup_{x \in X}$ fs-set[x] ▷ We recommend the probabilistic t-conorm
  $Y \leftarrow$ SpectralEmbedding(top-rep, d)
  $Y \leftarrow$ OptimizeEmbedding(top-rep, Y, min-dist, n-epochs)
  return Y

Figure 2.2: General outline of the UMAP algorithm. For the detailed functions related to the construction of the local fuzzy simplicial sets, the determination of the spectral embedding, and optimizing the embedding with regard to fuzzy set cross entropy, we refer to the original article by McInnes et al. from which this general outline is adopted. [83]

There is some resemblance to the Kullback-Leibler divergence (t-SNE cost function (2)) in the first part of the equation, however important to note here is that UMAP does not use the same definitions for v_{ij} and w_{ij} wherein i and j refer to two objects in the high (v_{ij}) and low (w_{ij}) dimensional space.

$$C_{t-SNE} = \sum_{i \neq j} p_{ij} log \frac{p_{ij}}{q_{ij}} \tag{2.2}$$

v_{ij} refers to the local fuzzy simplicial set memberships defined in the high dimensional space, based on the smooth nearest neighbour's distances, whereas w_{ij} refers to the low dimensional similarities between i and j. The t-SNE cost function (2) on the other hand seeks to minimize the Kullback-Leibler divergence between the joint probability distribution, p_{ij} in the high dimensional and the joint probability distribution, q_{ij}, in the low-dimensional space. The fact that both p_{ij} and q_{ij} require calculations over all pairs of points imposes a high computational burden on t-SNE. Therefore the improvements in the efficiency of methods such as the Barnes-Hut approximation and UMAP focuses on approximating these quantities. UMAP achieves an efficient approximate k-nearest-neighbour computation via the Nearest-Neighbor-Descent algorithm [41] for which an empirical complexity of O($p^{1.14}$) was reported. This in comparison to t-SNE O(p^2) and BH t-SNE O($plogp$), wherein p refers to the number of pixels. While the aforementioned time complexities indicate a better asymptotic scaling for BH t-SNE in comparison to UMAP, it is important to remind the reader that these complexities do not reflect the different constant time multipliers between BH t-SNE and UMAP. We thus want to emphasize that, for any contemporary data set of common size (i.e. <5M pixels), UMAP exhibits significantly shorter

run times than BH t-SNE using the implementations we benchmarked. For a detailed outline of the complete mathematical foundations of the algorithm we would like to refer the interested reader to the original manuscript, and in particular to appendix C, for a detailed comparison between UMAP and t-SNE. [128, 83]

Custom distance metric. The 'histomatch' distance metric is inspired by the Histogram Matching algorithm which forces the intensity distribution of an image to match the intensity distribution of a target [30, 48]. The 'histomatch' distance metric is specified according to the following formula:

$$d(x, y) = 1 - \sum_i \max(0, \min(x_i, y_i)) \tag{2.3}$$

Where x and y are spectra from different pixels. A schematic illustration is given in figure 2.1 B.

Data visualisation. For all manifold learning approaches, the locations of pixels were translated to RGB-color coding by varying the red, green and blue intensity linearly on the three independent axes, such that the minimum value on the axis is represented by a color intensity of 0, and the maximum value on the axis has intensity 255, which can be normalized to a scale of 0 to 1), this visualisation is referred to as hyperspectral visualisation [44].

Autocorrelation. Given the spatial nature of MSI data, a certain degree of spatial correlation is expected to occur. This means that neighbouring pixels are more likely to be correlated to each other than pixels located at a further distance [29]. As shown in figure 2.1, we have evaluated the obtained embeddings by subjecting them to a spatial autocorrelation function. By diagonally shifting each embedding over a number of pixels (ranges of 1-15 and 1-50 pixels), we can determine the correlation of pixels to their neighbours according to incremental distance. Technically, the autocorrelation function is used to compute the dot product of the original image (hyperspectral visualisation of tissue embedding in 3D) with the shifted image for increasing number of shifts in pixels. For each shift from the original image, the RGB autocorrelation index was calculated according to the following formula:

$$RGB_{index} = \sqrt{(r_R)^2 + (r_G)^2 + (r_B)^2} \tag{2.4}$$

The correlation vector, r, represents the correlation between shifted and original image for the Red (R), Green (G) and Blue (B) color components respectively.

Autocorrelation was performed using Python and all results are normalised using min/max normalisation.

Spectral similarity or chemical distance information. To visualise the relationship between the spatial autocorrelation of the embedding and the chemical data, we followed a similar approach as above. This was done by calculating the mean distance between the MSI spectral or intensity data (according to the relevant distance metric used) rather than correlating the pixels of the resulting embeddings. Here we assume that for pixels that are located closer to each other, that the spectral similarity or chemical relatedness is more likely to be larger for these pixels than for the ones located further from each other. We use this assumption to support the relative comparison of the embeddings obtained on the same tissue section but via different algorithms eg. t-SNE versus UMAP, and in particular as a benchmark for the introduced autocorrelation measure. From here on, we will refer to this measure as spectral similarity.

2.1.3 Results and Discussion

UMAP enables the analysis of large MSI datasets

We compare the performance of the UMAP algorithm for the analysis of pancreas and lymph node samples to the performance of t-SNE, the Barnes-Hut (BH) approximation and PCA (figures 2.3 and 2.4).

The embeddings obtained via the different manifold learning approaches are visualised using the RGB color scheme such that the colors in the image depend on each pixel's location in the model space. Hence similar colors or RGB values represent similar spectra or biochemical patterns.

UMAP strongly outperforms t-SNE but also its faster counterpart, BH t-SNE, with almost a fourfold decrease in runtime. A comparison of the runtimes obtained by the different algorithms is given in figure A.2. Besides being computationally more efficient, UMAP is also capable of delivering embeddings with at least the same quality as t-SNE, as can be seen in figures (2.3 and 2.4).

Figure 2.3 shows the hyperspectral visualisation of the two analysed pancreas samples for the 3D embeddings obtained using t-SNE, BH t-SNE and UMAP. For this comparative analysis all implementations were used with standard parameters. The original t-SNE implementation is not intended to be used with alternative distance metrics, which is why we used the Euclidean distance metric to make a cross-comparison possible.

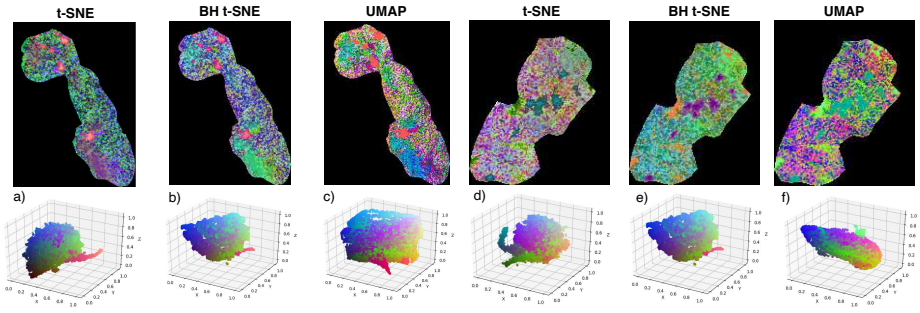


Figure 2.3: Comparison of t-SNE (a,d), Barnes-Hut (b,e), and UMAP (c,f) embeddings of two different pancreas tissue samples M1 (a-c) and M2 (d-f). Shown on top are the 3D embeddings of the tissues using hyperspectral visualisation with the corresponding scatter plots underneath. The Euclidean distance metric was used for all embeddings.

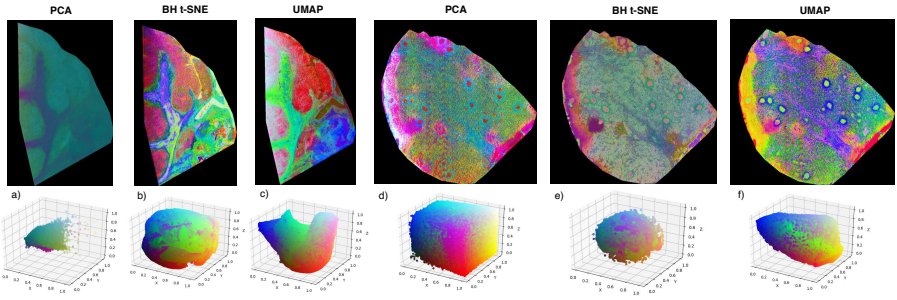


Figure 2.4: Comparison of PCA (a,d), Barnes-Hut (b,e), and UMAP (c,f) embeddings of two lymphoma tissue samples P1 (a-c) and P2 (d-f). Shown on top are the 3D embeddings of the tissues using hyperspectral visualisation with the corresponding scatter plots underneath. The Euclidean distance metric was used for all embeddings.

In figure 2.4, an overview is given of the analysed lymphoma samples. For the lymphoma datasets used in this paper (552,701 and 572,173 pixels x 8,000 m/z bins respectively), BH t-SNE required more than two weeks of computational time when not using another dimensionality reduction technique in advance. In an ideal world using PCA (or similar dimensionality reduction algorithms) in advance should not be necessary, because these methods are strong at preserving the global distance structure within the data, whereas methods such as t-SNE focus on the preservation of the local distance. In our case the composition of biological tissue is very heterogeneous, hence we are interested in preserving the local distances as well as possible. Ideally however we get the best of both worlds and the global structure is preserved as well. The UMAP algorithm is said to be competitive with t-SNE in terms of visualisation and to preserve more of the global structure in the data, while being more scalable towards large datasets [83]. We have indeed observed that for our larger datasets, UMAP is able to reduce 8,000 dimensions for over 500,000 data points to 2 or 3 dimensions in approximately 6 hours. In terms of memory usage for UMAP the median measured across 5 runs was 1.7 Gb for the pancreas tissue and 25 Gb for the lymphoma tissue sample. This is comparable to the observed median (n=5) memory usage for BH t-SNE (1.4 Gb) for the pancreas samples. Because of the high computational burden associated with BH t-SNE for the large lymphoma samples, an initial dimensionality reduction to 100 dimensions using PCA was needed. Therefore a fair comparison with UMAP in this regard was not possible.

Overall, UMAP shows a good performance and in comparison to PCA, as shown in figure 2.4, using NLDR techniques has the major advantage of compressing the spatial and molecular information into three dimensions resulting in a detailed visualisation.

2.1.4 Towards a relative comparison of the embeddings obtained by UMAP and t-SNE via spatial autocorrelation

Because it is difficult to compare the embeddings obtained through UMAP versus the other methods solely by visual inspection, we propose the use of spatial autocorrelation. This approach reflects the correlation between the values of a single variable strictly due to the proximity of these values in a geographical space by introducing a deviation from the assumption of independent observations of classical statistics [51]. For MSI data it is commonly assumed that close neighbouring pixels are likely to be more correlated than their more distant neighbours [29]. Our experiments empirically corroborate this assumption, since the correlation between a pixel and its neighbouring pixels is more likely to decrease with increasing distance.

Likewise, we assume that for pixels that are located closer to each other, the spectral similarity or chemical relatedness is more likely to be larger for these pixels than for the ones located further from each other. We therefore use this assumption as a benchmark indicator for the autocorrelation measured.

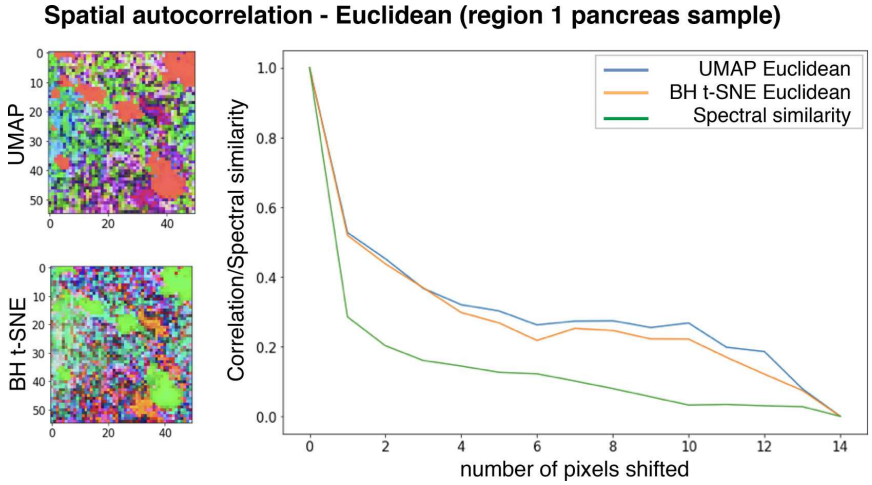


Figure 2.5: Spatial autocorrelation of the M2 pancreas embeddings is shown as the correlation in function of increasing distance or incremental number of pixels shifted. As expected the correlation values are maximal for a zero shift and decrease in function of the distance. Shown are the results for one region in the tissue when the Euclidean distance metric was used in the BH t-SNE (orange) and UMAP (blue) algorithms. Each graph also includes the spectral similarity between pixels according to the Euclidean distance metric (green) which serves as a benchmark indicator for the spatial autocorrelation. The same approach is shown in figure A.3 (a) for a larger region. The graphs show clearly that in every case the spatial autocorrelation behaves similarly for both BH t-SNE and UMAP which supports the fact that UMAP delivers embeddings of at least the same quality as BH t-SNE.

In figure 2.5, a comparison is given for the spatial autocorrelation of embeddings obtained based on the pancreas M2 sample using UMAP and BH t-SNE. Shown is the graph for a selected region using the Euclidean distance metric. As expected the correlation values are maximal for a zero shift and decrease in function of the distance. In addition to the spatial autocorrelation, the spectral similarity is also shown as a benchmark indicator. The fact that the spatial autocorrelation shows a similar behaviour for both BH t-SNE and UMAP further supports the fact that UMAP delivers embeddings of at least the same

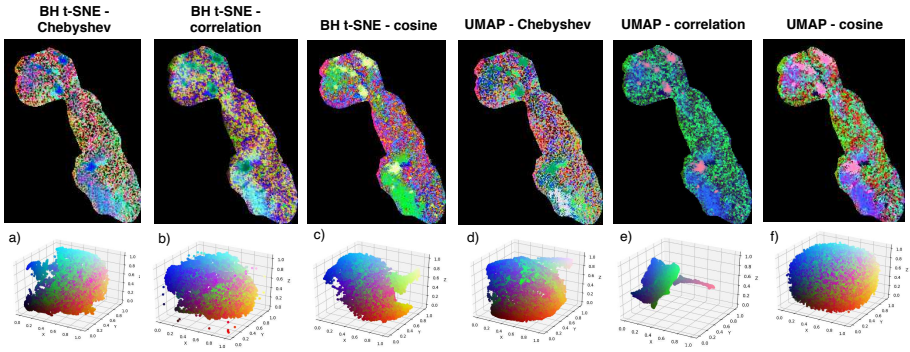


Figure 2.6: Comparison of the Chebyshev, correlation and cosine distance metrics using BH t-SNE and UMAP for the M2 pancreas samples. Shown on top are the 3D embeddings of the tissues using hyperspectral visualisation with the corresponding scatter plots underneath.

quality as BH t-SNE. This is also reflected in the additional results regarding the spatial autocorrelation as shown in figures A.3-A.6.

2.1.5 The underestimated importance of the distance metric

Besides the value of NLDR methods for the analysis of MSI data we can also question the suitability of the Euclidean distance measure in these analyses. As Aggarwal et al. [2] have noted previously, the choice of a particular distance metric may significantly improve the results of standard algorithms. Given that in the high-dimensional space the data becomes sparse and the concept of proximity or distance becomes less meaningful, the authors have shown that the Euclidean distance metric is not an ideal metric to be used in the high-dimensional space. Taken this information into account, we have evaluated the effect of using different distance metrics when applying BH t-SNE and UMAP to our data.

In figure 2.6, a comparison is made between UMAP and BH t-SNE using the Chebyshev, correlation and cosine distance metric respectively.

Figure 2.7 shows the embeddings for the lymphoma samples using the Chebyshev, correlation and cosine distance metrics. For the P2 sample it is clear to see how using the cosine metric strongly increases the level of detail observable in the resulting embedding. Moreover, as shown in figures 2.7 (e-f), using the correlation and cosine metric facilitates the detection of

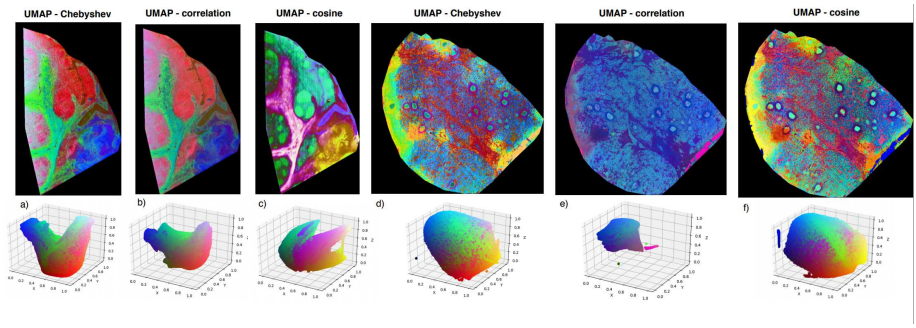


Figure 2.7: Comparison of the Chebyshev, correlation and cosine distance metrics using UMAP (a-c) for the P2 and P1 (d-f) lymphoma samples. Shown on top are the 3D embeddings of the tissue using hyperspectral visualisation with the corresponding scatter plots underneath. Panels e-f) show the importance of choosing a good distance metric: using correlation and the cosine distance metric, a series of outliers could be detected. Figure 2.8 shows the impact of removing these outliers on the hyperspectral visualisation due to the improved color space utilisation.

outliers which take up a major part of the available colorspace. Therefore their removal strongly improves the hyperspectral visualisation because it enables the utilisation of the complete colorspace. This is visualised in figure 2.8, where the outliers detected by using the correlation distance metric were removed as an example.

Important to note here is that all data was TIC normalised before applying the dimensionality reduction methods, because this step is commonly applied within the MSI field. However to ensure that this normalisation does not affect our observations we have also included a comparison of the different distance metrics applied to the data without TIC normalisation. These results are shown in figures A.7-A.10, A.15-A.18. Overall, we can conclude the cosine distance metric, independently of whether the data was TIC normalised or not, delivers good results across the different methods and tissues. This is in agreement with previous research by Winderbaum et al. who have observed that k-means clustering of MSI data using the cosine distance lead to superior results as opposed to the Euclidean distance metric [134].

Custom distance metric- histogram matching

As shown above, the choice of a specific distance metric may significantly alter the resulting embeddings. We have therefore evaluated the performance of an

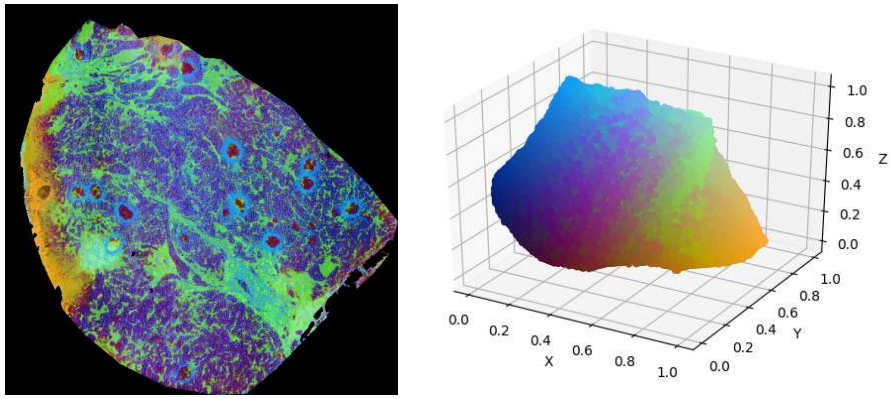


Figure 2.8: Hyperspectral visualisation of the lymphoma P1 sample with corresponding scatterplot. The improved visualisation is due to the removal of outliers detected by using the correlation distance metric in the UMAP algorithm as shown in figure 2.7 (e). The removal of these outliers makes the utilisation of the complete color space possible for the remaining points resulting in an enhanced hyperspectral visualisation of the data.

additional distance metric called histomatch. Histogram matching is a method that is often used in computer vision and forms an approximation of correlation, making it an interesting candidate for the analysis of MSI data.

The resulting embeddings for a lymphoma sample obtained according to the histomatch metric, in comparison to the cosine metric, are shown in figure 2.9, in figure A.15 the results are shown for the pancreas sample. As shown in figure 2.9, the hyperspectral visualisation of the P2 lymphoma sample is very similar to the one obtained using the cosine metric. This is also reflected in the spatial autocorrelation results for region 1 and 2. Therefore histomatch can be a valuable alternative distance metric for the analysis of MSI data. Our observations show that for different datasets divergent distance metrics can shed an alternative light on the same data, which makes experimenting with these metrics worthwhile.

2.2 Conclusion

We have shown that UMAP yields superior runtimes compared to t-SNE and Barnes-Hut t-SNE for the analysis of MSI data, while obtaining embeddings

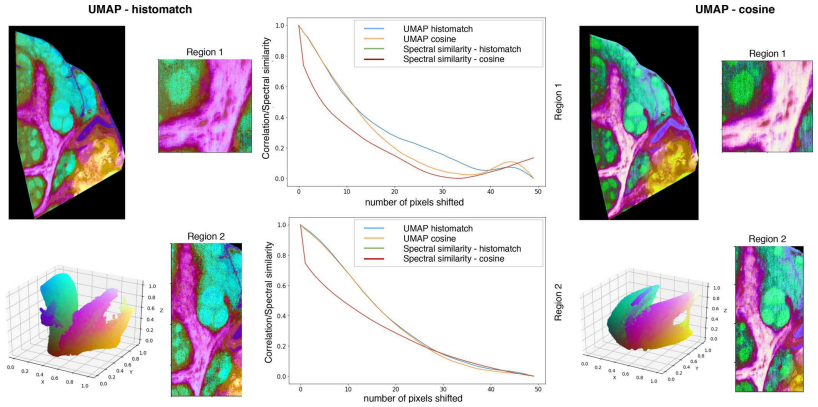


Figure 2.9: Hyperspectral visualisation of P2 lymphoma UMAP embedding using a custom distance metric called 'histomatch' with corresponding scatterplot underneath compared with the results obtained using the cosine distance metric. The graphs show the behaviour of the embedding in terms of spatial autocorrelation for the two selected regions according to the histomatch (blue) and the cosine distance metric (orange). The spatial autocorrelation is supported by the spectral similarity which is shown in green for the histomatch distance and in red for the cosine metric. The spectral similarities measured according to the cosine and histogram metric overlaps completely, this is supported by figure A.16 together with the spatial autocorrelation of pixels in terms of their raw chemical information or measured intensity values calculated based on the histomatch distance metric (green). Both the hyperspectral visualisation and the spatial autocorrelation plots resemble the results obtained with the cosine distance metric.

that are of at least the same quality as those obtained by (BH) t-SNE. In addition, we have demonstrated that spatial autocorrelation can be used for the relative comparison of the results obtained by different NLDR methods. Moreover, we have highlighted the importance of using different distance metrics for performing NLDR. Overall, we can conclude that for the analysis of MSI data the correlation and cosine distance metric achieve the best results and going for the Euclidean distance metric as the standard might not be the best idea. Furthermore we have presented 'histomatch' as an additional distance metric for the analysis of MSI data. In conclusion, we have shown the value that UMAP, spatial autocorrelation and different distance metrics can bring to the analysis of MSI data by which we hope to pave the way for future research in this area.

Chapter 3

Prioritization of m/z-values in mass spectrometry imaging data

This chapter is published and reprinted with permission from:

Smets, Tina; Waelkens, Etienne; De Moor, Bart, "Prioritization of m/z-values in mass spectrometry imaging profiles obtained using uniform manifold approximation and projection for dimensionality reduction", *Analytical chemistry*, 92, 7, 5240-5248, 2020,
Copyright 2020, American Chemical Society.

We build further on the previous chapter, starting from the visualizations obtained for a tissue sample using UMAP, we apply a bi-directional dimensionality reduction approach to prioritize the molecules driving the obtained visualizations.This approach enables the prioritization of m/z-values in individual tissue samples but also across different tissue samples as demonstrated on pancreas tissue obtained from wildtype mice.

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

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

3.1.1 Introduction

Mass spectrometry imaging (MSI) enables the untargeted measurement of biomolecular species and the visualization of their spatial distribution in a variety of tissue sections [28, 105] . The combination of these elements gives rise to a powerful tool that allows us to dissect and characterise the biological composition of tissues both in health and disease. To discern the spatial pattern of molecules measured with MSI, their distribution is typically visualized in the form of ion images. These are visualizations that employ a pseudo-color scale to the mass spectral intensities associated with a particular m/z -value resulting in a heat map that is intuitive to interpret. However with the large number of features being measured, visualizing thousands of ions or m/z-values makes it infeasible to gain rapid insight into potential patterns present in the data.
For this reason we can rely on a number of dimensionality reduction and clustering techniques that improve the interpretability of the data through a comprehensive decomposition or visualization thereof. Methods such as principal component analysis (PCA), probabilistic latent semantic analysis (pLSA) and non-negative matrix factorization (NMF) have been particularly useful in this regard [131]. While PCA seeks to determine the orthogonal eigenvectors associated with the largest variance in the data [76], NMF tries to resemble the original data matrix, X, as well as possible through iterative minimization of the residual squared distance between the products of the two factorized matrices, W and H ($X \approx WH$) [113, 125]. Yet another approach is taken by pLSA, which relies on a statistical mixture model to decompose the data of underlying latent variables via the iterative expectation-maximization (EM) method [54]. A variety of applications have also shown the value of a neural network approach in the form of Self - Organizing Maps (SOMs) [67, 49, 68].
Non-linear dimensionality reduction methods such as t-distributed Stochastic Neighbour Embedding (t-SNE) on the other hand have gained popularity as they were shown to outperform methods such as PCA and SOM due to their strong visualization capabilties [44, 128]. Uniform Manifold Approximation and Projection (UMAP) is an example of a recent non-linear dimensionality reduction method that is comparable to t-SNE but more scalable towards larger MSI datasets [83, 128]. Like t-SNE, UMAP does not impose the strong

assumption of a linear relationship between variables that is made by techniques such as Principal Component Analysis (PCA) [76], which is beneficial when working with biological models that are inherently non-linear.

While we have already shown the value of UMAP for the dimensionality reduction of MSI data in earlier work, we want to stress the fact that this method captures all features in the reduced space irrespectively of the number of components selected [115]. This is an important aspect when visualizing data in two or three dimensions because the number of features embedded in the reduced space is not restricted by these two or three dimensions, even if more features are present in the data. This characteristic empowers strong data visualizations and is absent from methods such as PCA or clustering approaches.

A downside of non-linear methods however is their limited explicability Although providing excellent and rapid insight into the major patterns or trends present in a tissue, they do not reveal which ions or m/z- values are driving these patterns and observations. In the end, it is through the identification of ions that are co-localized with certain regions of interest or differ in expression between samples of different conditions from which we can elucidate disease mechanisms or facilitate biomarker discovery [87].

It is with this idea in mind that we, as illustrated in figure 3.1, propose a workflow starting from the hyperspectral images obtained using UMAP to identify those ions that are driving the different profiles as visualized by a color gradient. We do this by extracting these profiles according to their corresponding color of interest, followed by a spectral and spatial-driven prioritization of m/z-values. For the spectral part we extract the median peak spectral intensities from the data associated with the pattern of interest and apply peak picking to select those peaks that differ the most from the background profile (i.e. the residual tissue pixels). While the latter enables us to build a spectral prioritization of m/z-values, we are also interested in a spatial prioritization such that we can detect co-localized ions that are specific for the extracted profile. For this spatial prioritization we start from the two-dimensional embedded pixel space, which we cluster using the Hierarchical Density-based Spatial Clustering for Applications with Noise (HDBSCAN) algorithm [84] to group similar ion images. This enables us to correlate the selected profile to the average correlation per cluster and gain rapid insights into promising co-localization patterns. We specifically aim to include both the spectral and spatial information because this enables a more comprehensive characterization of the information present in the different profiles, which would not be possible when considering these sources of information individually.

We demonstrate our method on MSI data collected from healthy mouse

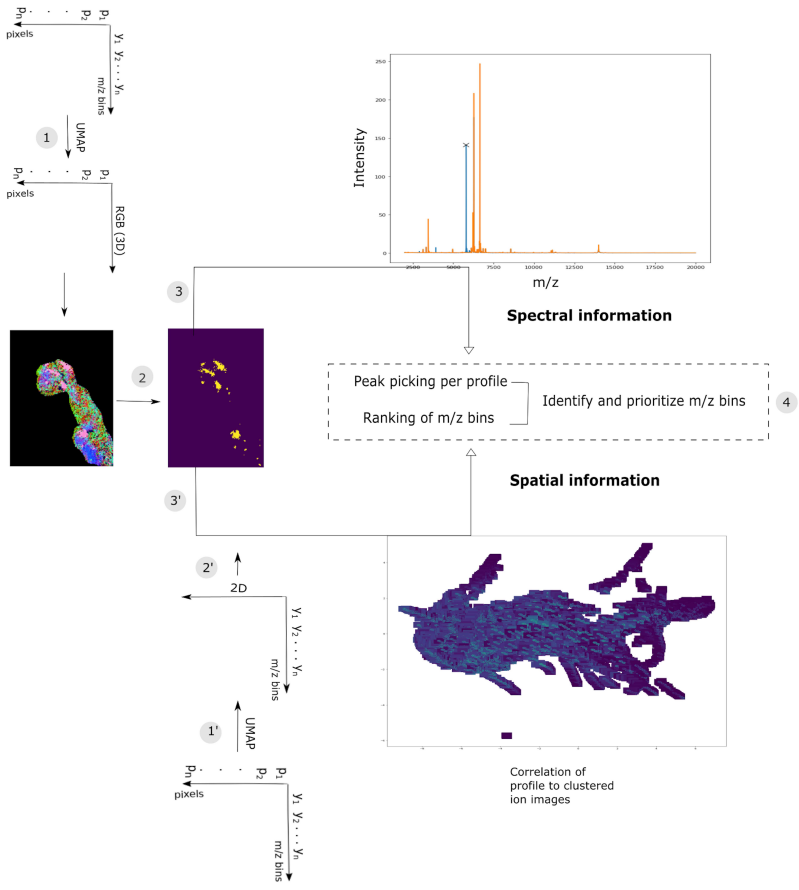


Figure 3.1: Method overview. We start by reducing the dimensionality of the MSI feature space (m/z bins) to three dimensions (1). This three-dimensional embedding is used for the hyperspectral visualization of the data which is then used to extract the different profiles according to their color representing similar biochemical content (2). The binary representation of a selected profile is then used (3) to identify peaks driving this profile in comparison to the overall tissue (spectral information). In parallel, this binary representation is used to match this pattern through correlation with clusters of similar ion images (3'). These are obtained through clustering (2') of the two-dimensional embedding of the pixel space rather than the m/z space (1'). By ranking the identified peaks obtained through the spectral and spatial information we are able to get a comprehensive view on the peaks that are driving the hyperspectral visualization obtained using UMAP (4). The combined ranking is obtained by combining the spectral and spatial ranking per tissue by giving a higher priority to those m/z-values that are i) prioritized both in a spatial and spectral manner within the same tissue and ii) observed in the same profile across multiple tissues

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

3.1.2 Experimental Section

3.1.3 Mass Spectrometry Imaging (MSI) data acquisition and processing

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

3.1.4 Data processing of Mouse Pancreas datasets

Data modeling and visualization. All data was normalized using total-ion count (TIC). The Python language was used for all data analyses. UMAP was used for dimensionality reduction to three dimensions, followed by hyperspectral visualization. By hyperspectral visualization we mean that based on their position in the obtained 3D embedding, the pixels were translated to RGB color coding by varying the red, green, and blue intensities linearly on the three independent axes, such that the minimum value on an axis is represented by a color intensity of 0, and the maximum value on an axis has an intensity of 255, which can be normalized to a scale of 0 to 1 [44]. The UMAP mapping to three dimensions was performed using the Python implementation (https://github.com/lmcinnes/umap) with the default parameters (n_neighbors=15, gamma=1.0, n_epochs=None, alpha=1.0, init='spectral', spread=1.0, min_dist=0.1, a=None, b=None, random_state=None, metric_kwds=, verbose=True) except for the cosine and Chebyshev distance metric. An elaborate evaluation of the performance of UMAP and different distance metrics can be found in our previous work [115].

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

Spectral information. Based on the biochemical profiles corresponding to certain colors, as selected using the k-means procedure, we determine the location median peak intensities of these profile-specific pixels. The peak finding and prominence measures give insight into the difference between profile-specific peaks in comparison to the other regions of the tissue. For this purpose, the find_peaks and peak _prominences functionality as provided in the Python SciPy package and signal processing library were used. [62] The prominence of a peak measures how much a peak stands out from the surrounding baseline of the signal and is defined as the vertical distance between the peak and its lowest contour line, whereas the find_peaks function finds all local maxima by simple comparison of neighbouring values.

Spatial information. For the biochemical profiles corresponding to certain color gradients, as selected using the k-means procedure, we also construct a binary image for this profile. This binary image is then used to rank all ion or m/z-images according to their Spearman correlation to this specific profile of interest. For this comparison, we first reduce the pixel-space to a two-dimensional embedding with UMAP (default parameters and cosine distance metric) which we then subject to the HDBSCAN clustering algorithm [84] to group similar ion or m/z-images. For the clusters obtained we calculate the average correlation value per cluster to the selected profile and we return the top N ranked ion images per cluster.
HDBSCAN is as the name suggests a hierarchical clustering variant of the DBSCAN algorithm [84, 43]. HDBSCAN aims to identify the presence of dense regions by using sliding windows that move towards the high density points. Important advantages of HDBSCAN over other clustering algorithms are i) the ability to identify clusters of data with varying shape ii) robustness to clusters of different densities and iii) the ability to perform density-based clustering, which eliminates the need to specify the number of clusters desired. The only hyperparameter that needs to be specified upfront is the minimum cluster size. This parameter is the primary parameter to effect the resulting clustering and intuitive to interpret since it refers to the smallest size grouping that should be considered as a cluster [84]. Our intention here is to apply clustering such that when correlating a certain profile or pattern of interest

to all available ion images, we can identify as many trends as possible by grouping together similar ion images, for which each cluster can still be evaluated in depth if desired. We therefore prefer a small minimal cluster size over a large one. To run the HDBSCAN algorithm we used the scikit-learn implementation (https://github.com/scikit-learn-contrib/hdbscan) using the default parameters and a minimal cluster size of 5.
The combined ranking is obtained by combining the spectral and spatial ranking per tissue by giving a higher priority to those m/z-values that are i) prioritized both in a spatial and spectral manner within the same tissue and ii) observed in the same profile across multiple tissues

Histogram equalization for contrast enhancement of ion or m/z-images. An ion image represents the mass spectra insensities for a particular m/z- value. These ion images are typically visualized using a pseudo-color scale wherein gradually changing colors are assigned to the intensities. Using this approach can lead to hot spots of pixels with artificially high intensities, distorting the pseudo-color scale such that other pixels will lack contrast. As illustrated in previous work [6] advanced contrast-enhancing procedures like histogram equalization are useful to alleviate this problem. We therefore applied Contrast Limited Adaptive Histogram Equalization (CLAHE) [142] to the ion images prior to visualization for contrast enhancement. To this end we relied on the CLAHE implementation in the scikit-image library for Python (https://scikit-image.org/) [129].

3.1.5 Results and Discussion

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

The hyperspectral visualizations for these samples are the starting point of our analysis and are shown in figure 3.2. To illustrate our approach, we show the results for two segmentations obtained from these images. The first example, as shown in figure 3.3, correlates with endocrine tissue more specifically, the anatomical location of the pancreatic islets or the Islets of Langerhans. The (binary) segmentations in panels D-F correspond to the pink, dark green and light blue regions of the hyperspectral visualizations of samples 1,2 and 3 respectively. These regions clearly overlap with the pancreatic islets, which are annotated in the H&E stainings shown in panels A-C. The combined rankings are presented below each segmentation. These

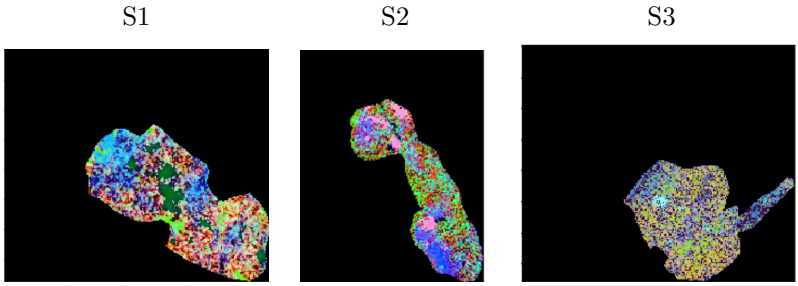


Figure 3.2: Overview of the hyperspectral visualizations for the three healthy pancreas tissue samples. From left to right are shown samples S1, S2 and S3. The different colors in the image are a representation of each pixel's location within the embedded space. As such similar colors or RGB values represent similar spectra or biochemical patterns within one sample. This is however not necessarily the case for different tissues hence the green profile in S1 is completely independent from the green profile in S2 or S3.

rankings are obtained by combining the spectral and spatial ranking per tissue by giving a higher priority to those m/z-values that are i) prioritized both in a spatial and spectral manner within the same tissue and ii) observed in the same profile across multiple tissues. The top three m/z-values are the same for all samples, while the top four is completely the same for samples 1 and 3. In all tissues the highest ranked ion is the m/z- value around 5800 Da, which likely corresponds to the full-length active insulin molecule after cleavage of the c-peptide from pro-insulin. This peak with a major abundance around 5800 Da, is observed to lie within the anticipated m/z-deviations, as can be expected from MALDI measurements, across all tissues. Moreover, the region-specific localization that corresponds to the islets of Langerhans, as shown also in the H&E stainings, further strenghtens this association (figures 3.3 and 3.4). For the values around 5840 and 5821 Da a putative association could be made to ATP5E (ATP synthase subunit epsilon) and VIP (Vasoactive Intestinal Peptide) respectively. Previous research regarding the mouse pancreatic islet proteome has highlighted the enrichment of proteins that play a role in oxidative phosphorylation [97] which might explain the detection of ATP5E. In addition, the stimulatory effect of VIP on insulin has previously been reported [3, 21]. The ubiquitously distributed m/z- value of 6276.22 detected in the third sample is likely to correspond to a part of the mouse kallikrein protein. This protein is known to play a role in the pancreas metabolism and has been observed both in the endocrine and exocrine pancreas.
Another key-player involved in glucose metabolism is glucagon which could putatively correspond to the m/z values around 3484. This molecule is observed

in all samples with a distribution around the pancreatic islets (Figure B.13) and ranked around position 15 for the three tissues which can be well explained given its lower abundance. This indicates that the results are likely to have a biological meaning in addition to the confirmed spatial co-localization as shown in the corresponding ion images. Moreover using this method a good and robust correspondence between tissues is observed as illustrated by the combined ranking. The particular ion images associated with the top three m/z-values for sample 1 are shown in figure 3.4 and are available for all samples in the supplementary information (B.1-B.12). Figure 3.4 shows additional confirmation regarding the spatial co-localization of the identified ions with the dark green segment, which is also supported by the overlay between the H&E and hyperspectral visualization.

For the second segmentation, which corresponds to the light and dark blue regions of the hyperspectral visualizations, we observe a similar prioritization across the different tissues corresponding to exocrine pancreas. As shown in figure 3.5, the top three of m/z- values is almost identical for the different tissue samples. The assumption that this profile represents exocrine tissue is supported by the fact that the m/z -value around 6651 Da has been associated with the acini in earlier work [88]. Moreover, the MASCOT engine identifies a match with Selenocysteine lyase (6643 Da), a protein associated with high expression in the exocrine pancreas according to the Human Protein Atlas (HPA). For the value around 6691 Da, we could identify a putative association to Single-pass membrane protein with coiled-coil domains 4 (SMCO4), a protein for which also a high expression in the exocrine pancreas is noted in the HPA. In addition to the biological relationship, the ion images also show a similar spatial distribution to the segmentation obtained from the hyperspectral visualizations, as shown in figure 3.6 for sample 1 and in the supplementary information for the other samples (B.1-B.12).

These kind of observations can be very insightful, not only to study similar molecular compositions across different samples, but also to assess the quality of the hyperspectral visualizations obtained using UMAP. An example of the latter use case is shown in figures 3.7-3.8 where instead of the cosine distance metric, the Chebyshev distance metric was used for constructing the embeddings. In grey, we can discern an additional profile within the blue region. Because we are not able to detect any ion images corresponding to this grey profile, we can assume that the cosine metric achieves a better representation of the underlying data. Moreover using this approach one can guide the level of detail that is desired in two ways i) by selecting the number of colors one wants to extract and ii) by selecting the number of ion images to be returned per cluster. This is not only important to give us more confidence in the visualizations obtained using UMAP but can also strongly support us

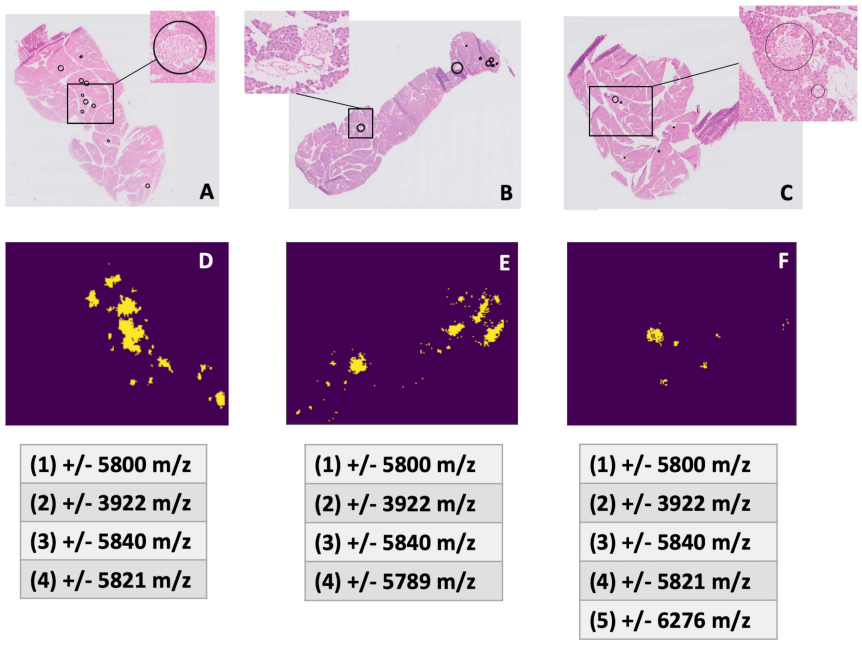


Figure 3.3: In panels, D-F, a segmented profile associated with samples 1,2, and 3 respectively, is shown. These (binary) segmentations correspond to the pink, dark green and light blue regions of the associated hyperspectral visualizations in figure 3.2. This segmentation correlates with endocrine tissue, more precisely the pancreatic islets, as indicated in the H&E stainings (panels A-C). The rankings presented below each segmentation, are obtained by combining the spectral and spatial ranking per tissue by giving a higher priority to those m/z-values that are i) prioritized both in a spatial and spectral manner within the same tissue and ii) observed in the same profile across multiple tissues. The close similarity of the prioritized m/z-values shows the value of this method to obtain intra-and inter-sample knowledge regarding the molecular composition supported by the visualizations obtained using UMAP.

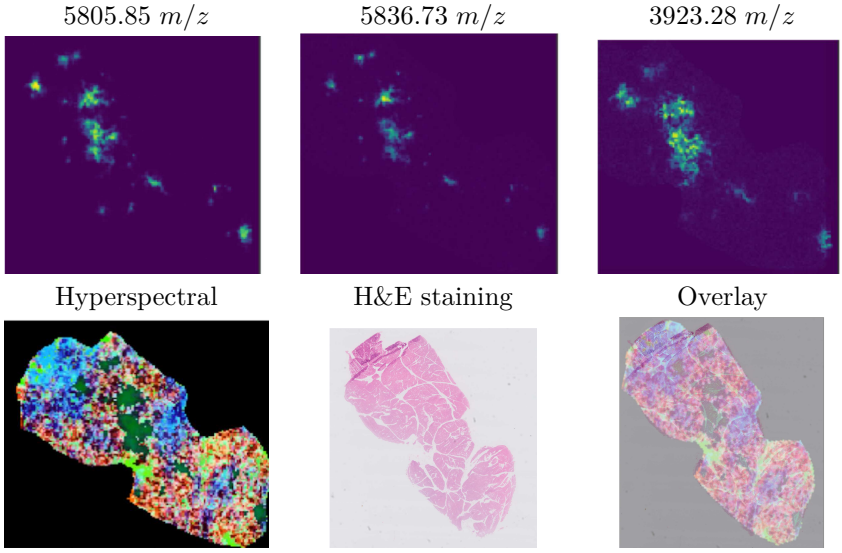


Figure 3.4: The m/z-value of 5805.85 Da, shown in panel A, likely corresponds to the full-length active insulin molecule after cleavage of the c-peptide from pro-insulin. For m/z 5821.29 Da, a putative association could be suggested to ATP5E (ATP synthase subunit epsilon), which plays a role in the oxidative phosphorylation process and VIP (Vasoactive Intestinal Peptide), a known player regarding insulin metabolism, is suggested to be associated with the m/z-value of 5836.73 Da. All prioritized ion images show a clear association with endocrine tissue and more specifically a distribution around the pancreatic islets (figure 3.2).

in finding biological meaning, in particular when dealing with heterogeneous tissue samples.

To the best of our knowledge it is the first time this approach is applied to the results obtained using UMAP. This approach combines the strong visualization and dimensionality reduction performance of UMAP with identifying m/z-values for a specific spatial region. In earlier work, the limitation of PCA in this regard has been highlighted. In particular it was noted that ion images found using PCA differ sometimes from the corresponding score images of the first principal components. At the same time, the added value of spatial segmentation maps with regard to clustering results to identify co-localized m/z- values was noted [6, 39]. Pancreatic tissue samples have the advantage that various endogenous peptides with key metabolic functions are within the mass range of MALDI-MS imaging experiments. For insulin, the proposed

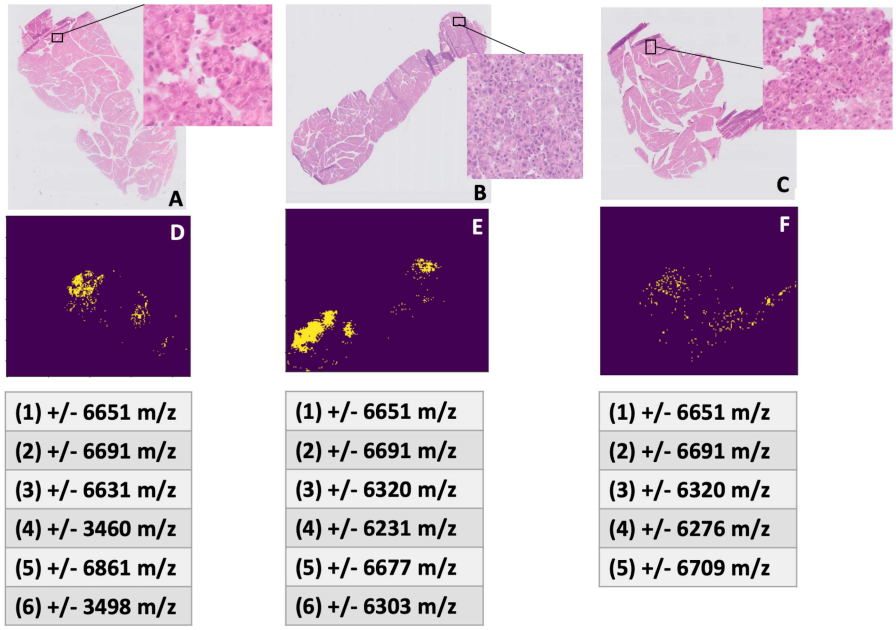


Figure 3.5: In panels D-F, a segmented profile associated with samples 1,2, and 3 respectively is shown. This profile corresponds to exocrine tissue as indicated in the H&E stainings. The rankings are obtained by combining the spectral and spatial ranking per tissue by giving a higher priority to those m/z-values that are i) prioritized both in a spatial and spectral manner within the same tissue and ii) observed in the same profile across multiple tissues. The close similarity of the prioritized m/z-values shows the value of this method to obtain intra- and inter-sample knowledge regarding the molecular composition supported by the visualizations obtained using UMAP. The particular ion images for sample 1 are shown in figure 3.5 and are available for all samples in the supplementary information (B.1-B.12).

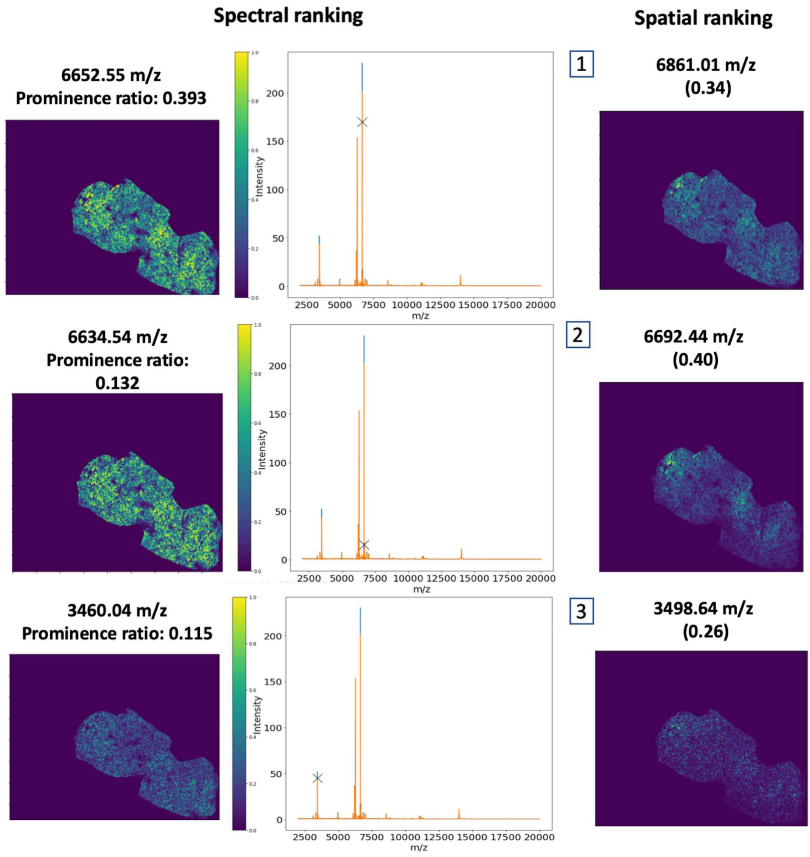


Figure 3.6: Top three of spectral (left) and spatial (right) results ranked vertically from top to bottom. For the spectral results the particular ion image corresponding to the highlighted peak is shown together with the median spectra intensities associated with the extracted profile (in blue) against the median overall spectra intensities (orange). For the spatial results the particular ion images are shown together with their correlation values for the highest ranked ion image per cluster which were also ranked according to their highest average correlation to the extracted profile. A clear association is visible to the extracted pattern which is assumed to be exocrine tissue. The m/z- value around 6650 Da has previously been associated with the acini.

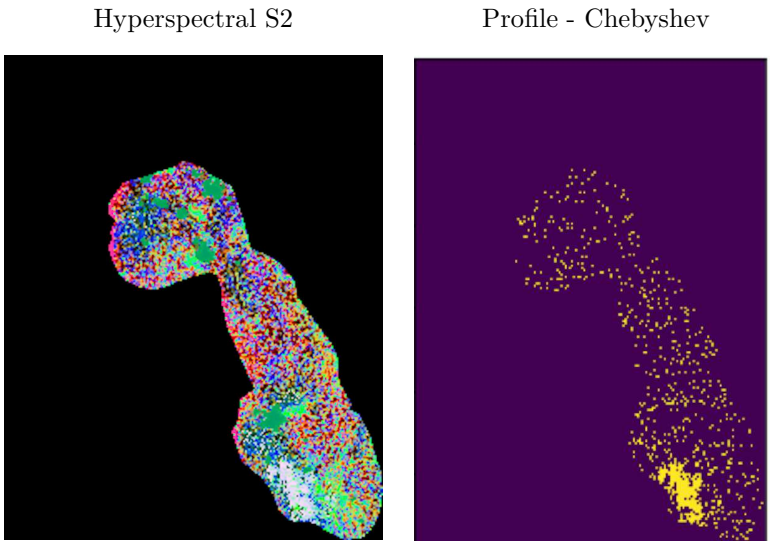


Figure 3.7: On the left the hyperspectral visualization for sample 2 obtained from a healthy mouse pancreas is shown with on the right the extracted profile corresponding to the grey region. This hyperspectral visualization was obtained by using the Chebyshev distance metric instead of the cosine distance metric.

identification is considered realistic because of its high abundance and the striking match with the localization of the pancreatic islets. This is a fine example underscoring the importance to consider both spatial information and spectral information in mass spectrometry imaging. The identity of the other co-localized m/z-values is highly speculative and more advanced identification tools such as immunohistochemistry are essential to further strengthen the proposed identifications, but the fact that we can obtain a robust ranking across tissues shows the value of our method for explorative analysis. Furthermore, in our earlier work, we have shown the enrichment of UMAP in comparison to methods such as for example PCA because it can take the non-linear nature of biological phenomena into account. Moreover, methods such as UMAP are able to capture the complete feature space into two or three dimensions, an attribute that other methods such as PCA and clustering methods lack and which facilitates strong data visualizations [115]. Interestingly, UMAP already alleviated a big limitation inherent to non-linear dimensionality reduction methods being their computational speed. With improvements at this level now being available through GPU support for UMAP in the NVIDIA RAPIDS implementation, we expect the application of UMAP in the analysis of large datasets such as MSI data to take even more importance [93].

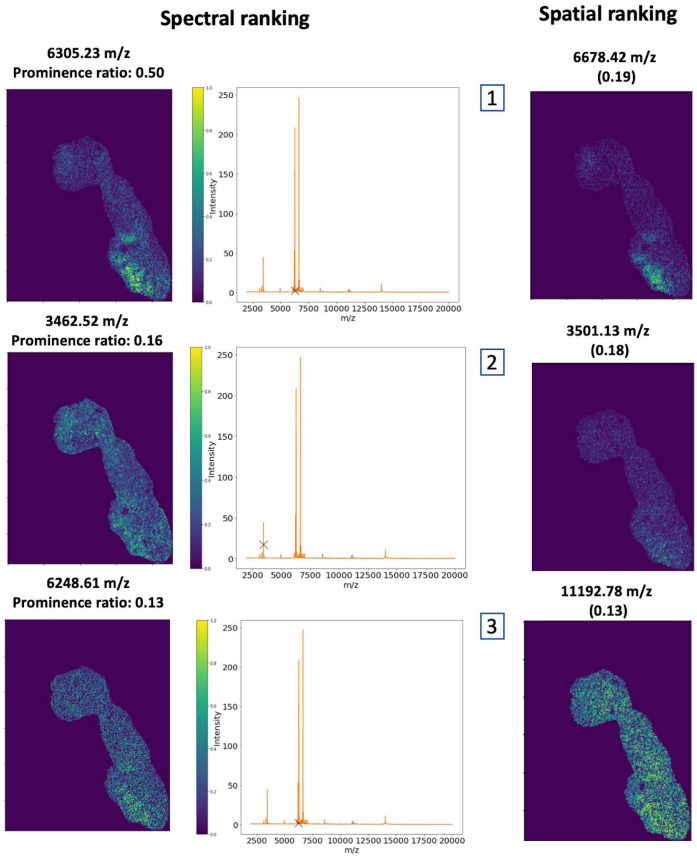


Figure 3.8: Top three of spectral (left) and spatial (right) results ranked vertically from top to bottom. For the spectral results the particular ion image corresponding to the highlighted peak is shown together with the median spectra intensities associated with the extracted profile (in blue) against the median overall spectra intensities (orange). For the spatial results the particular ion images are shown together with their correlation values for the highest ranked ion image per cluster which were also ranked according to their highest average correlation to the extracted profile. In comparison to the results obtained using the cosine distance metric, there is an additional grey region present for which we extracted the corresponding profile shown on the left. We are not able to detect this region in the spectral and spatial results. We can therefore assume that the observation of this region is probably a consequence of the distance metric rather than a reflection of the underlying data.

3.1.7 Conclusion

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

Chapter 4

Data fusion of Mass Spectrometry Imaging and histology data

The method presented in this chapter is filed for a patent (EP20189598.4). This chapter is also published and reprinted with permission from:

Smets, Tina; De Keyser, Tom; Tousseyn, Thomas; Waelkens, Etienne; De Moor, Bart, "Correspondence-aware manifold learning for microscopic and spatial omics imaging: a novel data fusion method bringing MSI to a cellular resolution", *Analytical chemistry*, 93, 7, 3452-3460,2021,
Copyright 2021, American Chemical Society.

This chapter proposes a correspondence-aware manifold learning approach for data fusion of MSI data with the microscopy image of the same tissue slide. Fusing MSI data (high molecular resolution, low spatial resolution) with the associated microscopy image (low molecular resolution, high spatial resolution) brings MSI to a higher spatial resolution, supporting the pathologist in interpreting the obtained molecular information.

High-dimensional molecular measurements are transforming the field of pathology into a data-driven discipline. While H&E stainings are still the gold standard to diagnose disease, the integration of microscopic and molecular information is becoming crucial to advance our understanding of tissue heterogeneity. To this end, we propose a data fusion method that integrates spatial omics and microscopic data obtained from the same tissue slide. Through correspondence-aware manifold learning, we can visualise the biological trends observed in the high-dimensional omics data at microscopic resolution. While data fusion enables the detection of elements that would not be detected taking into account the separate data modalities individually, out-of-sample prediction makes it possible to predict molecular trends outside of the measured tissue area. The proposed dimensionality reduction-based data fusion paradigm will therefore be helpful in deciphering molecular heterogeneity by bringing molecular measurements such as MSI to the cellular resolution.

4.1 Correspondence-aware manifold learning for microscopic and spatial omics imaging: a novel data fusion method bringing Mass Spectrometry Imaging to a cellular resolution

4.1.1 Introduction

Pathologists have been relying on morphology-based methods for decades to study and diagnose disease. While such staining approaches enable the assessment of one or two markers in a single tissue slide, spatial transcriptomic and proteomic studies make it possible to evaluate many thousands of molecules simultaneously. The number of studies gathering high-dimensional omics measurements keeps growing in an effort to understand the complex interactions taking place in biological systems. These studies have moved from focusing on single components (e.g. gene) to encompassing the entire genome, and even evaluating complementary omics measurements in parallel (e.g. transcriptomics, proteomics,..) [13, 80].

Increasingly, these components are being evaluated in terms of their spatial organisation as well. A prominent example is the field of Mass Spectrometry Imaging (MSI), which is capable to detect thousands of endogenous (small metabolites, lipids, peptides and proteins) or exogenous (drugs and drug metabolites) species in their spatial context [23]. Another important example is the spatial transcriptomics field which spatially resolves the distribution of gene expression profiles to improve our molecular understanding of tissues [25].

Often, alongside these molecular measurements other imaging modalities, such as high-resolution microscopy images, are being collected as well. Different modalities obtained from the same sample can provide relevant complementary information that cannot be obtained from a single modality [112]. Typically, histological or microscopic images (eg. haematoxylin and eosin stainings) are overlaid with the molecular profiles obtained by MSI. These microscopy images give insight into the correlation between structure and (pathological) function of cells and tissues that is complementary to the molecular information obtained using molecular imaging. Creating these overlays however is challenging because it is difficult to register two images obtained at different spatial resolutions. It is therefore important to overcome these challenges and truly integrate these heterogeneous data sources, a concept referred to as data fusion. By generating a single view or image from a set of source images, we can get a more complete picture of the complex interdependencies present in biological phenomena. In this paper we present a novel data fusion method called

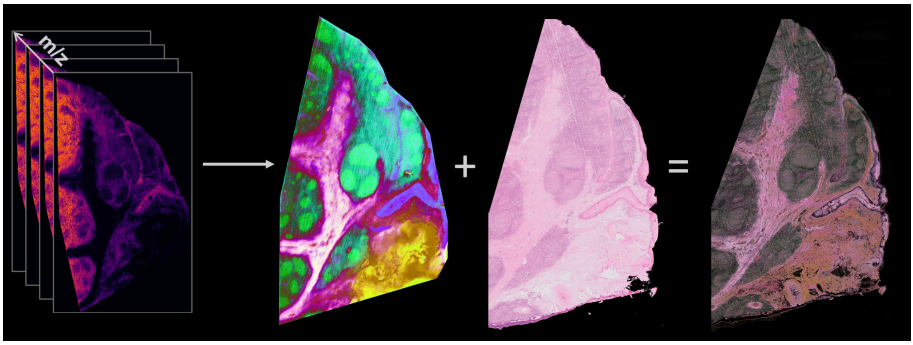


Figure 4.1: Conceptual overview. Correspondence-aware manifold learning is able to fuse the complete molecular feature space of high-dimensional omics data with their corresponding microscopy image. An example is shown for MSI data where the dataset is first reduced to three dimensions and this representation is fused with the microscopy data such that all molecular trends are visualized at a much higher resolution.

'correspondence-aware manifold learning' that builds on recent developments in the dimensionality reduction field.

Non-linear dimensionality reduction methods such as t-distributed Stochastic Neighbor Embedding (t-SNE) [77] are often used for the visualisation of high-dimensional biological data [66]. Not only are these methods capable of detecting non-linear trends, they can also capture the complete feature space when reducing data to two or three components, which is not always the case for methods such as for example Principal Component Analysis (PCA) [46]. Recently, Uniform Manifold Approximation and Projection (UMAP) [85] was introduced to this family of methods with major improvements in terms of scalability, enabling the analysis of large spatial omics data such as MSI.

In earlier work, we have shown how the hyperspectral visualisations obtained using UMAP can reflect the molecular trends present in an entire tissue sample [115]. Connecting these molecular trends to histological information is essential to support clinical settings. These images, obtained from the same subject or tissue sample but acquired in different ways, are expected to show some level of correspondence. The same anatomical structures can be displayed in the microscopic image but can also be reflected by the molecular trends. Where the microscopic information will typically have a lower 'chemical resolution' but a very high spatial resolution, the molecular trends are complementary in this regard, as they offer very rich chemical information but at a lower spatial resolution. With correspondence-aware manifold learning we are now able to

visualise these molecular trends at a higher spatial resolution by fusing the molecular and microscopic data (Figure 4.1). Moreover, using out-of-sample prediction we can predict the distribution of these molecules for regions not measured by the molecular measurements such that we can enrich a complete microscopy slide with the biological signals available. We demonstrate our approach for the fusion of representative spatial omics and optical microscopy data.

4.2 Methods

4.2.1 UMAP outline

UMAP creates a topological structure that represents the high-dimensional data by assembling approximations of local manifolds, and assembles an equivalent topological structure for a low-dimensional representation of the data. It then optimizes the low-dimensional representation to the high-dimensional data by minimizing the cross entropy between the two topological structures [85]. The algorithm innovates by its mathematical foundations that allows it to make interesting assumptions on the data. An important assumption often used in manifold approximation is a uniform distribution of the data on the manifold [16]. For real world data this is usually not the case. UMAP addresses this problem by creating local Riemannian manifold approximations on which the data is assumed to be uniformly distributed and patching them together into a fuzzy simplicial set representation of the data.

UMAP uses fuzzy set cross entropy to compare the two fuzzy simplicial set representations, (A, v) for the high-dimensional data and (A, w) for the low-dimensional data. A low-dimensional embedding can be optimized using the cross entropy loss with the membership strength functions v and w as catalysts for attraction and repulsion:

$$C_{UMAP}((A, v), (A, w)) = \sum_{a \in A} v(a) \log\left(\frac{v(a)}{w(a)}\right) + (1 - v(a)) \log\left(\frac{1 - v(a)}{1 - w(a)}\right). \tag{4.1}$$

In general UMAP fits well within the family of algorithms such as t-SNE [77] or LargeVis [91]. These algorithms rely on different mathematical principles although their implementations have lots of common ground. Like t-SNE and LargeVis, manifold approximations are implemented as weighted k-neighbour graphs. It is explained in detail in [85] that the main equations from these

algorithms also share similarities. Any of these algorithms would suit the data fusion method explained here.

In earlier work we have shown the strong visualisation capabilities of UMAP for MSI data, making it an excellent choice as a general purpose algorithm for high-dimensional omics data [115, 83]. UMAP is therefore used and extended to fit the desired data fusion goals. We use UMAP as a dimensionality reduction algorithm for high-dimensional omics data and adapt UMAP to fuse the resulting low-dimensional representation with high-resolution imaging.

4.2.2 Capturing correspondence

Our goal is to capture spatial correspondence between high-resolution and high-dimensional data and leverage this information into the manifold learning process. Let us use the matrix $A_{n\times p}$ to denote the flattened high-resolution data and $B_{m\times q}$ for the low-resolution spectral data. The correspondence between these two data sets is recorded in the matrix $C_{n\times m}$ such that

$$C_{ij} = \begin{cases} 1 & \text{if } A_i \text{ corresponds to } B_j \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

Finding matching pairs can be achieved with registration techniques. Geometric transformation algorithms can estimate a projection between the two coordinate spaces using only a small set of matching pixel pairs. The output is a transformation matrix that can be applied to warp all other pixels from one image to the other. In this work we relied on the Python Scikit-image library for image processing to perform registration and obtain a transformation matrix through affine transformation. More specifically we identified only four landmark points in both images to estimate the transformation matrix. We have observed that CAML is robust against small registration deviations.

4.2.3 Correspondence-aware manifold learning (CAML)

Correspondence-aware manifold learning (CAML) creates a fused representation of two datasets such that its corresponding instances lie close to each other in the fused representation. Specifically, CAML models the fusion task as an optimisation problem that aims to (1) preserve local distances within the first dataset, and to (2) minimize distances of corresponding instances with the second dataset.

Consider the task of fusing high-resolution data matrix $A_{n\times p}$ and high-dimensional data matrix $B_{m\times q}$ based on a correspondence matrix $C_{n\times m}$. The information of the correspondence matrix $C_{n\times m}$ is reconstructed as a mapping γ between the index sets of both datasets.

Definition 1. *Define $\gamma : I \to 2^J$, the correspondence map for matrices $A_{n\times p}$ and $B_{m\times q}$ and their respective index sets $I = \{1, 2, \ldots, n\}$ and $J = \{1, 2, \ldots, m\}$, such that*

$$\forall i \in I : \gamma(i) = \{\ j \mid j \in J \text{ and } A_i \text{ corresponds to } B_j\}$$

As a consequence, $\gamma(i) = \phi$ when there are no corresponding instances for A_i in B.

We capture the concept of distance between corresponding points as an interplay of attraction and repulsion. CAML aims to minimize the repulsion between corresponding points. We formally define the repulsion between A and B:

Definition 2. *Consider matrices $A_{n\times d}$ and $B_{m\times d}$. Let γ be the correspondence map between the index sets of A and B. Define $\Psi : \mathbb{R}^d \to \mathbb{R}$, the repulsive strength between these matrices, as*

$$\Psi_i(A) = \frac{1}{|\gamma(i)| + \epsilon} \sum_{j\in\gamma(i)} \|A_i - B_j\|_2.$$

The repulsion Ψ is an average because a single instance in $A_{n\times p}$ can have multiple corresponding instances in $B_{m\times q}$ when $n < m$. All corresponding instances receive equal weight. A small number ϵ is used to avoid division by zero when calculating the average.

Definition 2 defines repulsion by comparing instances of the two datasets, which is done by choosing a shared dimension d for both datasets. In our case, $d = p = 3$, similar to RGB color space. The second dataset can then be reduced to d dimensions using any dimensionality reduction algorithm, resulting into the two datasets two datasets $A_{n\times d}$ and $B_{m\times d}$.

Given two datasets $A_{n\times d}$ and $B_{m\times d}$ and their correspondence map γ, CAML can be formulated as a constrained optimization problem for a manifold learning cost function C_{MAN}.

$$\min_A \quad C_{MAN}(A) \quad \text{subject to} \quad \Psi_i(A) = 0, \quad i \in \{1, 2, ..., n\}, \tag{4.3}$$

Instead of solving the constrained optimization problem (4.3), we consider here the problem (4.4) with soft constraints instead of hard constraints [1].

$$Q(A;\mu) = C_{MAN}(A) + \frac{\mu}{2}\sum_{i=1}^{n}\Psi_i^2(A). \tag{4.4}$$

The first term in Equation 4.4 focuses on preserving local distances within the data. In our case, $C_{MAN} = C_{UMAP}$ (see Equation 4.1). The second term penalises the repulsion between the corresponding instances in the two datasets. Because the penalty term in $Q(A;\mu)$ is smooth, we can use unconstrained optimisation methods to find a solution.

The penalty parameter μ controls the balance between the two terms. A high value for μ increases the importance of the correspondence information while a low value for μ focuses on the manifold projection of A. Figure S1 provides an overview of running the algorithm using different values for μ. The optimal value for μ can also be learned using iterative methods, however optimizing $Q(A;\mu)$ is costly depending on the underlying manifold learning algorithm. Better means of evaluating low-dimensional embeddings have been published instead of calculating $Q(A;\mu)$ directly. We used the following as an alternative:

$$E(A;\mu) = 1 - \left((1-T)^2 + \sum_{i=1}^{n}\Psi^2(A_i)\right) \tag{4.5}$$

We want to capture both the quality of the manifold embedding $A_{n\times d}$ and the similarity between the low-resolution image $B_{m\times d}$. The first concept can be tackled using the measure of trustworthiness [130]. Trustworthiness evaluates to what extent the local structure within data is retained in a manifold embedding. For the second concept, we already defined repulsion to calculate similarity between corresponding points of images. Equation (4.5) thus expresses the trade-off between the trustworthiness T of the embedding and the repulsion Ψ of corresponding points in a value between 0 and 1. A represents the fused result after solving the optimization problem. A high value for $E(A;\mu)$ corresponds to a high trustworthiness and a low repulsion. Figure S1 presents the values for $E(A;\mu)$ for each of the embeddings.

Currently CAML has been implemented as an extension of the UMAP algorithm because of its general applicability. The implementation is based on the model implementation of UMAP by the original author. Although CAML

[1] This sentence has been adapted and was originally published as 'This equality-constrained problem can be transformed into the following quadratic penalty function, which concludes the CAML cost function:'

expands the algorithm, to our knowledge the implementation does not impose additional theoretical complexities on UMAP and does not remove performance improvement made to the algorithm. The UMAP algorithm also makes it possible to embed data based on an existing embedding. Leveraging this data transformation option together with image slicing releases the computational burden imposed by very large images.

All images are converted to the LAB color space prior to fusing. The LAB color space is a perceptually uniform color space that is often considered for microscopy imaging analysis, and it fits well with the current cost function. We have seen that this workflow provides good results and limits the computational burden on the system.

Only the dimensionality reduction step for the lymphoma MSI dataset has been done on an Intel Xeon CPU E5-2660 v2 2.20 GHz machine with 10 cores and 128 GB RAM. All other experiments have been done on a MacBook Pro with a 2,8 GHz Intel Core i7 CPU and 16 GB RAM.

4.2.4 Data measurements

For the human-lymph-node sample, cryosections of 5 μm thickness were prepared and mounted on ITO glass slides. 2,5-Dihydroxybenzoic acid (2,5-DHB) was used as the matrix and applied using sublimation. The pixel size was set to 10 μm, and the recorded m/z range was 620-1200 Da in positive reflector mode. The acquisition was performed with 200 lasershots/pixel and a laser repetition rate of 10 kHz, resulting in an acquisition speed of 32 pixels/s. For the mouse brain spatial transcriptomics samples, H&E images and count matrices were downloaded from https://www.spatialresearch.org/resources-published-datasets/ licensed under the Creative Commons Attribution license. The dimensions of the datasets are as follows: for the lymphoma dataset 500 000 pixels x 8000 m/z values and for the spatial transcriptomics data 281 pixels x 16 416 transcripts.

4.3 Results

With correspondence-aware manifold learning we:

i. project the high-dimensional molecular features to a low dimensional space,

ii. capture spatial correspondence between the high-resolution microscopic image and the high-dimensional molecular measurements of the same tissue sample,

iii. perform correspondence-aware manifold projection using both data modalities to obtain a fused image reflecting both modalities in one visualisation.

The general methodology is depicted in Figure 4.2. The high-dimensional molecular data is first reduced to three dimensions. After a registration step, this hyperspectral visualisation is used as a constraint to transform the microscopy image, resulting in a fusion of the molecular information with the microscopy data. By modeling the manifold, and piece-wise transforming the full-resolution microscopy image while taking into account the properties of the molecular data, we are able to visualise the molecular image at a much higher resolution. As such we leverage the complementarity of the high spatial resolution offered by optical microscopy with the high-dimensional but lower spatial resolution molecular imaging data.

We demonstrate the correspondence-aware manifold learning approach for data fusion of molecular measurements and their corresponding microscopy images. We show:

i. the prediction of molecular trends at a higher spatial resolution through data fusion,

ii. the prediction of molecular trends outside of the tissue area measured with out-of-sample prediction,

iii. the general applicability of the method.

4.3.1 Correspondence-aware manifold learning for data fusion of molecular and microscopy images

In Figure 4.3, we show the low dimensional representation of reactive lymphoid tissue in a human tonsil MSI dataset (500.000 pixels $\times$ 8000 m/z features, measured at $10\mu m$ resolution). This hyperspectral visualisation represents the complete 8000 m/z feature space compressed into 3 dimensions, such that each color is connected to a molecular trend present in the data. This molecular, hyperspectral visualisation is subsequently used to perform data fusion with the corresponding microscopy image. In Figure 4.4, the fused results show that the molecular trends in the data can be visualised at a much higher resolution.

The goal of performing data fusion is to exploit the complementarity present between the different data modalities such that the resulting visualisation goes beyond the information offered by a single modality. In figure 4.5, a fused detail is shown of the extracellular matrix and stroma comprising glycoproteins and proteoglycans with below the multi-layered squamous epithelium (panel A). Panel B shows a detailed view of a cluster of plasma cells. The nuclei of these cells received a black colour after data fusion, while the green colour seems to correspond to the cytoplasm. In active plasma cells a high density of the golgi apparatus is required for the synthesis of immunoglobulins, which explains the larger amount of cytoplasm present. A closer look reveals that these plasma cells are infiltrating the epithelium (panel C) and display a kind of integration with the present epithelial cells (purple colour) such that they do not overlay or compress these keratinocytes. Panel D shows that the keratinocyte cells further away from the basement membrane have become larger in comparison to the ones closer to this basal layer, which can be explained by the progressive maturation process taking place inside of the squamous epithelium. This maturation process is associated with changes in the composition of the cytoplasm with mainly an increase in the number of cytokeratines, which are part of the cytoskeleton. These findings are supported by the corresponding H&E stainings in panels B'-D', and show the power of data fusion to surpass the regional or subregional level of interpretation offered via MSI by bringing the molecular information to the cellular level.

In the Supplementary figures additional examples highlight the potential of data fusion. In Supplementary figure C.3, two blood vessels are shown where we can see that a venule on the left is surrounded by a thin layer of endothelial cells, while the arteriole on the right is lined by a layer of smooth muscle cells. In light pink we can also perceive some collagen fibers. Figure C.2 demonstrates a secondary B-cell follicle where the reactive germinal center is surrounded by a lymphocyte corona, highlighted by an orange and green dashed line respectively. Supplementary figure C.4 shows how data fusion can support us in distinguishing artefacts from true biological signals. The epithelium contains a small, rounded structure where the number of cells is increased. While this structure is visible in the H&E image, it becomes more pronounced upon data fusion. It is regarded as an artefact created during the cutting process.

All findings are supported by the corresponding H&E images. Moreover, these results highlight the potential of our method to distinguish individual cells in their micro-environment instead of being limited to the interpretation of regional or subregional molecular trends measured by MSI. This could be very valuable when evaluating for instance the invasiveness of individual tumor cells and their interaction with the tumor micro-environment.

4.3.2 Out-of-sample prediction

In addition to performing data fusion for the area covered by the MSI measurements, we can predict the molecular distributions for the entire microscopy image through out-of-sample prediction, as shown in Figure 4.6. This approach enables the interpretation of a much larger microscopic area based on a limited amount of molecular information. This can be a valuable asset given the high costs associated with molecular measurements or the limited amount of tissue that is often available.

4.3.3 Data fusion for spatial transcriptomics data

We illustrate the general applicability of the method for spatial omics data. In supplementary figure C.5, we show the hyperspectral visualisation of the low dimensional representation obtained for a spatial transcriptomics mouse brain measurement and the corresponding microscopy image with the fused result on the right. We show these results to highlight the potential of this method for other spatial omics technologies. Notwithstanding the low spatial resolution of the molecular measurement (281 pixels $\times$ 16.416 features), we can see that the green coloured cell nuclei are embedded in a purple background of the tissue center. Given the low spatial resolution, we want to emphasize the restricted potential for biological interpretation. However given the fast technical improvements that are being made in terms of spatial resolution, we believe the proposed method holds a lot of potential for data fusion of spatial omics measurements. In this light, in supplementary figure C.6, we can also observe that the highlighted regions show a clear correspondence of the green dots with the cell nuclei, as stained by hematoxylin, across the associated H&E image.

4.4 Discussion

Scalable and powerful dimensionality reduction methods have become indispensable to deal with the growing number of high-dimensional datasets. Non-linear dimensionality reduction methods, such as t-SNE and UMAP, have brought and continue to bring significant value for the biomedical sciences in this regard [78, 40]. Due to their strong visualisation capabilities these methods have become a standard for the analysis of high-dimensional datasets [15]. And while the number of high-dimensional and spatial omics measurements keeps on growing, the computational methods capable of fusing the multi-modal

measurements acquired from the same sample are lagging behind. In this work we present a novel data fusion method that is able to compress and fuse the complete molecular and microscopic feature spaces toward a combined image. This work builds on the framework of non-linear dimensionality reduction methods to enable the fusion of molecular and microscopic data obtained from the same tissue sample. We demonstrate our results according to the UMAP framework, but the same principle could be applied starting from similar methods such as for example t-SNE. By constraining the projection step based on the molecular target information, we are able to transform the corresponding pixels in the microscopy image accordingly, resulting in a fused representation presenting the molecular information at a much higher spatial resolution.

In Figure 4.4, we highlight the potential of our method for the integration of MSI data with the corresponding microscopy images. While we show that this enables us to improve the resolution of the MSI data, these colours reflect in fact an underlying group of biomolecules. As such, in previous work, we have shown that it is possible to prioritize and identify those molecules associated with a molecular trend or color [116]. This could support researchers in finding correlations between underlying biological actors and their histopathological architecture. Recent work has shown that it is possible to correlate single cell morphological features based on microscopic images with molecular information [110]. Given that the proposed method is capable of retaining the cellular morphology in the fused results, we believe it holds potential for this area of study as well. Moreover, in figure 4.5 we show that the proposed method can leverage MSI measurements to study tissues at the cellular level such that we can move beyond the regional or subregional insights and evaluate the presence of aberrant cells in their micro-environment. This will be of growing importance with technological advancements in terms of spatial resolution but also with the increasing demand to integrate multi-modal data measurements. In this regard, we have also included a spatial transcriptomics sample as an example of a rapidly evolving domain with a lot of potential [118]. While the molecular measurements are at the moment still of a lower resolution, it is yet possible to show that the method is widely applicable and will be able to offer more value with increasing spatial resolution. An additional advantage data fusion has to offer is the ability to better distinguish artefacts from true biological signals (supplementary figure C.4). These advantages will be useful not only in the domain of molecular imaging but in general for example when dealing with other imaging technologies such as for example MRI, CT, PET, etc. ...).

Earlier work in this domain has been focused on the modelling of linear relationships between the different modalities. Van de Plas et al. have focused on the fusion of MALDI MSI data with optical images of H&E stainings

through the mapping between two modalities based on linear regression models [98]. And recently, a novel method was introduced by Race et al. based on dimensionality reduction through non-negative matrix factorization [102]. While providing good results, given the non-linear nature of biological data, taking into account these complex relationships is preferable. This is in particular the case for MALDI MSI data which can suffer from artifacts caused by the non-linear ionization process or for instance by ion suppression. Therefore, the method proposed in this work performs data fusion through a non-linear dimensionality reduction approach. In addition to the benefit of being able to take more complex interactions into account, the data fusion result is not hindered by an imperfect multi-modal registration, which can also be seen in Figure 4.6. Given that registration is often a time-consuming and difficult step, this constitutes an important advantage.

To illustrate the general applicability of the method we have also performed a study on a multi-modal MNIST dataset, an extension of the well known MNIST database of handwritten digits. In Supplementary Figure C.7, we show that we can perform data fusion on single multi-modal digits and we can also perform out-of-sample prediction based on this initial trained data fusion model, with additional experimental verification for the MSI data in figures C.8-C.9.

In conclusion, data fusion facilitates the combination of complimentary data sources to obtain insights that would not be obtained from a single modality alone. Given the growing interest towards spatial multi-omics studies, this method will be valuable to enable the mapping of molecular measurements to the underlying tissue architecture at cellular resolution. Moreover, given the large costs associated with state-of-the-art molecular measurements, the out-of-sample prediction can expand the amount of information obtained from conducted experiments. Finally, due to its broad applicability we hope that the proposed paradigm will be valuable to researchers coming from different domains.

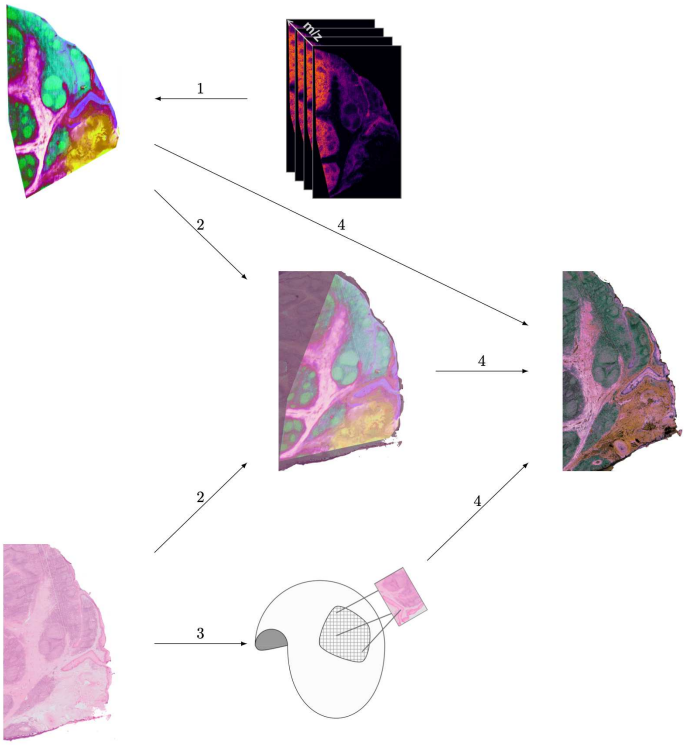


Figure 4.2: Method overview.
(1) The molecular data represented by matrix $B_{m\times q}$ is reduced to a matrix $B_{m\times 3}$ target embedding. (2) The pixel coordinates of the molecular and the microscopy image undergo a registration step such that we obtain a correspondence matrix. (3) Subsequently, the microscopy image is subjected to a dimensionality reduction step, wherein each pixel is evaluated in function of its correspondence to the target embedding. (4) Specifically, the projection step in the dimensionality reduction method is constrained based on the target embedding, causing pixels in the microscopy image to receive a similar color based on the reduced target embedding of the molecular data. This approach enables the transfer of information obtained from a complete high-dimensional molecular dataset to a single microscopy slide, but can also be used to transfer the information from a single feature or molecular image to the microscopy slide.

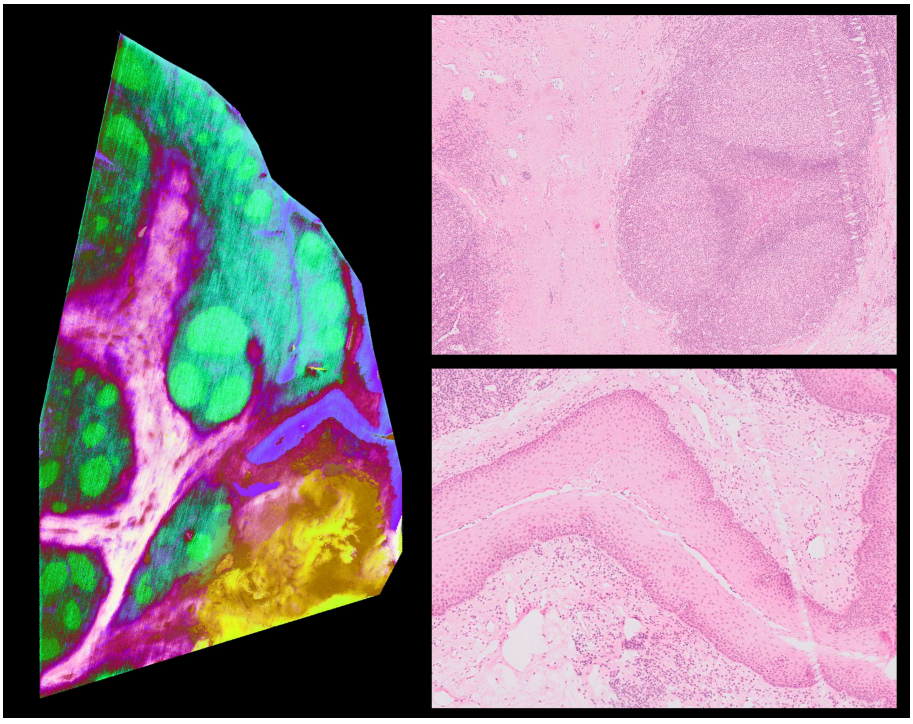


Figure 4.3: Low-dimensional representation of lymphoma MSI dataset and corresponding H&E image. (a) Shown on the left: the low dimensional representation of a human lymphoma MSI dataset (500.000 pixels $\times$ 8000 m/z features, 10μm resolution). The different colours reflect the molecular trends present in the data. On the right, two parts of the corresponding microscopy image are shown.

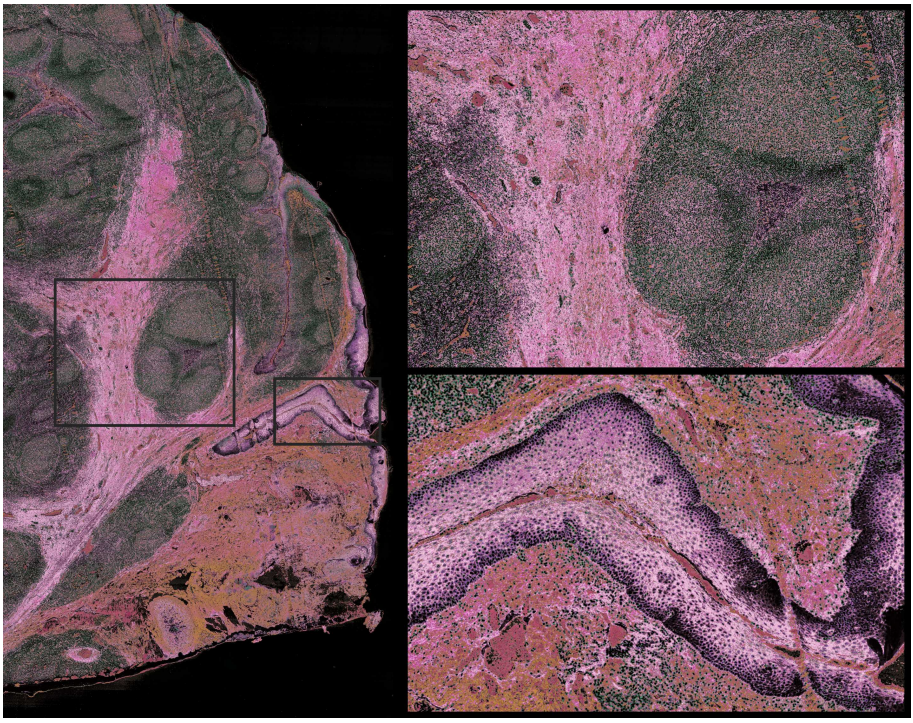


Figure 4.4: Fusion of mucosa-associated lymphoid tissue of the tonsil MSI dataset and corresponding H&E image. On the left the data fusion result for the molecular and microscopy images is shown. On the right the fusion details are shown for two regions.

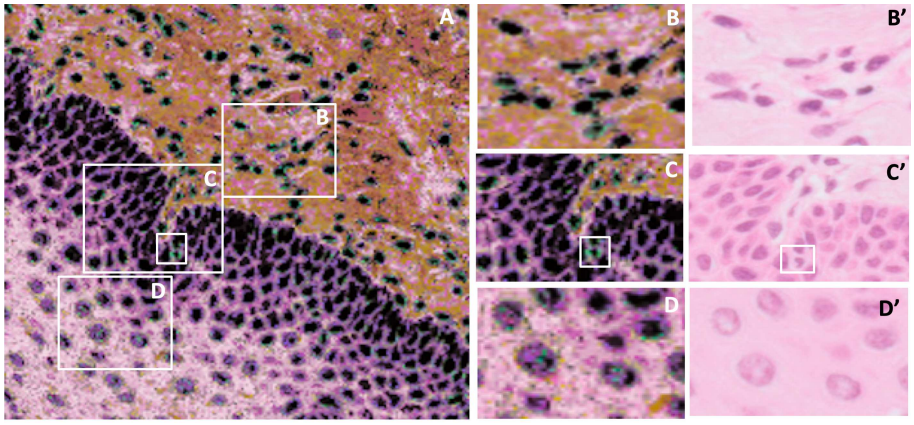
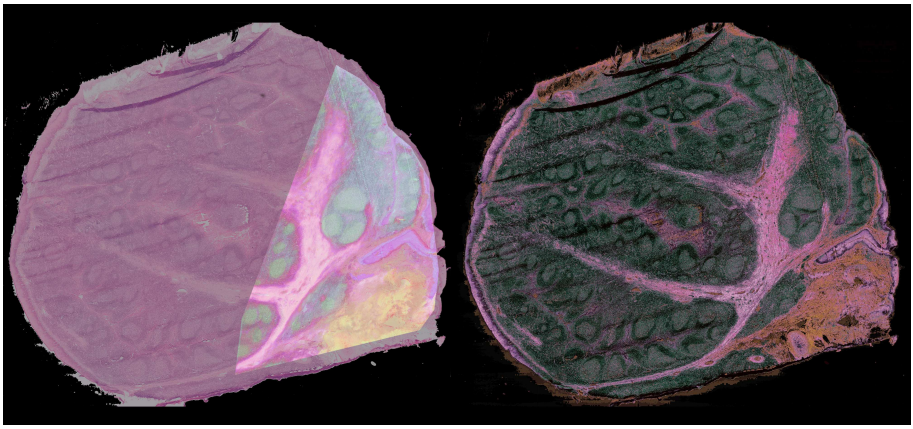


Figure 4.5: A detail of the fused data. Panel A shows the extracellular matrix and stroma comprising glycoproteins and proteoglycans, as well as plasma cells on top and the multi-layered squamous epithelium below. Panel B shows a detailed view of a cluster of plasma cells. The nuclei of these cells received a black colour after data fusion, while the green colour seems to correspond to the abundant cytoplasm. In active plasma cells a high density of the golgi apparatus is required for the synthesis of immunoglobulins, which explains the larger amount of cytoplasm present. A closer look in panel C reveals that these plasma cells are infiltrating the epithelium, as they display a kind of integration with the present keratinocytes (purple colour) such that they do not overlay or compress these cells. Panel D shows that the keratinocyte cells further away from the basement membrane have become larger in comparison to the ones closer to this basal layer, which can be explained by the progressive maturation of these cells. These findings are supported by the corresponding H&E stainings in panels B'-D'. Note that these results are provided at the cellular level: individual cells with their particular nuclei and cytoplasm are shown. This demonstrates the power of data fusion to surpass the regional or subregional level of interpretation offered via MSI by bringing the molecular information to the cellular level.

Figure 4.6: Example of out-of-sample prediction. On the left an overlay is shown of the region in the lymphoma tissue measured by MSI with the microscopy slide. On the right, the result of out-of-sample prediction shows the fused result for the complete microscopy or H&E image.



Chapter 5

Conclusion and future perspectives

5.1 Conclusion

Mass Spectrometry Imaging (MSI) is an emerging and powerful molecular imaging technology enabling a true label-free tracking of hundreds to thousands of biomolecules simultaneously throughout flat samples such as biological tissue sections. Continuous technological improvements are pushing the data size boundaries causing manual exploration to become infeasible, and datamining more complex but also more important than ever. A vital component in the MSI workflow is extracting the rich information captured in these measurements, as evidenced by the growing number of publications regarding the computational analysis of MSI data.

In this thesis we have contributed to the MSI community by focusing on three areas of MSI data analysis, namely non-linear dimensionality reduction of MSI data, the prioritization of m/z-values in individual tissue samples and/or across multiple tissue samples and the fusion of MSI data with microscopy images to enhance the spatial resolution of the molecular imaging data by introducing the concept of correspondence-aware manifold learning.

In chapter 2, we discussed the use of Uniform Manifold Approximation and Projection or UMAP for non-linear dimensionality reduction of MSI data. The application of non-linear approaches for dimensionality reduction has a strong advantage in terms of visualization capabilities but in particular for biological data with its inherent non-linear relationships between data points. We have performed an extensive comparison of UMAP with its older neighbouring method, t-SNE, using spatial autocorrelation as a benchmark. While t-SNE was not scalable for large datasets such as in the MSI field, recent computational advances are even further improving UMAP in terms of computational speed which is a bottleneck of most non-linear models. With the recent introduction of UMAP in the RAPIDS framework provided by NVIDIA, the computational speed for a 30 Gb MSI dataset was even reduced from a 6 hour runtime to a few minutes only thanks to GPU acceleration. Figures 5.1 and 5.2 illustrate the importance of UMAP for large datasets and an overview of the dimensionality associated with the demonstrated datasets in this thesis is given. While currently UMAP has become very popular in the biomedical sciences, we were the first to apply and evaluate the method for MSI data and one of the earliest applications in the biomedical field in general [115, 15]. Moreover, we have shown the importance of the distance metric used to reduce the dimensionality of the data. The choice of a particular distance metric is not always easy and strongly depends on the application and the underlying data, which we have observed in our work as well. Overall, the cosine distance metric resulted in the best visualizations according to the associated microscopy images. Looking for automatic and less empirically driven ways to support the choice of a particular distance metric or even the design of a customized distance metric for dealing with MSI data are relevant topics for further research.

Dimensionality MSI data	**t-SNE runtime**	**UMAP runtime**
10 000 pixels x 14 000 m/z **2 Gb**	~ 20 minutes	~ 5 minutes
500 000 pixels x 8000 m/z **30 Gb**	Days	~ 6h (128 Gb RAM) **~ 5 minutes using GPU** (rapids)

Figure 5.1: Overview of average dimensionality associated with the demonstrated MSI datasets and associated runtimes.

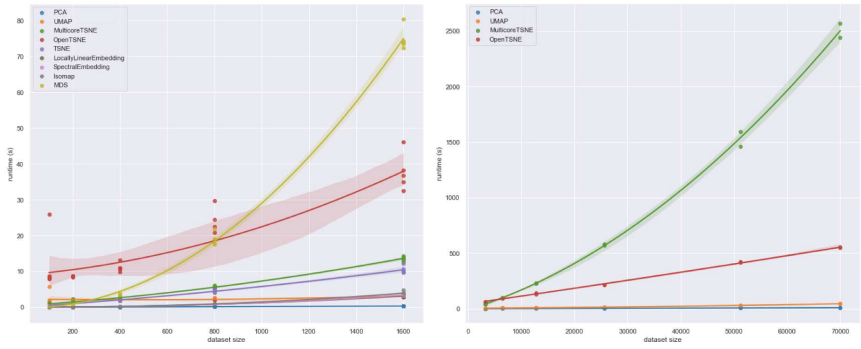


Figure 5.2: Experimental comparison of runtimes associated with a number of dimensionality reduction methods in function of increasing dataset sizes. The graphs illustrate the superiority of UMAP in particular for large datasets.

In chapter 3, we proposed the application of non-linear dimensionality reduction in a bi-directional way to prioritize the m/z-values associated with the hyperspectral visualization of tissue samples. Using this bi-directional approach, we are able to incorporate both the spatial and the spectral information into the prioritization process. We have shown that this enables a robust ranking of molecular features per individual sample or across different samples as demonstrated for a number of mouse pancreas tissue sections. The repeated representation of the glucagon molecule around position 13 in all rankings highlights the robustness of the approach. Given the strong improvements relating to computational speed, this method can provide quick insights into the molecular players driving the visualization of each sample and aims to alleviate the disadvantage of limited interpretability in non-linear dimensionality reduction methods [116].

In chapter 4, we introduced a novel data fusion paradigm called correspondence-aware manifold learning for fusing MSI and microscopy data. We showed that this approach is capable of bringing the molecular data to a higher spatial resolution [114]. It was demonstrated that a single cell infiltrating into the epithelium could be differentiated from the epithelial cells. As such this method

can play an important role to give quick insights, for instance regarding a cancer biopsy, where typically single aberrant cells have to be identified. As far as we are aware, this is the first data fusion method in the MSI field to incorporate non-linearities or to make data fusion possible based on manifold learning. Moreover, this approach can be extended to any (molecular) imaging approach that might benefit from data fusion.

5.2 Future perspectives

The number of publications related to the computational analysis of MSI data has increased steadily over the last years. As shown in the introduction of this thesis the focus has shifted from (pre)-processing of the data towards more downstream and statistical analysis, but also to the support of identification of molecules measured in MSI. Nevertheless, there are still many outstanding computational challenges to be addressed in order for the MSI field to reach its full potential. Here we give an overview of some interesting movements and directions that will require specific attention in the coming years.

5.2.1 Technical improvements leading to novel, more complex or larger datasets

Ion Mobility Separation (IMS) can be integrated into an imaging mass spectrometer between source and analyzer, and as such adds another molecular chemical dimension by separating ions in gas phase based on their structural properties, namely, their collisional cross section (CCS), while flying through a buffer gas [108, 58, 27, 72, 5]. Previous IMS applications in bulk proteomics,metabolomics, and lipidomics, have demonstrated an increase in specificity (separation of structural isomers by their CCS and matching signals to either measured or theoretically predicted CCSs. Moreover, IMS has also shown to increase the sensitivity enabling the analysis of every CCS peak individually thereby avoiding ion suppression which is an important problem in MSI as well [58, 5]. Recently technical developments in this direction were presented by several vendors, therefore we expect a surge in the demand for computational methods to deal with this additional dimension. Machine learning has already been successfully applied in CCS prediction for both metabolites and lipids and will be useful for the MS imaging field as well [139, 140, 55, 5] Besides the challenge of incorporating the information obtained from this additional dimension, it will be important to keep up with the ever-increasing data sizes of MSI data due to improvements of the spatial and spectral resolutions. High-mass resolution mass spectrometers such as FT-ICR and Orbitrap instruments and micron to sub-micron spatial resolutions will drive dataset sizes into the terabyte range. The research presented in this thesis will therefore be valuable for the visualization and interpretation of these large datasets.

5.2.2 The multi-omics and single cell era: integration with other data modalities

While it is known that tumors are composed of an abundant number of cell types, histopathology is still the standard method for grading and evaluating a tissue biopsy. Enrichment with molecular information is in this context still mainly focused on histological approaches such as H&E stainings or immunohistochemistry, enabling the simultaneous assessment of a few markers only. Along with a progressive shift towards molecularly targeted approaches to treat human disease, treatment decisions are increasingly based on the molecular characterization of clinical tissues. Despite major advances in multi-omics approaches, one of the key hurdles in this analysis is the high degree of intra-tissue heterogeneity that in bulk omics approaches often masks important changes and hampers a proper interpretation of the data. These more untargeted approaches in the multi-omics domain (eg. standard bulk RNA sequencing or mass spectrometry (MS)) are very informative in terms of molecular information but suffer from a loss of spatial context due to the need of dissociating the tissue prior to analysis, in addition to bulk averaging of molecular information. Novel technologies such as multiplexed and molecular imaging-based techniques are able to move beyond the detection of a few markers, while retaining spatial information as well. In addition, novel single-cell (multi-)omics sequencing and imaging-based approaches are able to profile cellular heterogeneity and cell-cell cooperation down to the single-cell level within tissues. This is crucial to study the complex interactions taking place between cells in the tumor microenvironment. A single modality is however not sufficient to understand the complex tumor microenvironment and gathering cutting-edge spatially characterized single cell data across these different molecular layers is very costly. Therefore the computational development of dedicated data fusion approaches will become more important to exploit the complementarity of different molecular modalities measured (figure5.3). An example of additional modalities that can be combined is given below with a summary overview in figure 5.4.

Bulk MS-lipidomics Lipidomics involves the analysis of lipids on extracts from tissues or liquid biopsies MS-based methods have the ability to analyze intact complex lipids with the resolving power to analyse individual lipid species. In a typical experiment approximately 2000 species can be identified and quantitated. Although it lacks spatial information, it comes with an unsurpassed resolving power in terms of lipid species and molecular identification. As such it is highly complementary to MSI for molecular species identification and confirmation, which is more difficult to achieve when using MSI by itself.

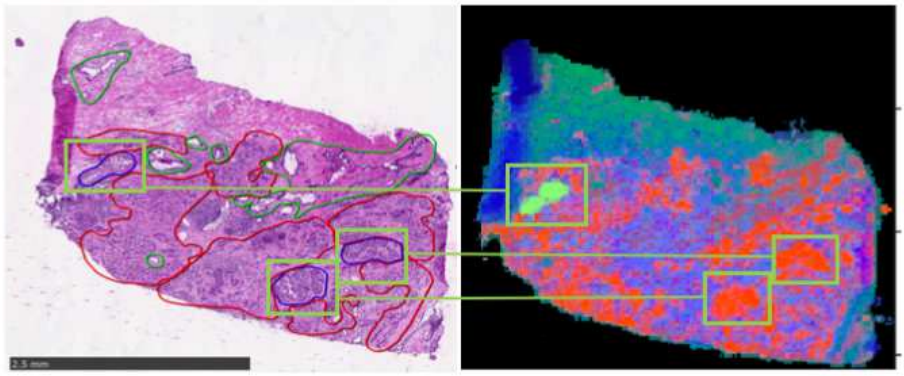


Figure 5.3: On the left an annotated H&E staining is shown, with in blue morphologically similar glands. On the right, spatial lipidomics uncovers distinct molecular profiles for the same tumor specimen across morphologically comparable glands in prostate cancer tumor tissue but requires stronger discriminatory power to unravel the underlying mechanisms.

Spatial transcriptomics (VISIUM) The 10x Genomics Visium spatial technology is an untargeted (20 455 genes) spatial transcriptomics method that identifies in situ captured and barcoded polyadenylated transcripts using next generation sequencing at a spatial resolution of 55 μm diameter.

Spatial proteomics (MILAN) This method relies on antibody staining through repetitive cycles, enabling the spatial proteomic profiling of several dozens of biomarkers at the single cell level. Antibody-based labeling of the targets comes with a high specificity but depends strongly on the available antibodies and is limited to 50-100 markers. Images are collected in each cycle and therefore computational analysis makes it possible to extract single cell features facilitating the interpretation of cell-cell interactions.

Single-cell RNA-plus-ATAC-seq The newly developed 10x Genomics Chromium Single Cell Multiome ATAC + Gene expression technology is a novel multi-omics sequencing-based method that allows obtaining both open chromatin and transcriptome information from the same single cells in high throughput, enabling to link use of gene regulatory elements with gene expression to explore gene regulatory networks driving cell differentiation, development, and disease. Furthermore, using computational tools as NicheNet and CellPhoneDB, cell-cell interaction and communication can be predicted based on expression patterns of ligands and receptors across the single cells,

and then validated using spatial transcriptomics and proteomics techniques.

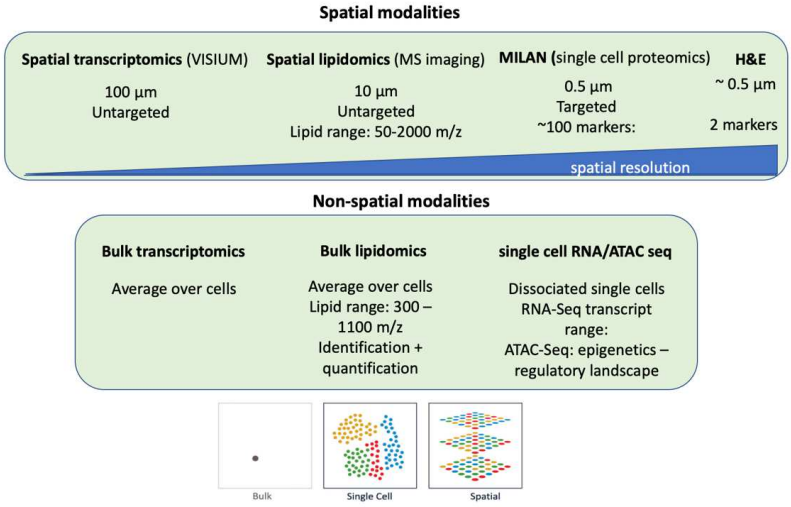


Figure 5.4: Strengths and weaknesses associated with each data modality. Using data fusion we can for example increase the spatial resolution of the molecular imaging methods to the resolution of microscopic H&E images.

5.2.3 Molecular identification

The inability of collecting spatially resolved MS/MS spectra in an untargeted fashion combined with the complexity of spectra due to the lack of separation turns molecular identification into one of the biggest challenges for MSI. Metabolite identification is also referred to as metabolite annotation which is typically taken care of by m/z-matching ie. matching spectral signals or peaks of interest against a molecular database by their m/z-value. This process however leads to very ambiguous results and therefore requires further experimental validation. A first step towards a validated database has been taken by the METASPACE initiative with an approach that uses a novel metabolite/signal match score that exploits both spectral and spatial information from MSI data and matches it to theoretically predicted properties of a metabolite such as its fine isotopic pattern.

However with only 1% of the data being annotated, the majority of the data has no label assigned yet due to our limited understanding of ionization pathways or principles of how an analyte forms a signal in MSI. In bulk metabolomics machine learning has been successfully employed to conduct

automated metabolite identification [38, 89]. Unfortunately, no library of authentic molecular standards, similar to MS/MS spectral libraries in LC-MS/MS metabolomics, is available yet. Once this bottleneck is lifted, research in this area would enable the detection of unannotated molecules and potentially shed a light on ionization pathways to support the integration of signals corresponding to a particular metabolite [5].

Taken together, many upcoming technological advancements and improvements for molecular identification will contribute to the practical utility and popularity of MSI.

In this work, we focused on :

i An enhanced visualization of large MSI datasets with particular attention to uncovering the underlying tissue heterogeneity in an intuitive way

ii The prioritization of the molecular drivers of these visualizations and,

iii Data fusion of the molecular visualization with the associated microscopy image in order to bring the molecular visualization to a higher spatial resolution. Through this data fusion approach it is possible to identify phenotypical differences at the single cell level, which can be an important factor to support the pathologist in the subclassification of tumor samples.

Appendix A

Supporting Information for Evaluation of distance metrics and spatial autocorrelation in Uniform Manifold Approximation and Projection applied to Mass Spectrometry Imaging data

A.1 Overview of distance metrics

A.2 Overview of runtimes associated with t-SNE, Barnes-Hut t-SNE and UMAP

A.3 Pancreas M2: relative comparison of obtained embeddings through spatial autocorrelation

A.4 Pancreas M2: UMAP and Barnes-Hut t-SNE applied with different distance metrics

A.5 Distance metric evaluation on MSI data without TIC normalisation

A.6 Custom distance metric - histomatch

Distance measures	Formula
Euclidean	$\mathrm{d}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
Correlation	$d(x,y) = 1 - r$ $r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - y)^2}}$
Cosine	$d(x,y) = 1 - cosine\ similarity$ $cosine\ similarity = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$
Chebyshev	$d(x,y) = \lim_{p\to\infty}\sum_{i=1}^{n}(\lvert x_i - y_i \rvert^p)^{1/p}$
Histomatch	$d(x,y) = 1 - \sum_{i=1}^{n} max(0, min(x_i - y_i))$

Figure A.1: Overview of different distance metrics used.

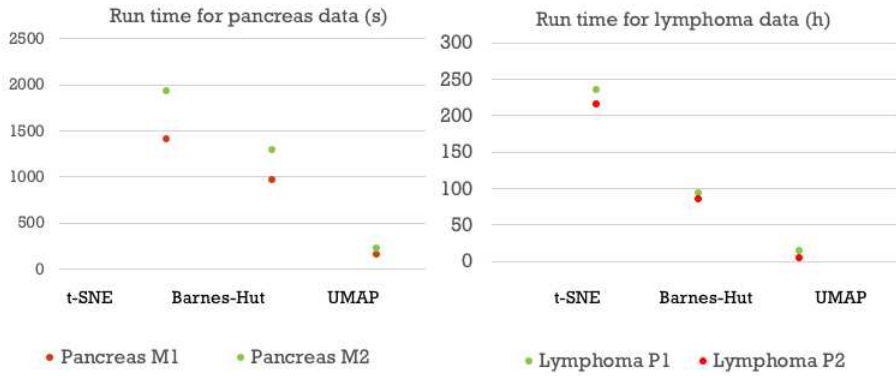


Figure A.2: Overview of runtimes obtained for the pancreas and lymphoma datasets using the t-SNE, Barnes-Hut approximation thereof and the UMAP algorithm. UMAP achieves much faster runtimes than t-SNE, and even the Barnes-Hut approximation thereof. For the pancreas tissues the average result is shown across 5 runs. For the lymphoma tissue only one run is visualised due to the high computational cost associated with the t-SNE algorithm. To make the analyses feasible for these large lymphoma datasets, an initial dimensionality reduction by PCA to 100 dimensions was required to run t-SNE or BH t-SNE.

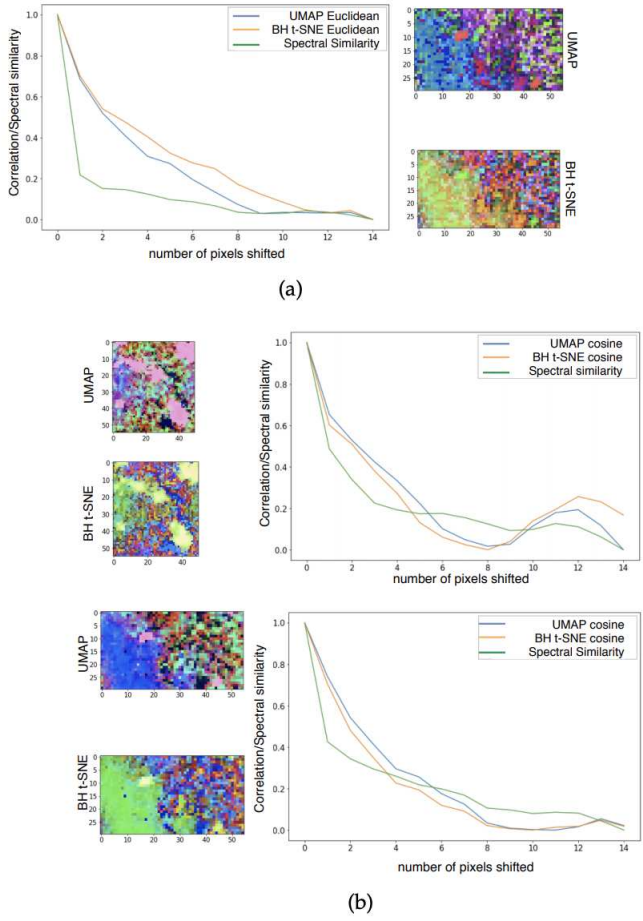


Figure A.3: Spatial autocorrelation of the M2 pancreas embeddings is shown as the correlation in function of increasing distance or incremental number of pixels shifted. As expected the correlation values are maximal for a zero shift and decrease in function of the distance. Shown are the results for two regions in the tissue for the Euclidean (a) and cosine (b) distance metric. The spatial autocorrelation is shown for the embedding obtained on the same tissue section using BH t-SNE (orange) and UMAP (blue). Each graph also includes the spectral similarity (green) according to these metrics as a benchmark indicator for the spatial autocorrelation.

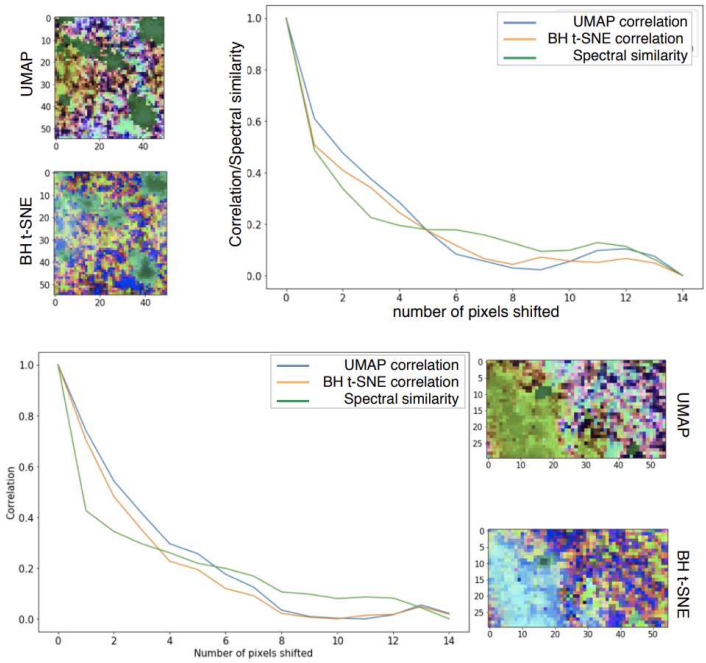


Figure A.4: Spatial autocorrelation of the M2 pancreas embeddings is shown as the correlation in function of increasing distance or incremental number of pixels shifted. As expected the correlation values are maximal for a zero shift and decrease in function of the distance. Shown are the results for two regions in the tissue for the correlation distance metric. The spatial autocorrelation is shown for the embedding obtained on the same tissue section using BH t-SNE (orange) and UMAP (blue). Each graph also includes the spectral similarity (green) according to these metrics as a benchmark indicator for the spatial autocorrelation.

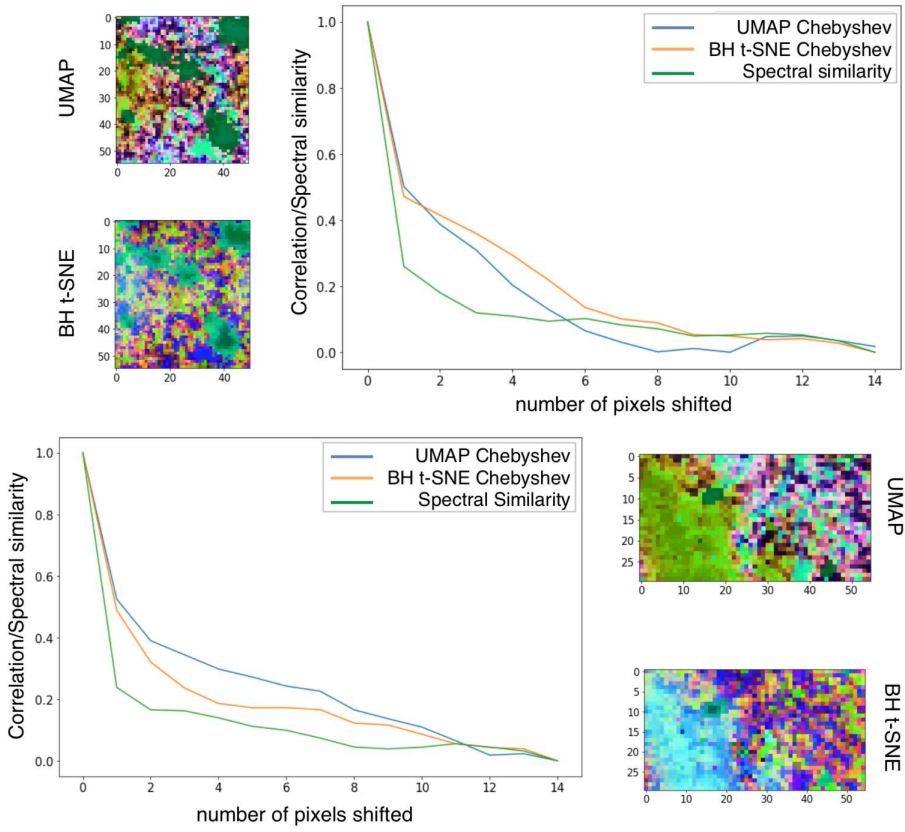


Figure A.5: Spatial autocorrelation of the M2 pancreas embeddings is shown as the correlation in function of increasing distance or incremental number of pixels shifted. As expected the correlation values are maximal for a zero shift and decrease in function of the distance. Shown are the results for two regions in the tissue for the Chebyshev distance metric. The spatial autocorrelation is shown for the embedding obtained on the same tissue section using BH t-SNE (orange) and UMAP (blue). Each graph also includes the spectral similarity (green) according to these metrics as a benchmark indicator for the spatial autocorrelation.

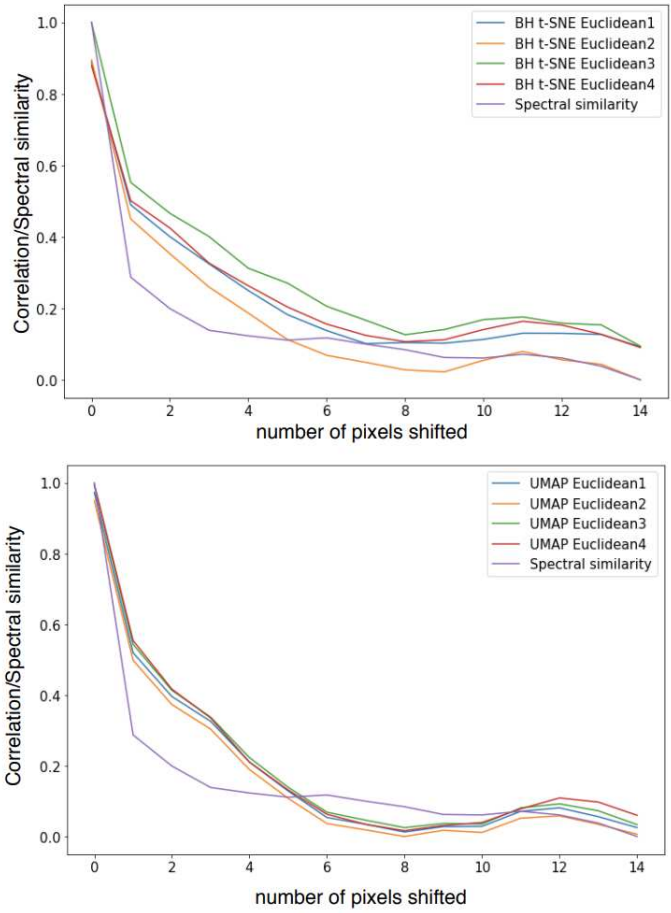


Figure A.6: Spatial autocorrelation of the M2 pancreas embeddings is shown as the correlation in function of increasing distance or incremental number of pixels shifted. As expected the correlation values are maximal for a zero shift and decrease in function of the distance. Shown are the results for the small region in the pancreas tissue. The spatial autocorrelation is shown for the embedding obtained on the same tissue section using BH t-SNE (top) and UMAP (bottom) for 4 repetitions using the Euclidean distance metric. Each graph also includes the spectral similarity (purple) according to the Euclidean distance metrics as a benchmark indicator for the spatial autocorrelation. The goal was to check robustness of the spatial autocorrelation when repeating the same experiment. We can derive from this figure that UMAP shows a very robust spatial autocorrelation in this regard, whereas the result for BH t-SNE is less robust in comparison to UMAP, but the autocorrelation clearly follows a similar trend.

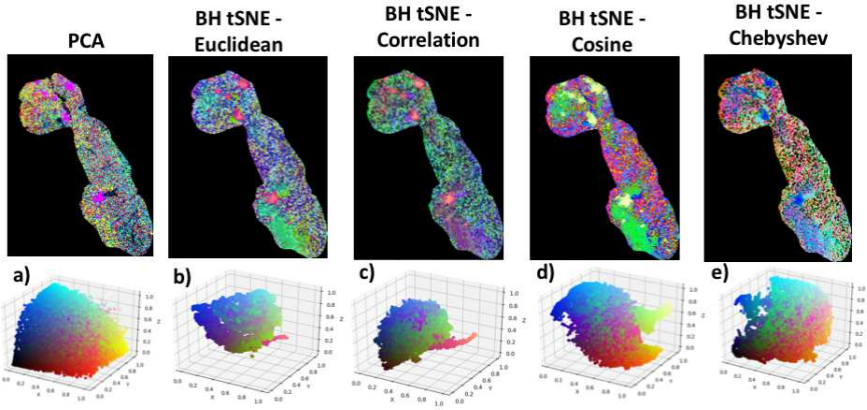


Figure A.7: Overview of pancreas (M2) embeddings obtained using PCA (a) and Barnes-Hut t-SNE applied with different distance metrics: Euclidean (b), correlation (c), cosine (d), and Chebyshev (e). Shown on top are the 3D embeddings of the tissue using hyperspectral visualisation with the corresponding scatter plots underneath.

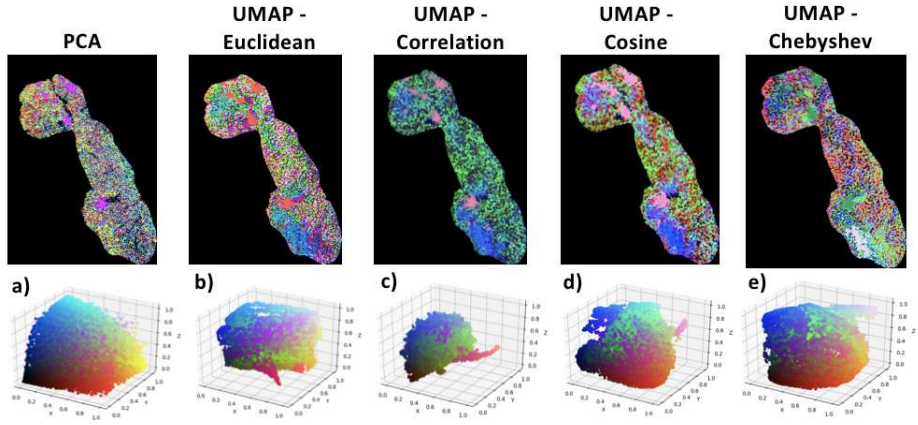


Figure A.8: Overview of pancreas (M2) embeddings obtained using PCA (a) and UMAP applied with different distance metrics: Euclidean (b), correlation (c), cosine (d), and Chebyshev (e). Shown on top are the 3D embeddings of the tissue using hyperspectral visualization with the corresponding scatter plots underneath.

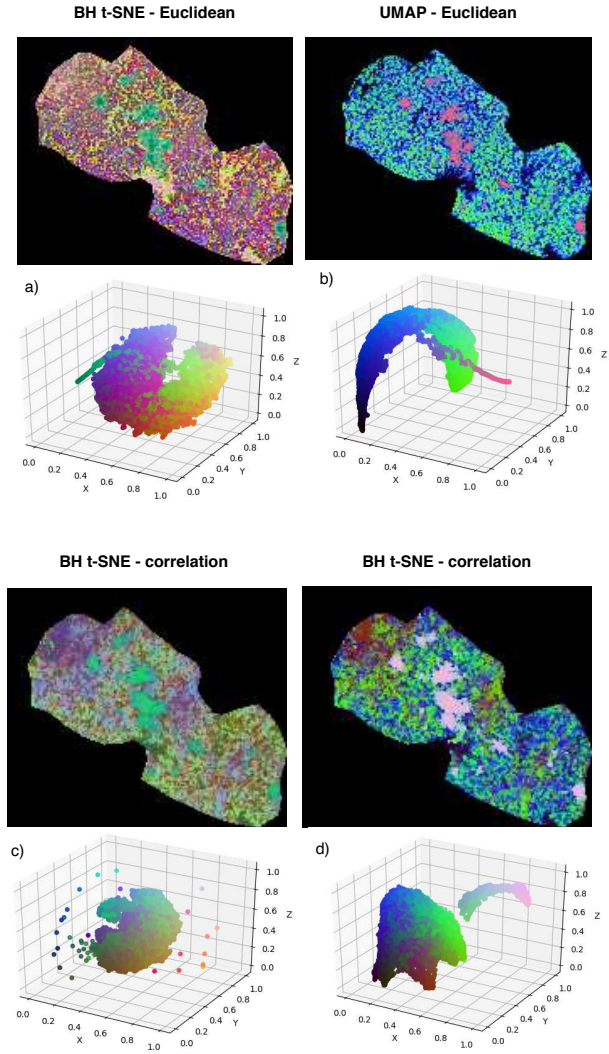


Figure A.9: Overview of pancreas (M1) embeddings without TIC normalization obtained using BH t-SNE (a-c) and UMAP (b-d) applied with the Euclidean (a-b) and correlation (c-d) distance metric. Shown on top are the 3D embeddings of the tissue using hyperspectral visualization with the corresponding scatter plots underneath.

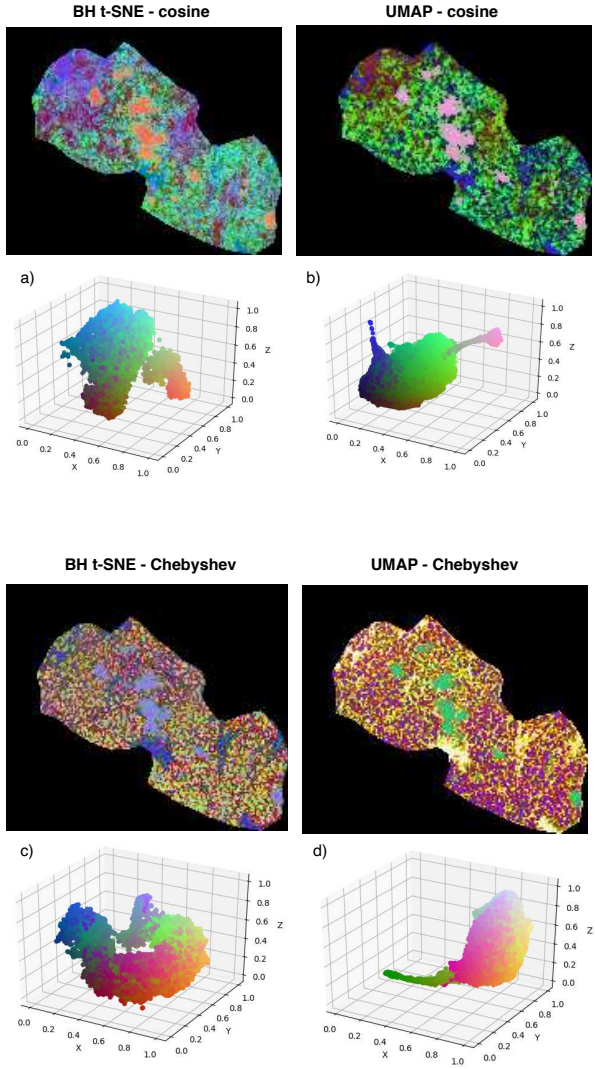


Figure A.10: Overview of pancreas (M1) embeddings without TIC normalisation obtained using BH t-SNE (a-c) and UMAP (b-d) applied with the cosine (a-b) and Chebyshev (c-d) distance metric. Shown on top are the 3D embeddings of the tissue using hyperspectral visualization with the corresponding scatter plots underneath.

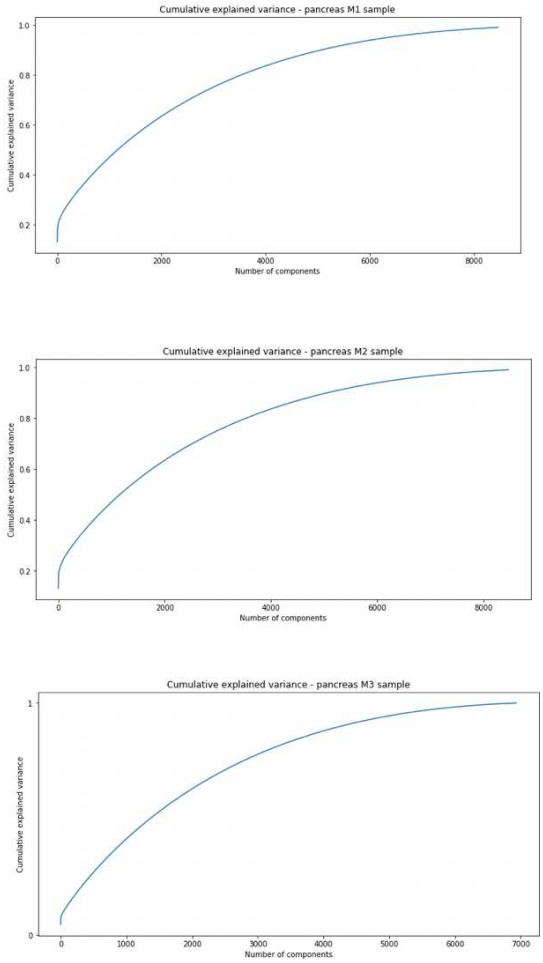


Figure A.11: The figures show the cumulative variance explained in function of the principal components. For all pancreas samples at least 10 components are needed to explain only 20 percent of the variance present in the data. This shows the importance of alternative visualization approaches that are capable of approximating the complete feature space into three or two dimensions.

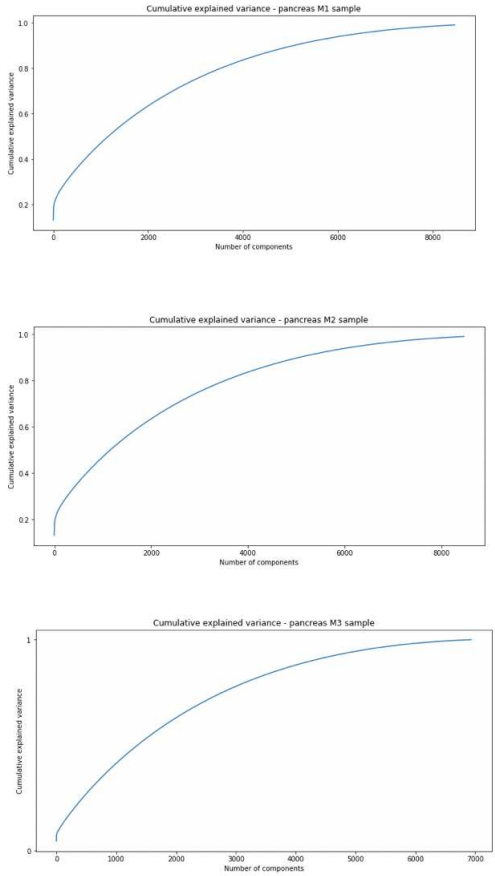


Figure A.12: The figures show the cumulative variance explained in function of the principal components. For all pancreas samples a large number of components are rrequired to explain 50 percent of the variance present in the data.

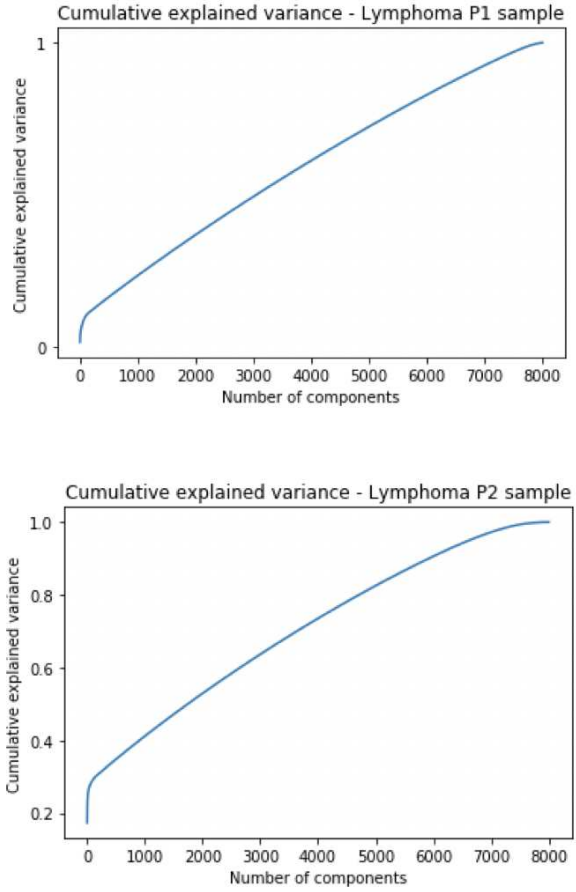


Figure A.13: The figures show the cumulative variance explained in function of the principal components. For all lymphoma samples at least 100 components are needed to explain only 10 or 30 percent of the variance present in the data. This shows the importance of alternative visualization approaches that are capable of approximating the complete feature space into three or two dimensions.

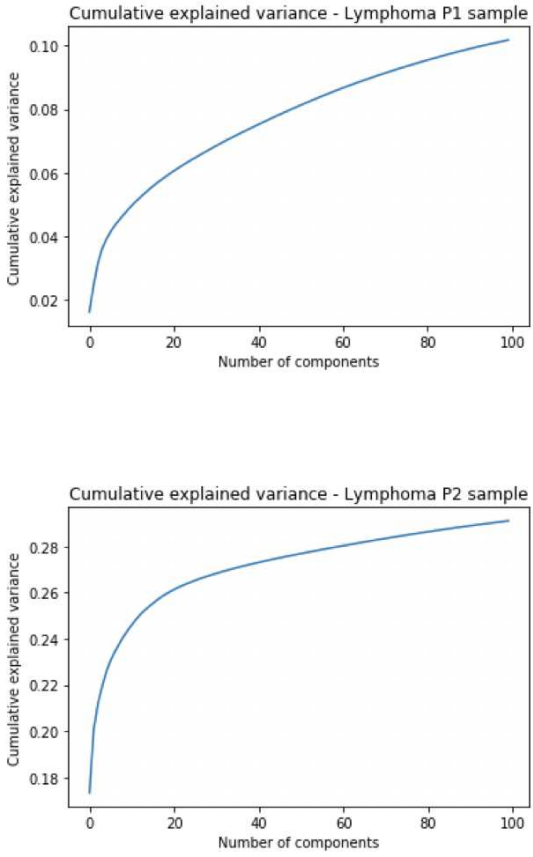


Figure A.14: The figures show the cumulative variance explained in function of the principal components. For all lymphoma samples at least 1000 components are needed to explain 50 percent of the variance present in the data.

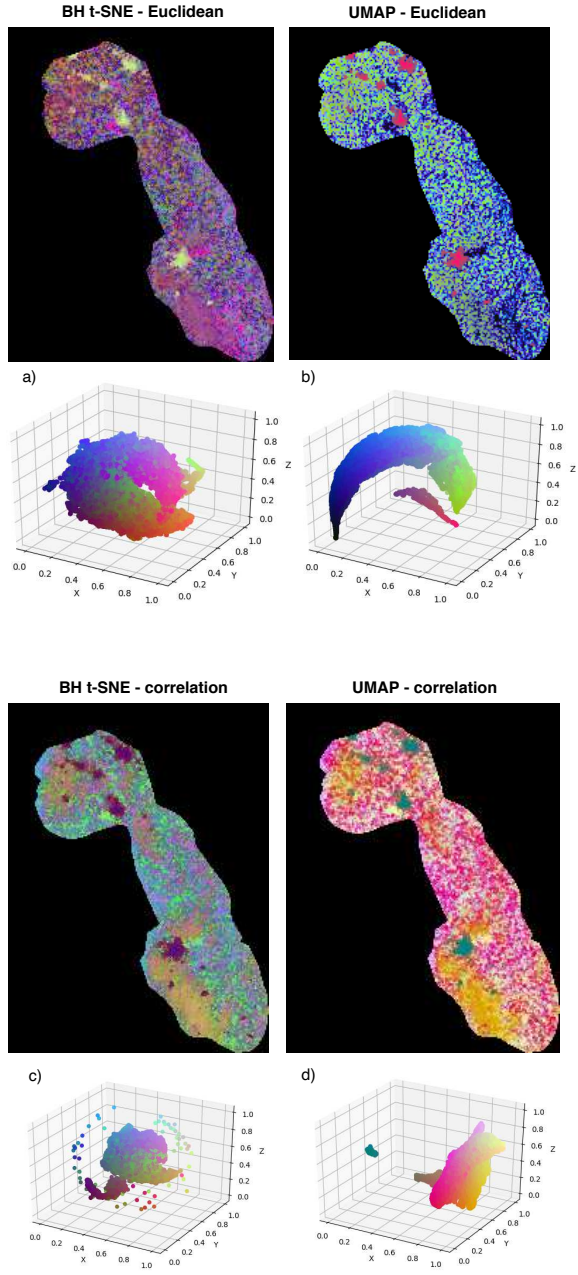


Figure A.15: Overview of pancreas (M2) embeddings without TIC normalisation obtained using BH t-SNE (a-c) and UMAP (b-d) applied with the Euclidean (a-b) and correlation (c-d) distance metric. Shown on top are the 3D embeddings of the tissue using hyperspectral visualisation with the corresponding scatter plots underneath.

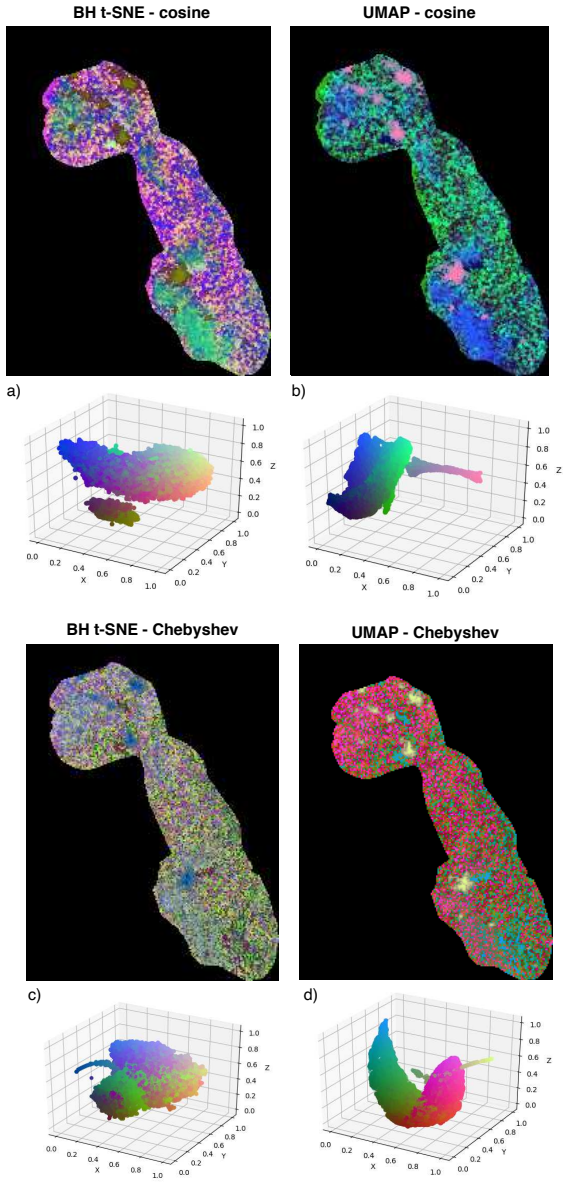


Figure A.16: Overview of pancreas (M2) embeddings without TIC normalisation obtained using BH t-SNE (a-c) and UMAP (b-d) applied with the cosine (a-b) and Chebyshev (c-d) distance metric. Shown on top are the 3D embeddings of the tissue using hyperspectral visualisation with the corresponding scatter plots underneath.

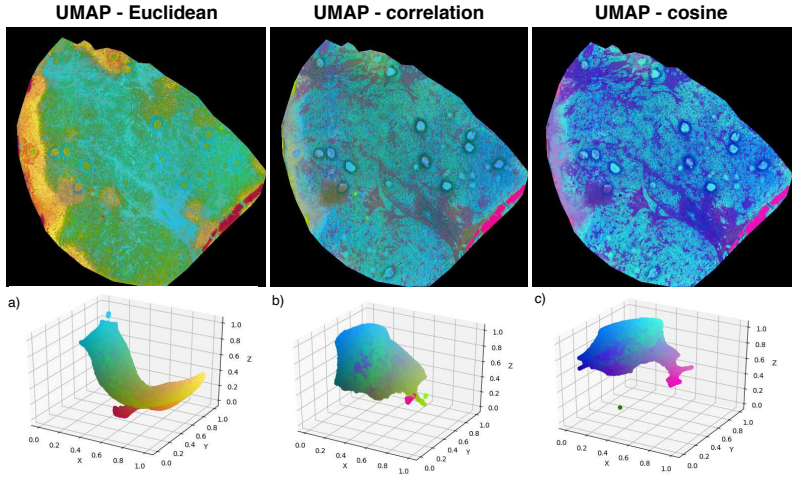


Figure A.17: Overview of lymphoma (P1) embeddings without TIC normalisation obtained using UMAP (a-c) applied with the Euclidean (a), correlation (b) and cosine (c) distance metric. Shown on top are the 3D embeddings of the tissue using hyperspectral visualisation with the corresponding scatter plots underneath.

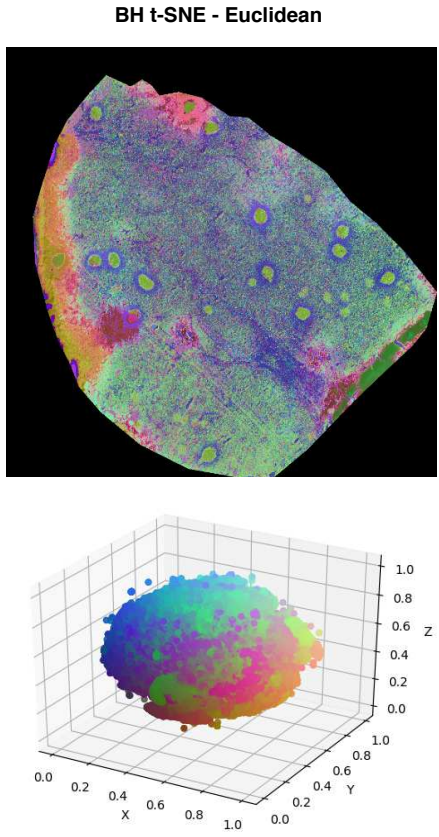


Figure A.18: Lymphoma (P1) embedding without TIC normalisation obtained using BH t-SNE applied with the Euclidean distance metric. Shown on top is the 3D embeddings of the tissue using hyperspectral visualisation with the corresponding scatter plot underneath.

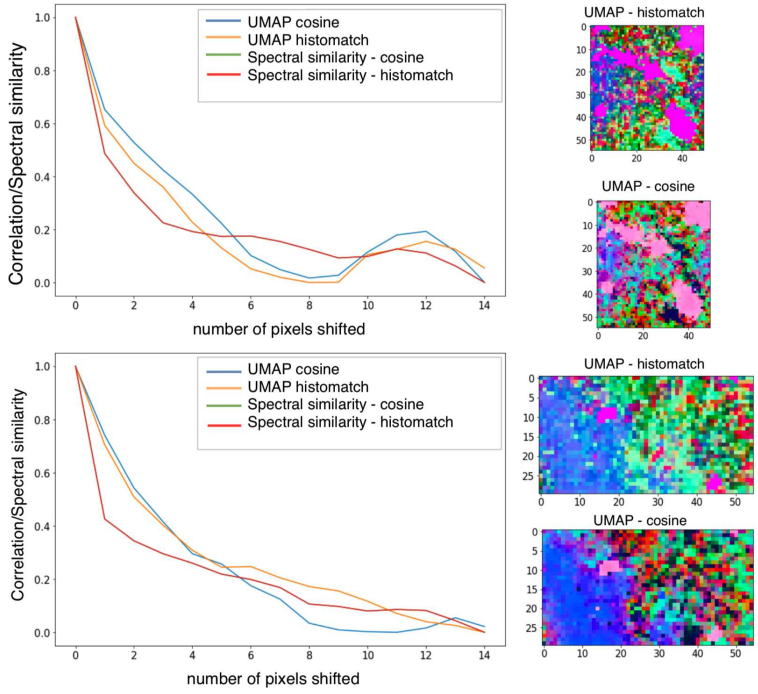


Figure A.19: Spatial autocorrelation of the M2 pancreas embeddings shown as the correlation in function of increasing distance or incremental number of pixels shifted. As expected the correlation values are maximal for a zero shift and decrease in function of the distance. Shown are the results for two regions in the tissue comparing the embeddings obtained using UMAP combined with the histomatch (orange) and the cosine (blue) distance metric . Each graph also includes the spectral similarity according to both the histomatch (red) and cosine (green) distance metric as a benchmark indicator for the spatial autocorrelation. The spectral similarity for these two distance metrics completely overlaps in this graph.

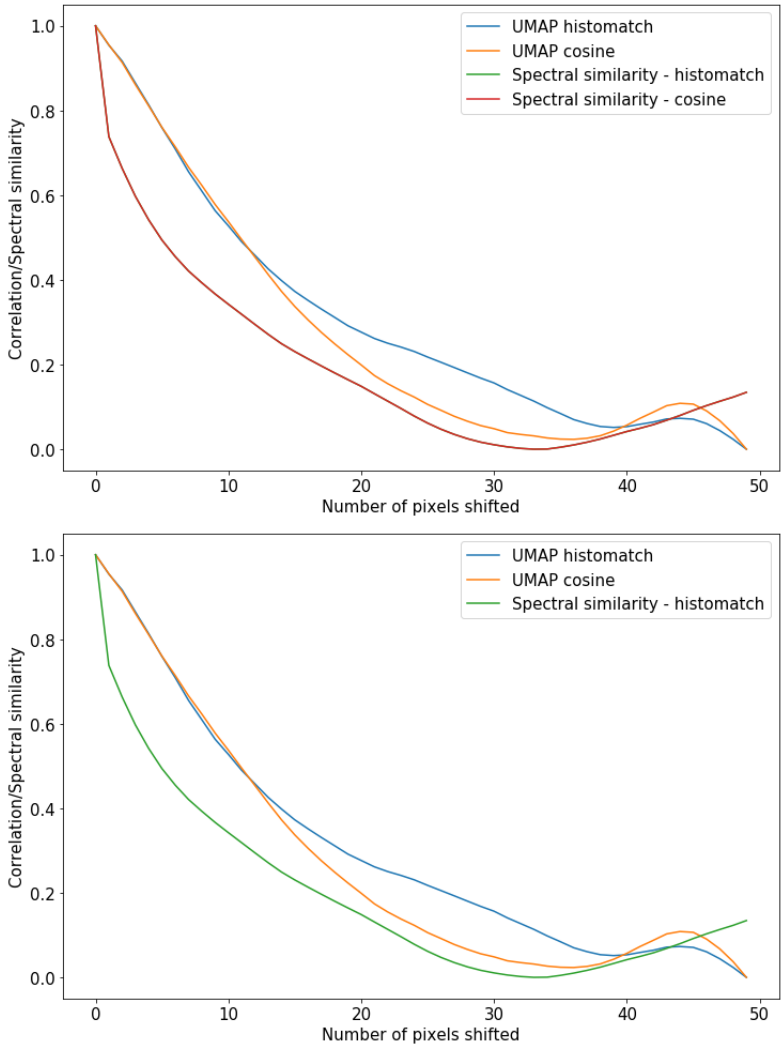


Figure A.20: Spatial autocorrelation of the P2 lymphoma embeddings shown as the correlation in function of increasing distance or incremental number of pixels shifted. As expected the correlation values are maximal for a zero shift and decrease in function of the distance. Each graph also includes the spectral similarity as a benchmark indicator for the spatial autocorrelation(orange). Panels (a-b) show that the spectral similarity associated with the histomatch (green, panel b) and cosine distance metric overlap (red, panel a)

Appendix B

Prioritization of m/z values in mass spectrometry imaging profiles obtained using UMAP for dimensionality reduction

Hyperspectral S1 Profile 3

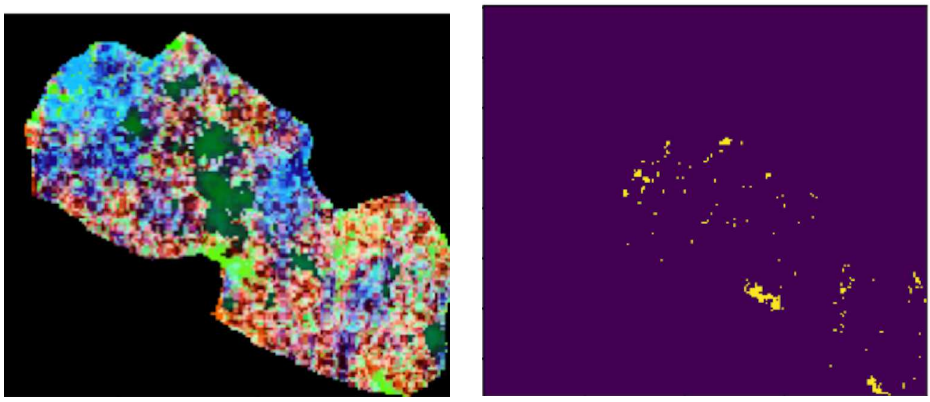


Figure B.1: On the left the hyperspectral visualization for sample 1 obtained from a healthy mouse pancreas is shown with on the right the extracted profile corresponding to the light green region.

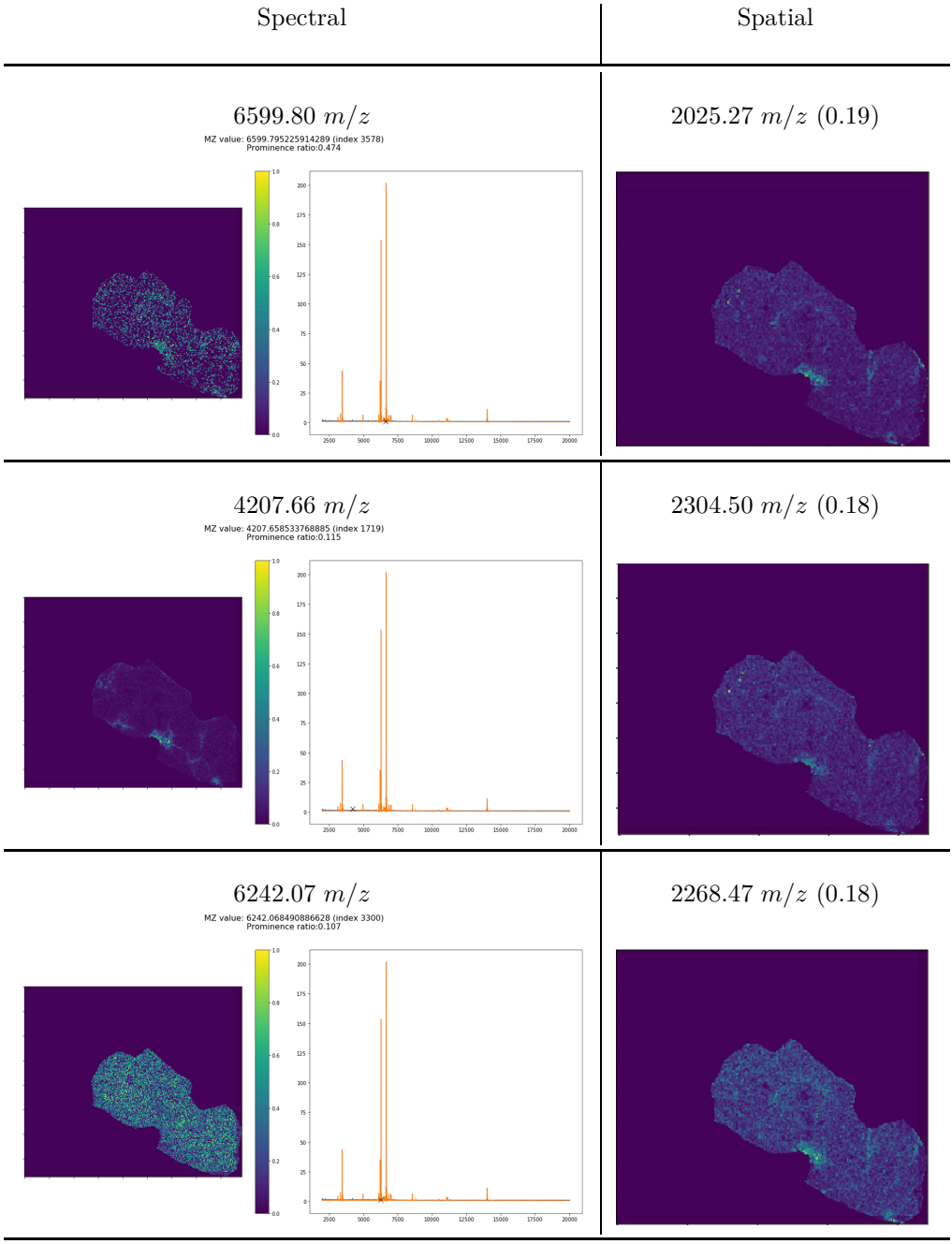


Figure B.2: Top three of spectral (left) and spatial (right) results ranked vertically from top to bottom. For the spectral results the particular ion image corresponding to the highlighted peak is shown together with the median spectra intensities associated with the extracted profile (in blue) against the median overall spectra intensities (orange). For the spatial results the particular ion images are shown together with their correlation values for the highest ranked ion image per cluster which were also ranked according to their highest average correlation to the extracted profile.

Hyperspectral S1 Profile 4

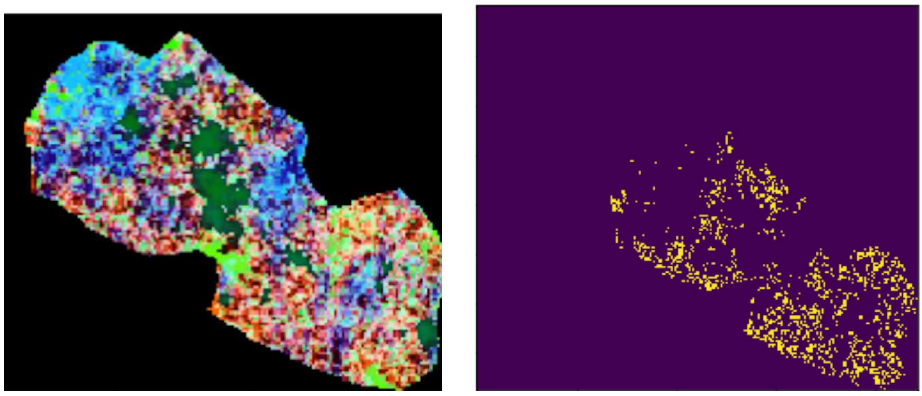


Figure B.3: On the left the hyperspectral visualization for sample 1 obtained from a healthy mouse pancreas is shown with on the right the extracted profile corresponding to the red region.

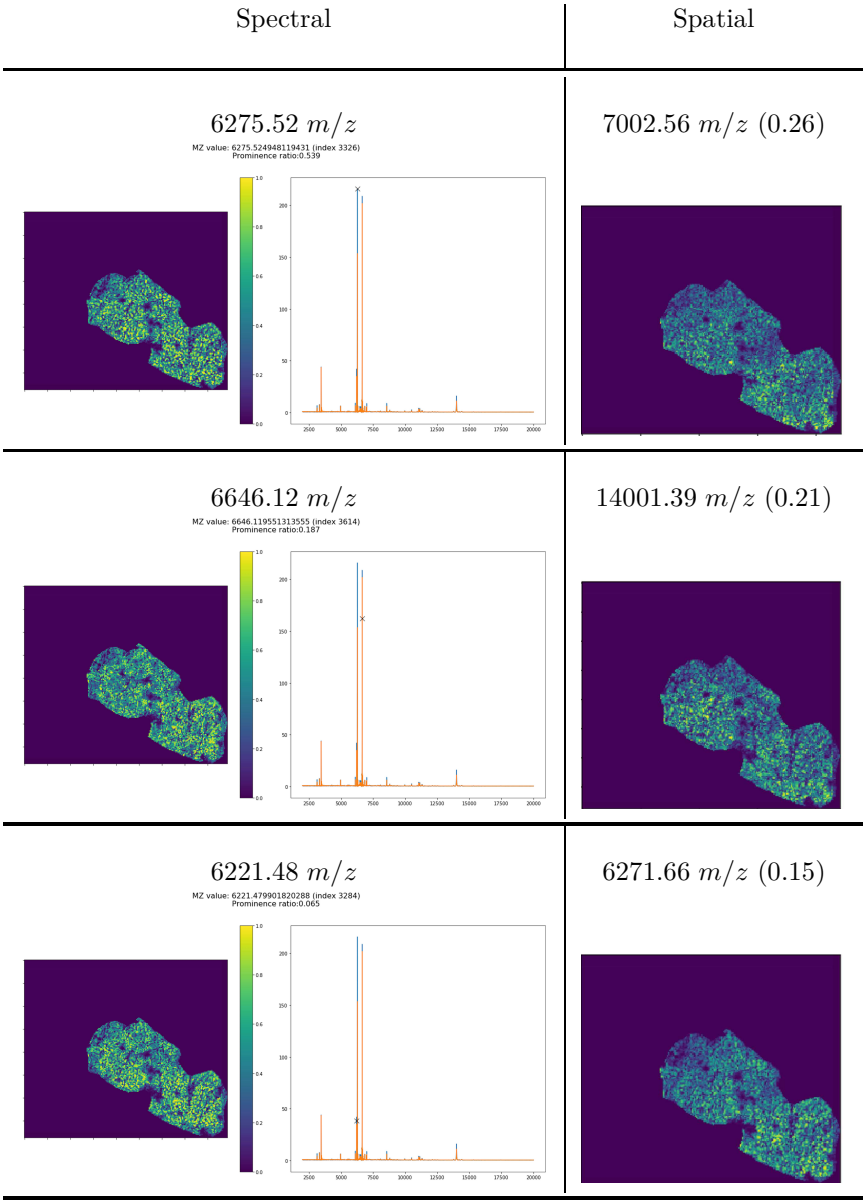


Figure B.4: Top three of spectral (left) and spatial (right) results ranked vertically from top to bottom. For the spectral results the particular ion image corresponding to the highlighted peak is shown together with the median spectra intensities associated with the extracted profile (in blue) against the median overall spectra intensities (orange). For the spatial results the particular ion images are shown together with their correlation values for the highest ranked ion image per cluster which were also ranked according to their highest average correlation to the extracted profile.

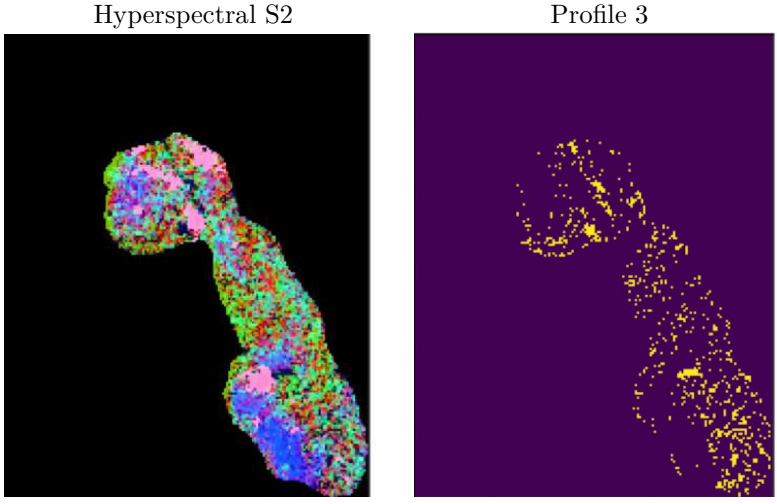


Figure B.5: On the left the hyperspectral visualization for sample 2 obtained from a healthy mouse pancreas is shown with on the right the extracted profile corresponding to the dark blue region.

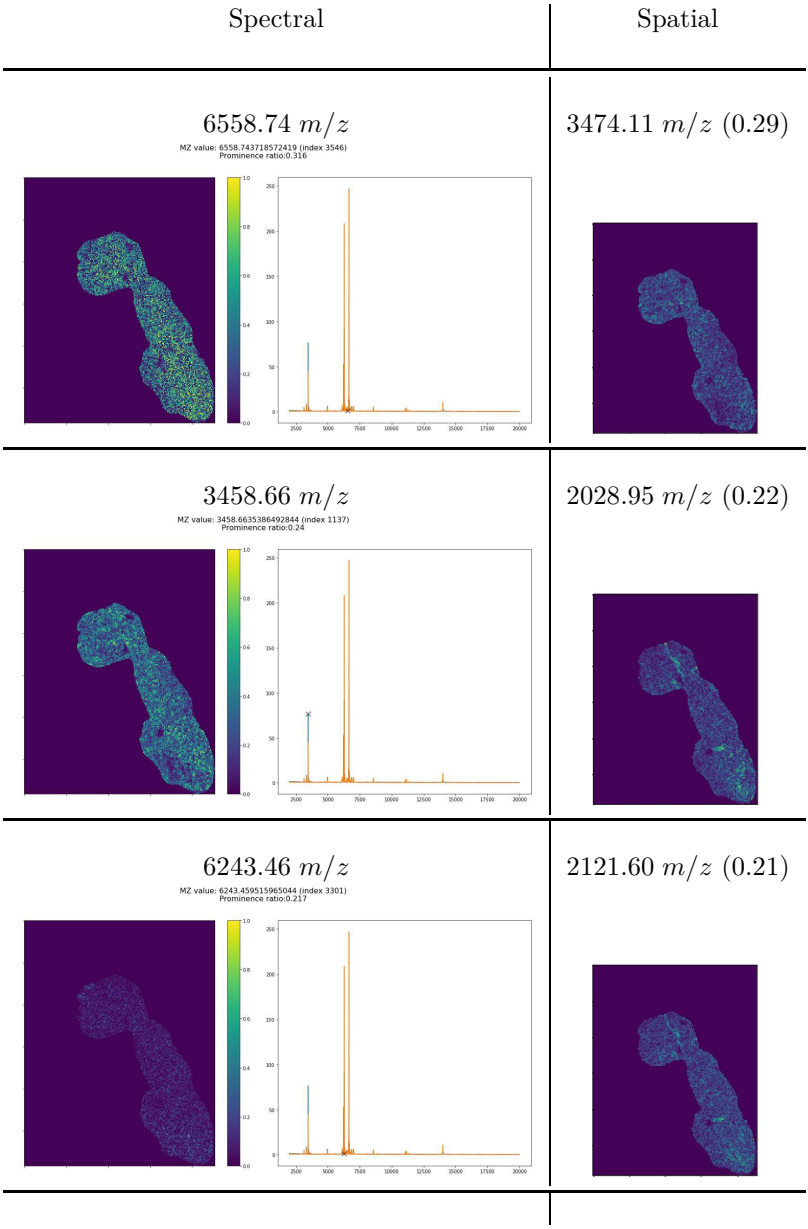


Figure B.6: Top three of spectral (left) and spatial (right) results ranked vertically from top to bottom. For the spectral results the particular ion image corresponding to the highlighted peak is shown together with the median spectra intensities associated with the extracted profile (in blue) against the median overall spectra intensities (orange). For the spatial results the particular ion images are shown together with their correlation values for the highest ranked ion image per cluster which were also ranked according to their highest average correlation to the extracted profile.

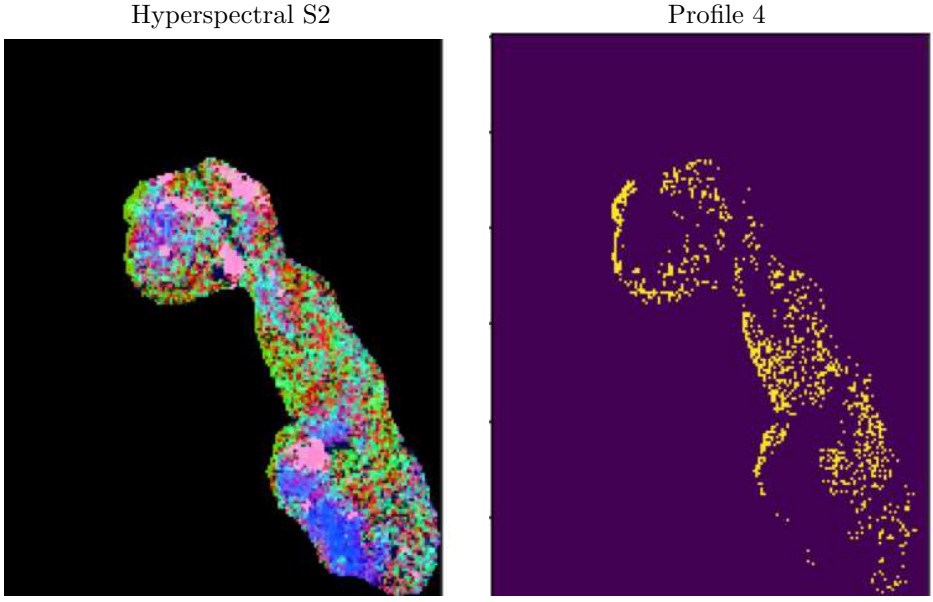


Figure B.7: On the left the hyperspectral visualization for sample 2 obtained from a healthy mouse pancreas is shown with on the right the extracted profile corresponding to the light green region.

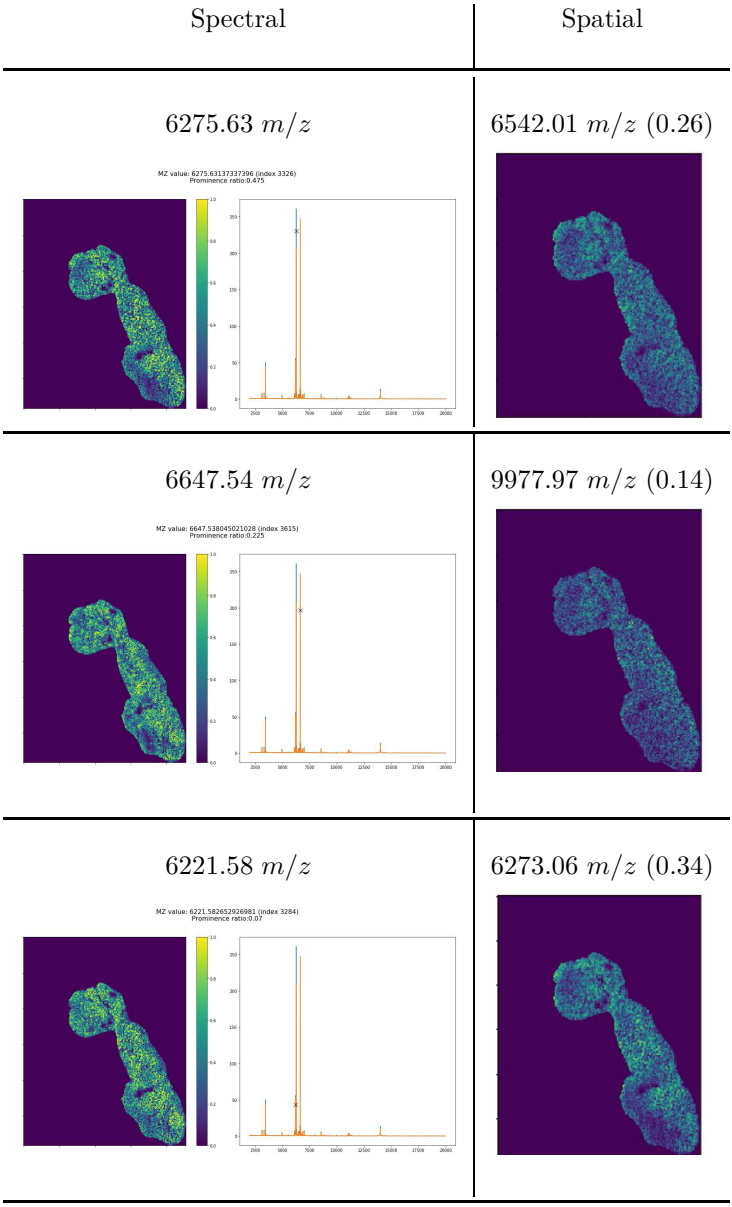


Figure B.8: Top three of spectral (left) and spatial (right) results ranked vertically from top to bottom. For the spectral results the particular ion image corresponding to the highlighted peak is shown together with the median spectra intensities associated with the extracted profile (in blue) against the median overall spectra intensities (orange). For the spatial results the particular ion images are shown together with their correlation values for the highest ranked ion image per cluster which were also ranked according to their highest average correlation to the extracted profile.

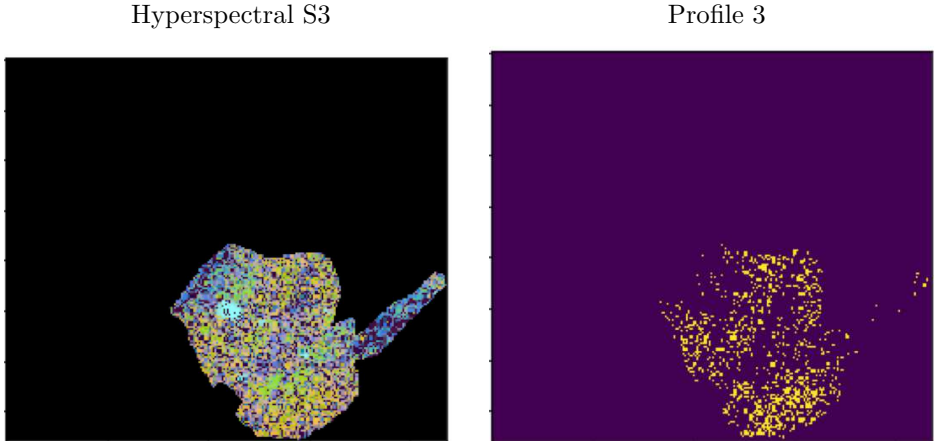


Figure B.9: On the left the hyperspectral visualization for sample 3 obtained from a healthy mouse pancreas is shown with on the right the extracted profile corresponding to the yellow region.

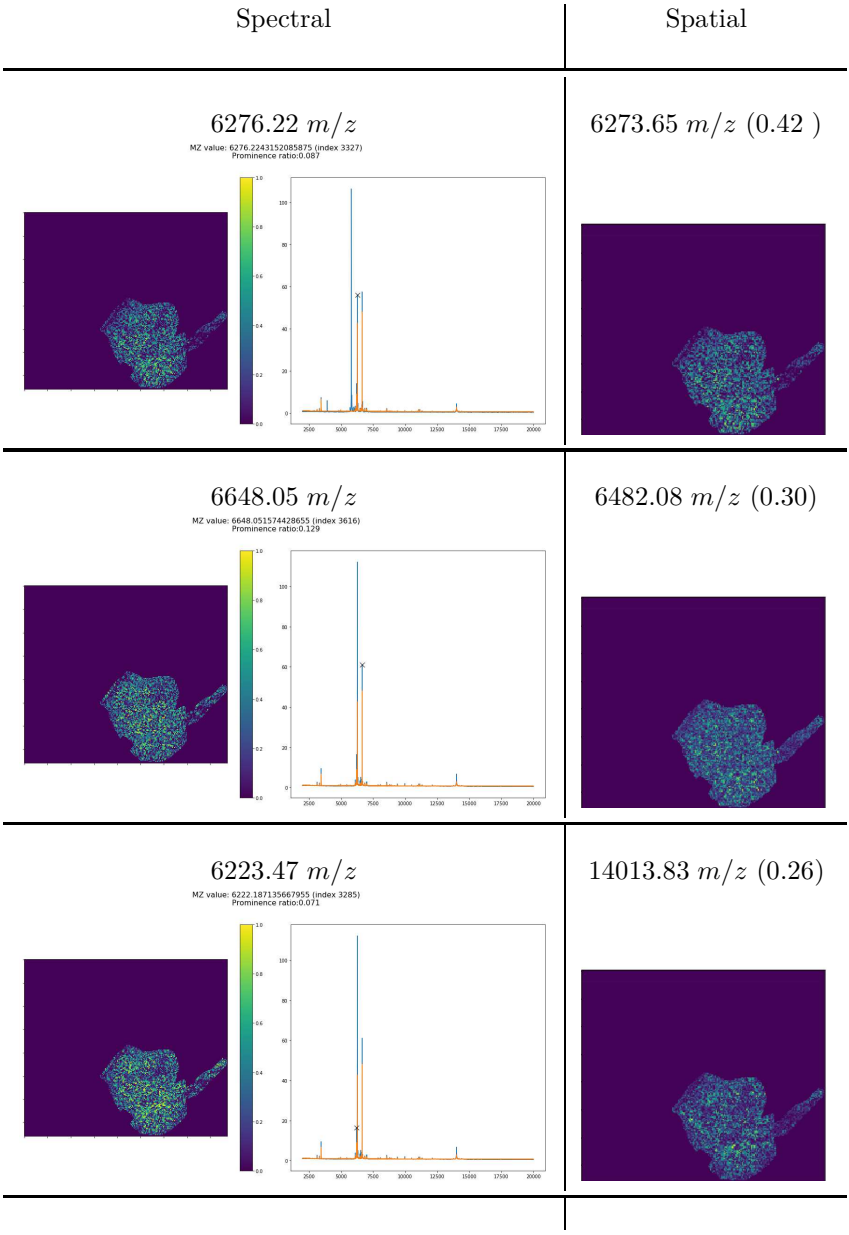


Figure B.10: Top three of spectral (left) and spatial (right) results ranked vertically from top to bottom. For the spectral results the particular ion image corresponding to the highlighted peak is shown together with the median spectra intensities associated with the extracted profile (in blue) against the median overall spectra intensities (orange). For the spatial results the particular ion images are shown together with their correlation values for the highest ranked ion image per cluster which were also ranked according to their highest average correlation to the extracted profile.

Hyperspectral S3 Profile 4

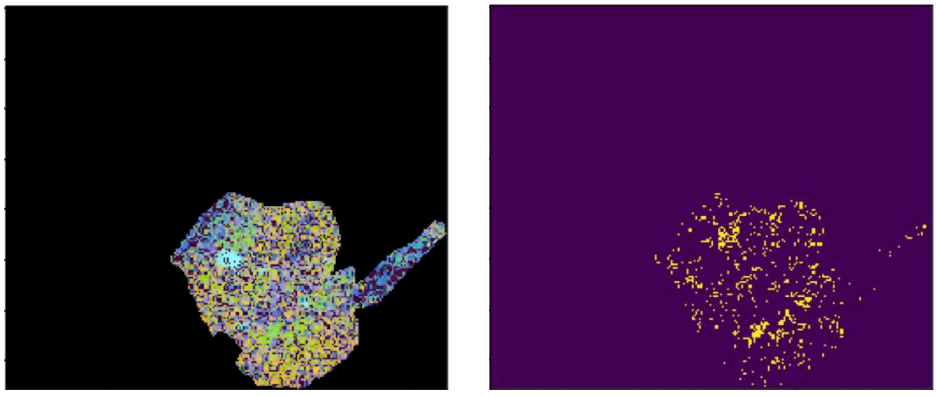


Figure B.11: On the left the hyperspectral visualization for sample 3 obtained from a healthy mouse pancreas is shown with on the right the extracted profile corresponding to the light green region.

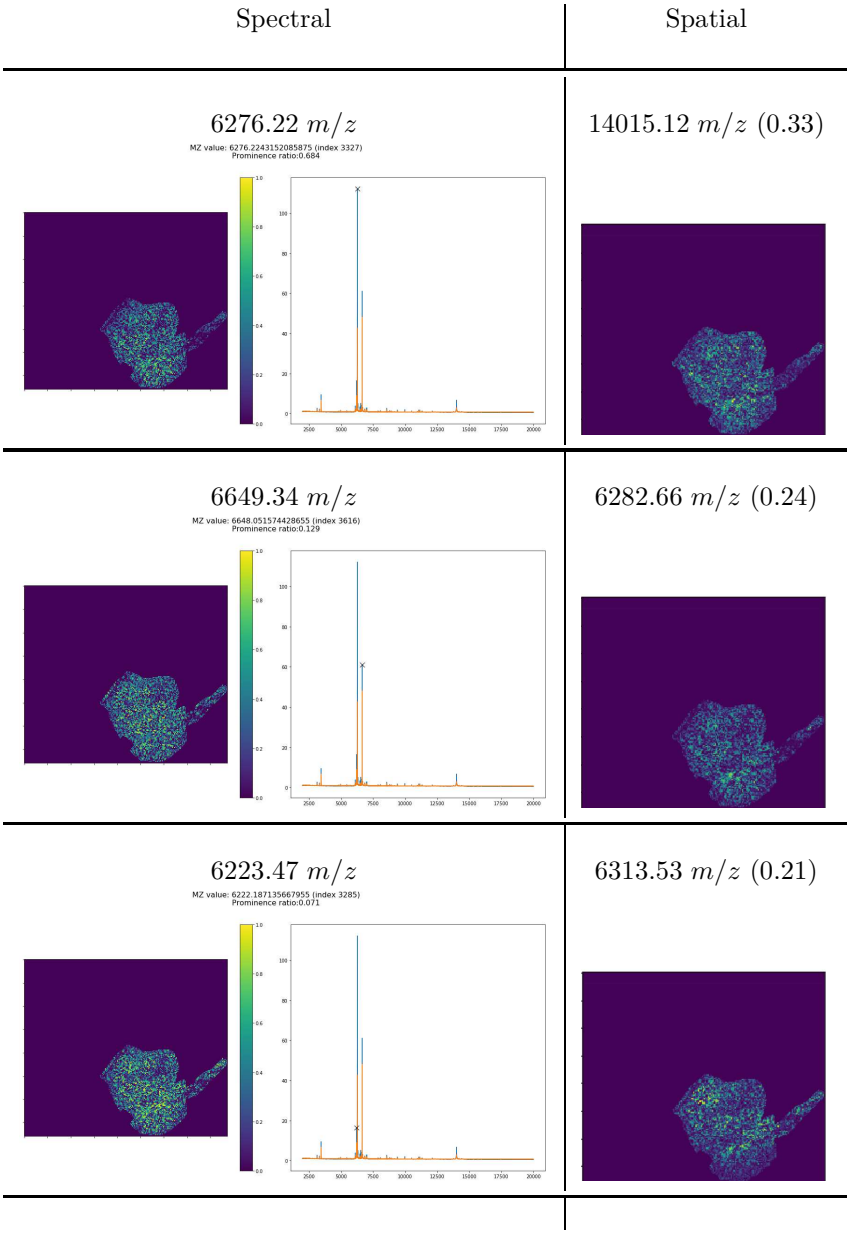


Figure B.12: Top three of spectral (left) and spatial (right) results ranked vertically from top to bottom. For the spectral results the particular ion image corresponding to the highlighted peak is shown together with the median spectra intensities associated with the extracted profile (in blue) against the median overall spectra intensities (orange). For the spatial results the particular ion images are shown together with their correlation values for the highest ranked ion image per cluster which were also ranked according to their highest average correlation to the extracted profile.

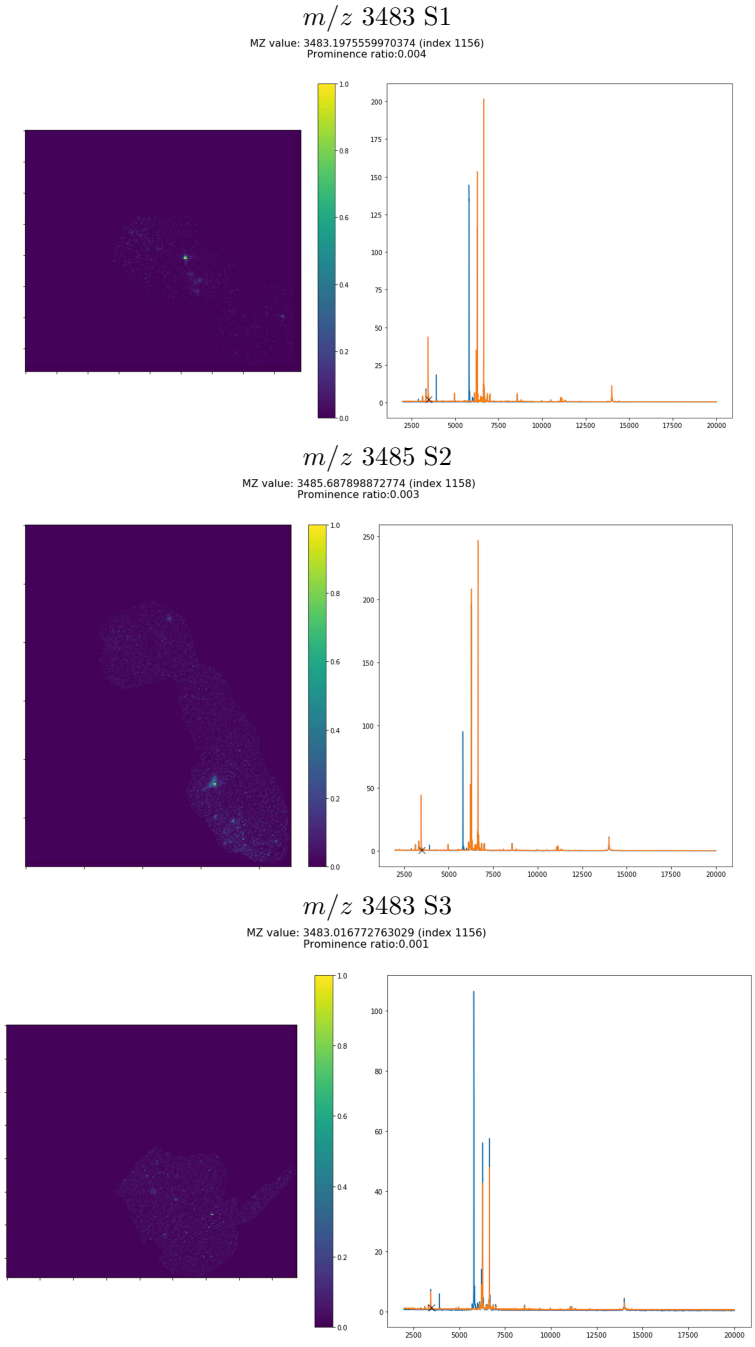
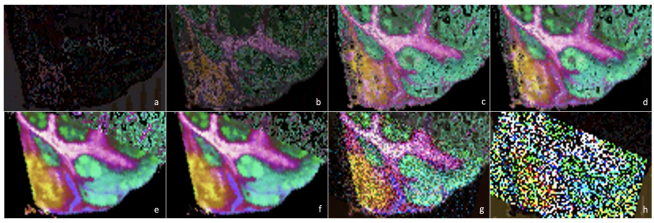


Figure B.13: The m/z value around 3484 Da could putatively be associated to glucagon, which is a key player in glucose metabolism. This molecule is present at a much lower abundance and is therefore ranked at a lower position (rank 16 for S1, 17 for S2 and 20 for S3 respectively). This result is obtained for all tissues based on the binary profile corresponding to the islets of Langerhans.

Appendix C

Supporting Information for Correspondence-aware manifold learning for microscopic and spatial omics imaging: a novel data fusion method bringing Mass Spectrometry Imaging to a cellular resolution



#	μ	$E(A;\mu)$
a	0.0001	0.9298
b	0.001	0.9659
c	0.01	0.9758
d	**0.015**	**0.9799**
e	0.1	0.9717
f	1.0	0.9641
g	10	0.9562
h	100	0.91

Figure C.1: Overview of the behavior of an increasing penalty parameter μ. The second column in the table contains the values for the penalty parameter μ that were used. The next column includes the values for the measure $E(A;\mu)$. This measure expresses a trade-off between the trustworthiness of the embedding and the repulsion of the data. Based on $E(A;\mu)$ the optimal value for μ is 0.015.

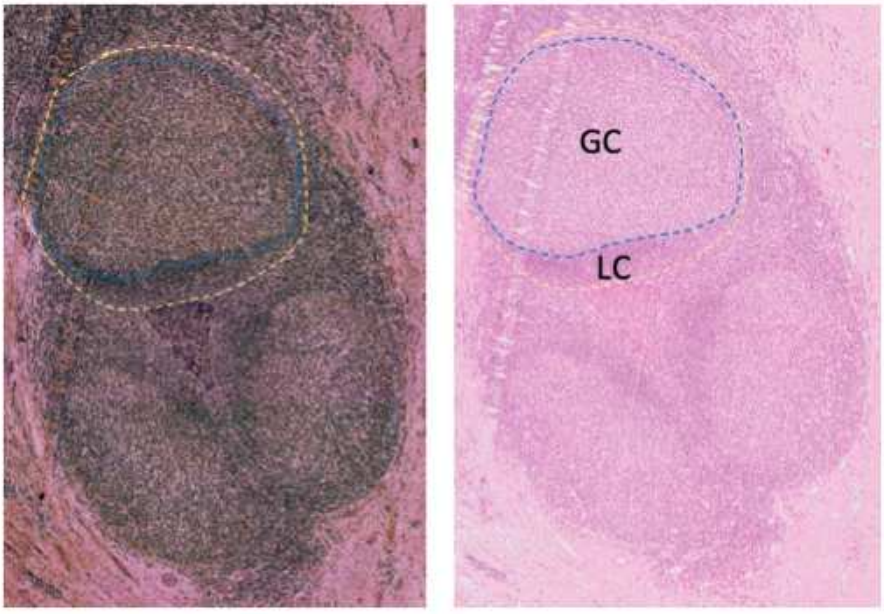


Figure C.2: The left panel shows a detail of the fused data with on the right the corresponding H&E image. Highlighted with a yellow dashed line are the secondary B follicles, the blue dashed line encircles the germinal center (GC) and the lymphocyte corona (LC) respectively. The nuclear-cytoplasmic ratio is higher in the LC than in the GC, which explains the denser/darker appearance of the LC. This is explained by the presence of smaller lymphoid cells, corresponding to mature, naive B-cells in the LC, compared to the larger centrocytes and centroblasts in the GC.

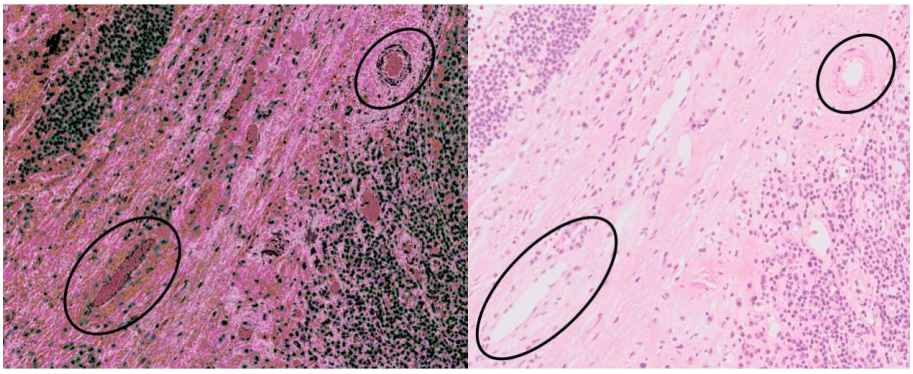


Figure C.3: The left panel shows a detail of the fused data with on the right the corresponding H&E image. Highlighted in both images on the left is a venule with a thin wall that seems to correspond to a single cell layer of endothelial cells. Highlighted in both images on the right is a small arteriole with a thicker wall that also seems to contain some smooth muscle cells. The lumina of the vessels are stuffed with red blood cells. In light pink, in the stromal area between the venes and artery, we can also observe some collagen fibers.

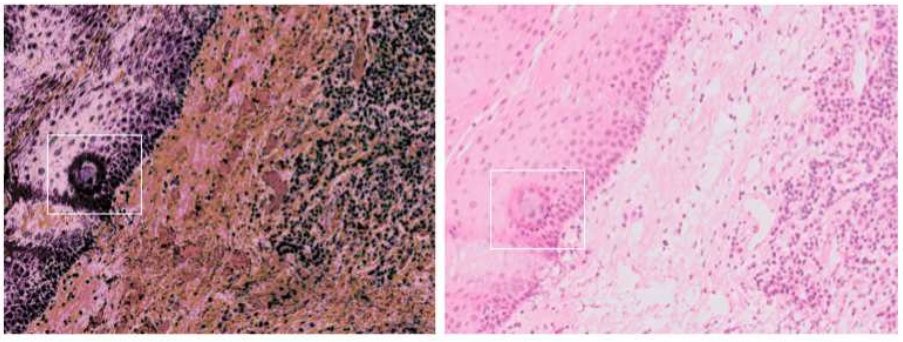


Figure C.4: The left panel shows a detail of the fused data with on the right the corresponding H&E image. The epithelium contains a small, rounded structure where the intensity is increased. While this structure is also visible in the H&E image, it becomes more pronounced after data fusion. It is regarded as an artefact created during the cutting process and not a cellular structure. As such data fusion can support in distinguishing artefacts from true biological signals.

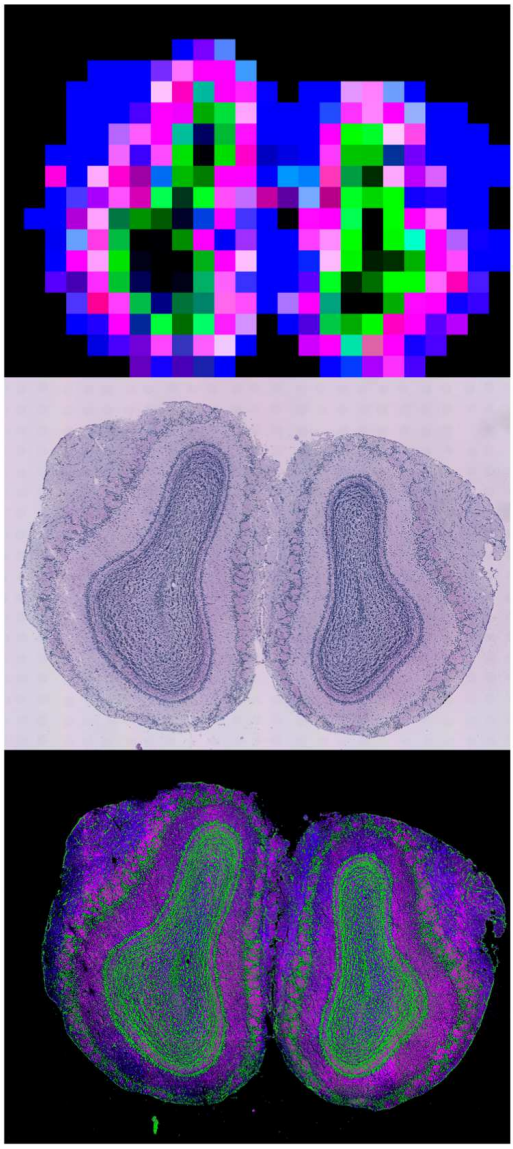


Figure C.5: An example demonstrates the application of our method to a public spatial transcriptomics dataset. Shown on top:the low dimensional representation of a mouse brain (281 pixels $\times$ 16416 features, 100μ m resolution). The different colours reflect the molecular trends present in the data obtained using UMAP. Below, the corresponding microscopy image is shown, with underneath the fused result to demonstrate the general applicability. Improvements in terms of spatial resolution will make it possible to obtain results similar to the ones shown for MSI data (10 μ m resolution).

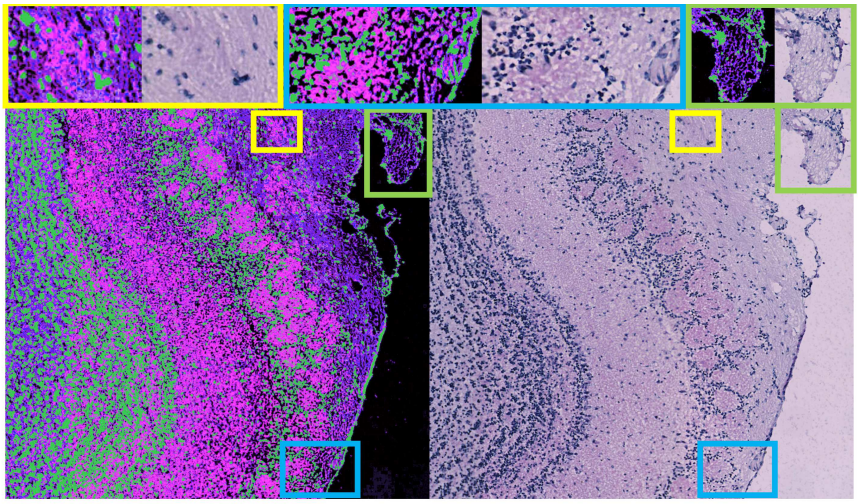


Figure C.6: Illustration of consistent mapping of cell nuclei by green color upon data fusion of spatial transcriptomics mouse brain data with corresponding H&E staining. A higher spatial resolution of the data could yield more detailed results.

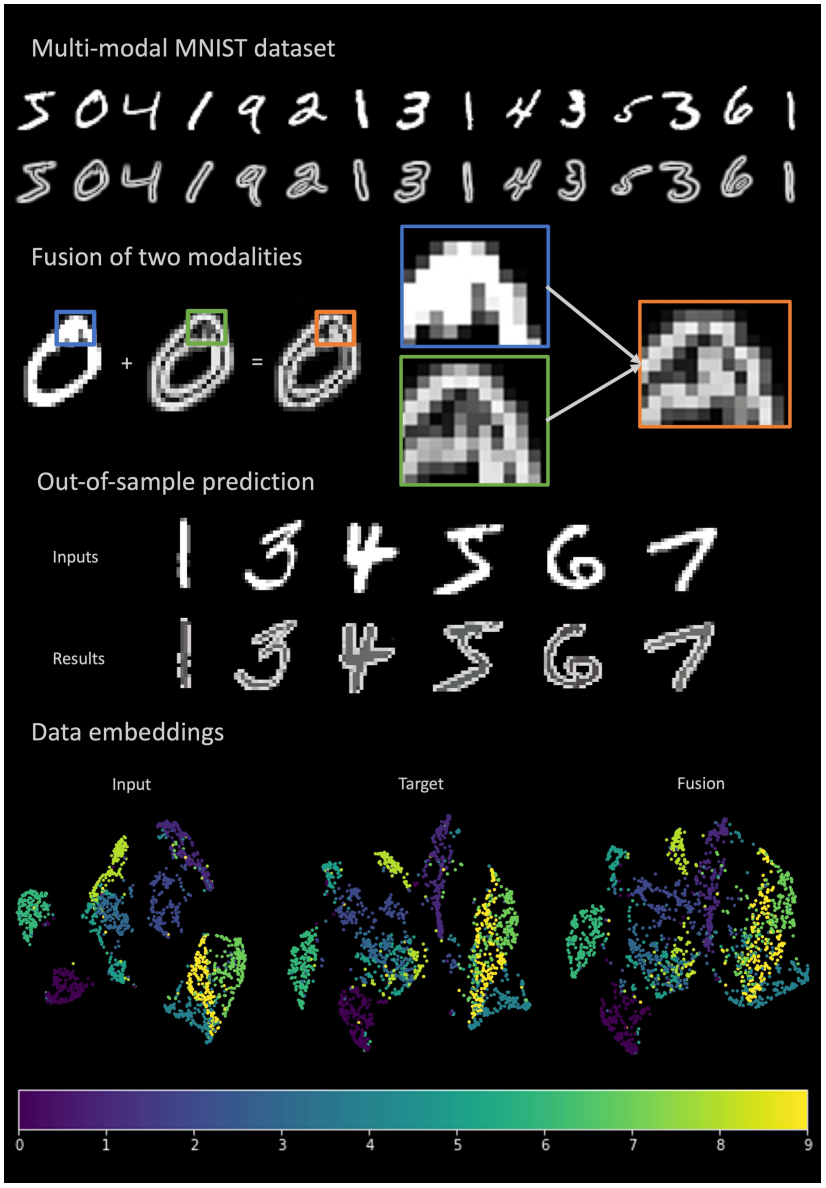


Figure C.7: CAML validation and general applicability. We built a multi-modal dataset based on MNIST, the handwritten digits dataset. (1) Using a Sobel filter we created corresponding data with a shared latent space. (2) Fusion – correspondence matrix is the diagonal matrix. (3) Transform instances based on the learned model from (2). (4) UMAP embeddings of the multi-modal dataset and the fused data. We see that UMAP can still distinguish groups from the fused dataset based on the shared latent space of the two original datasets.

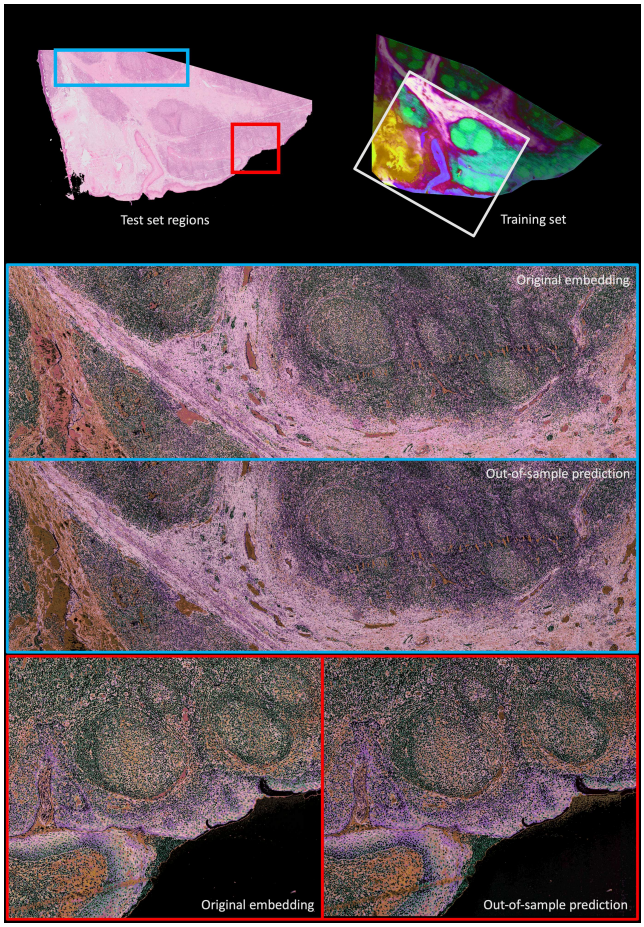


Figure C.8: Validation of CAML out-of-sample prediction. On the right, the unsupervised, hyperspectral visualization of the MSI data is shown (ie. after dimensionality reduction using UMAP) with the white square corresponding to the selected training data. The blue and red squares on the left highlight the regions predicted using out-of-sample prediction. For both regions, the results show that the out-of-sample prediction is very similar to the result obtained with CAML based on the original embedding for the MSI data. This suggests that CAML provides robust out-of-sample predictions.

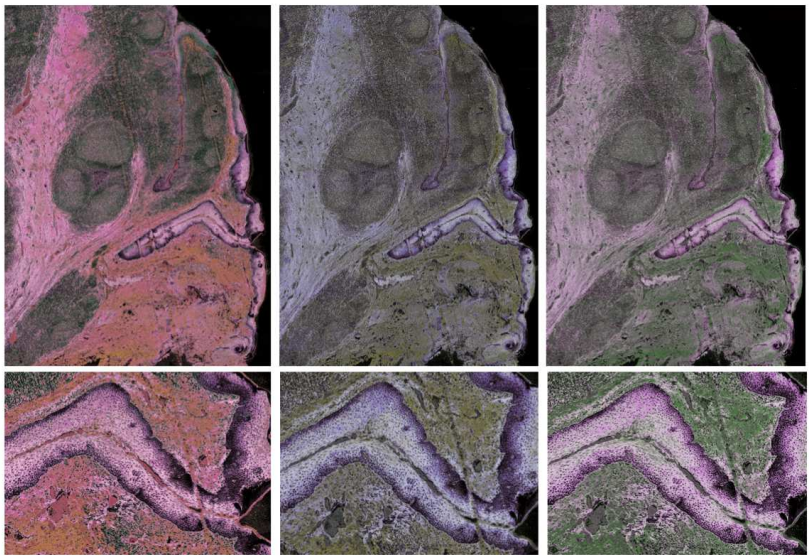


Figure C.9: Evaluation of the impact on final result when changing the order of the RGB axes.The first figure on the left shows the original result (RGB) to be compared with the figures towards the right which are the GRB and BGR versions respectively. We can see that the color perception is indeed different as to be expected but the overall results are unaffected by changing the ordering of the RGB axes.

Bibliography

[1] Aggarwal, C. C. *Machine Learning for Text*, 1st ed. Springer Publishing Company, Incorporated, 2018.

[2] Aggarwal, C. C., Hinneburg, A., and Keim, D. A. On the surprising behavior of distance metrics in high dimensional spaces. In *Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings.* (2001), pp. 420–434.

[3] Ahrén, B., Wierup, N., and Sundler, F. Neuropeptides and the regulation of islet function. *Diabetes 55*, Supplement 2 (2006), S98–S107.

[4] Aichler, M., and Walch, A. Maldi imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. *Laboratory investigation 95*, 4 (2015), 422–431.

[5] Alexandrov, T. Spatial metabolomics and imaging mass spectrometry in the age of artificial intelligence. *Annual Review of Biomedical Data Science 3*.

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

[7] 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 (2010), 6535–6546.

[8] Alexandrov, T., Chernyavsky, I., Becker, M., von Eggeling, F., and Nikolenko, S. Analysis and interpretation of imaging mass spectrometry data by clustering mass-to-charge images according to their spatial similarity. *Analytical chemistry 85*, 23 (2013), 11189–11195.

[9] Alexandrov, T., and Kobarg, J. H. Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. *Bioinformatics 27*, 13 (2011), i230–i238.

[10] Anderton, C. R., and Gamble, L. J. Secondary ion mass spectrometry imaging of tissues, cells, and microbial systems. *Microscopy today 24*, 2 (2016), 24–31.

[11] Balluff, B., Rauser, S., Meding, S., Elsner, M., Schöne, C., Feuchtinger, A., Schuhmacher, C., Novotny, A., Jütting, U., Maccarrone, G., et al. Maldi imaging identifies prognostic seven-protein signature of novel tissue markers in intestinal-type gastric cancer. *The American journal of pathology 179*, 6 (2011), 2720–2729.

[12] Barnes, K. The pixelation of mass spectrometry. *Nature Methods 12*, 1 (2015), 12–12.

[13] Barsoum, I., Tawedrous, E., Faragalla, H., and Yousef, G. M. Histo-genomics: digital pathology at the forefront of precision medicine. *Diagnosis 6*, 3 (2019), 203 – 212.

[14] Bauer, J. A., Chakravarthy, A. B., Rosenbluth, J. M., Mi, D., Seeley, E. H., Granja-Ingram, N. D. M., Olivares, M. G., Kelley, M. C., Mayer, I. A., Meszoely, I. M., et al. Identification of markers of taxane sensitivity using proteomic and genomic analyses of breast tumors from patients receiving neoadjuvant paclitaxel and radiation. *Clinical Cancer Research 16*, 2 (2010), 681–690.

[15] Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., and Newell, E. W. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology 37*, 1 (2019), 38–44.

[16] Belkin, M., and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. 1373–1396.

[17] Bemis, K. D., Harry, A., Eberlin, L. S., Ferreira, C. R., van de Ven, S. M., Mallick, P., Stolowitz, M., and Vitek, O. Probabilistic segmentation of mass spectrometry (ms) images helps select important ions and characterize confidence in the resulting segments. *Molecular & Cellular Proteomics 15*, 5 (2016), 1761–1772.

[18] Bereman, M. S., Williams, T. I., and Muddiman, D. C. Carbohydrate analysis by desorption electrospray ionization fourier transform ion cyclotron resonance mass spectrometry. *Analytical chemistry 79*, 22 (2007), 8812–8815.

[19] Bergman, H.-M., Lundin, E., Andersson, M., and Lanekoff, I. Quantitative mass spectrometry imaging of small-molecule neurotransmitters in rat brain tissue sections using nanospray desorption electrospray ionization. *Analyst 141*, 12 (2016), 3686–3695.

[20] Berthomieu, C., and Hienerwadel, R. Fourier transform infrared (ftir) spectroscopy. *Photosynthesis research 101*, 2-3 (2009), 157–170.

[21] Bishop, A., Polak, J., Green, I., Bryant, M., and Bloom, S. The location of vip in the pancreas of man and rat. *Diabetologia 18*, 1 (1980), 73–78.

[22] 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 (2018), 240–265.

[23] 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 (January 2018), 240âĂŤ265.

[24] Buck, M. D., Sowell, R. T., Kaech, S. M., and Pearce, E. L. Metabolic instruction of immunity. *Cell 169*, 4 (2017), 570–586.

[25] Burgess, D. J. Spatial transcriptomics coming of age. *Nature Reviews Genetics 20*, 6 (2019), 317–317.

[26] Burgoyne, T. W., and Hieftje, G. M. An introduction to ion optics for the mass spectrograph. *Mass spectrometry reviews 15*, 4 (1996), 241–259.

[27] Burnum-Johnson, K. E., Zheng, X., Dodds, J. N., Ash, J., Fourches, D., Nicora, C. D., Wendler, J. P., Metz, T. O., Waters, K. M., Jansson, J. K., et al. Ion mobility spectrometry and the omics: distinguishing isomers, molecular classes and contaminant ions in complex samples. *TrAC Trends in Analytical Chemistry 116* (2019), 292–299.

[28] Caprioli, R. M., Farmer, T. B., and Gile, J. Molecular imaging of biological samples:âĂĽ localization of peptides and proteins using maldi-tof ms. *Analytical Chemistry 69*, 23 (1997), 4751–4760. PMID: 9406525.

[29] Cassese, A., Ellis, S. R., Ogrinc PotoÄĐnik, N., Burgermeister, E., Ebert, M., Walch, A., van den Maagdenberg, A. M. J. M., McDonnell, L. A., Heeren, R. M. A., and Balluff, B. Spatial autocorrelation in mass spectrometry imaging. *Analytical Chemistry 88*, 11 (2016), 5871–5878. PMID: 27180608.

[30] Castleman, K. R. *Digital Image Processing*. Prentice Hall Press, Upper Saddle River, NJ, USA, 1996.

[31] Chen, H., Talaty, N. N., Takáts, Z., and Cooks, R. G. Desorption electrospray ionization mass spectrometry for high-throughput analysis of pharmaceutical samples in the ambient environment. *Analytical chemistry 77*, 21 (2005), 6915–6927.

[32] Chughtai, S., Chughtai, K., Cillero-Pastor, B., Kiss, A., Agrawal, P., MacAleese, L., and Heeren, R. M. A multimodal mass spectrometry imaging approach for the study of musculoskeletal tissues. *International Journal of Mass Spectrometry 325* (2012), 150–160.

[33] Cicerone, M. T., and Camp, C. H. Histological coherent raman imaging: a prognostic review. *Analyst 143*, 1 (2018), 33–59.

[34] Claude, E., Jones, E. A., and Pringle, S. D. Desi mass spectrometry imaging (msi). In *Imaging Mass Spectrometry*. Springer, 2017, pp. 65–75.

[35] Clish, C. B. Metabolomics: an emerging but powerful tool for precision medicine. *Molecular Case Studies 1*, 1 (2015), a000588.

[36] Cotter, R. J. *Time-of-flight mass spectrometry*. ACS Publications, 1993.

[37] Council, N. R., et al. *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*. National Academies Press, 2011.

[38] da Silva, R. R., Dorrestein, P. C., and Quinn, R. A. Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences 112*, 41 (2015), 12549–12550.

[39] Deininger, S.-O., Becker, M., and Suckau, D. Tutorial: Multivariate statistical treatment of imaging data for clinical biomarker discovery. *Methods in molecular biology (Clifton, N.J.) 656* (07 2010), 385–403.

[40] Diaz-Papkovich, A., Anderson-Trocme, L., and Gravel, S. Revealing multi-scale population structure in large cohorts. *bioRxiv* (2019), 423632.

[41] Dong, W., Moses, C., and Li, K. Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th*

International Conference on World Wide Web (New York, NY, USA, 2011), WWW '11, ACM, pp. 577–586.

[42] Eming, S. A., Wynn, T. A., and Martin, P. Inflammation and metabolism in tissue repair and regeneration. *Science 356*, 6342 (2017), 1026–1030.

[43] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (1996), KDD'96, AAAI Press, pp. 226–231.

[44] Fonville, J. M., Carter, C. L., Pizarro, L., Steven, R. T., Palmer, A. D., Griffiths, R. L., Lalor, P. F., Lindon, J. C., Nicholson, J. K., Holmes, E., and Bunch, J. Hyperspectral visualization of mass spectrometry imaging data. *Analytical Chemistry 85*, 3 (2013), 1415–1423. PMID: 23249247.

[45] Friia, M., Legros, V., Tortajada, J., and Buchmann, W. Desorption electrospray ionization-orbitrap mass spectrometry of synthetic polymers and copolymers. *Journal of mass spectrometry 47*, 8 (2012), 1023–1033.

[46] F.R.S., K. P. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2*, 11 (1901), 559–572.

[47] Gaber, T., Strehl, C., and Buttgereit, F. Metabolic regulation of inflammation. *Nature Reviews Rheumatology 13*, 5 (2017), 267–279.

[48] Gonzalez, R. C., and Woods, R. E. *Digital image processing*. Prentice Hall, Upper Saddle River, N.J., 2008.

[49] Gorzolka, K., Kölling, J., Nattkemper, T. W., and Niehaus, K. Spatio-temporal metabolite profiling of the barley germination process by maldi ms imaging. *PLoS One 11*, 3 (2016).

[50] Greaves, M., and Maley, C. C. Clonal evolution in cancer. *Nature 481*, 7381 (2012), 306–313.

[51] Griffith, D. A., Wong, D. W. S., and Whitfield, T. Exploring relationships between the global and regional measures of spatial autocorrelation. *Journal of Regional Science 43*, 4, 683–710.

[52] Guerquin-Kern, J.-L., Wu, T.-D., Quintana, C., and Croisy, A. Progress in analytical imaging of the cell by dynamic secondary ion mass spectrometry (sims microscopy). *Biochimica et Biophysica Acta (BBA)-General Subjects 1724*, 3 (2005), 228–238.

[53] Gurdak, E., Green, F. M., Rakowska, P. D., Seah, M. P., Salter, T. L., and Gilmore, I. S. Vamas interlaboratory study for desorption electrospray ionization mass spectrometry (desi ms) intensity repeatability and constancy. *Analytical chemistry 86*, 19 (2014), 9603–9611.

[54] Hanselmann, M., Kirchner, M., Renard, B. Y., Amstalden, E. R., Glunde, K., Heeren, R. M. A., and Hamprecht, F. A. Concise representation of mass spectrometry images by probabilistic latent semantic analysis. *Analytical Chemistry 80*, 24 (2008), 9649–9658. PMID: 18989936.

[55] Harris, R. A., Leaptrot, K. L., May, J. C., and McLean, J. A. New frontiers in lipidomics analyses using structurally selective ion mobility-mass spectrometry. *TrAC Trends in Analytical Chemistry 116* (2019), 316–323.

[56] Heeren, R., Kleinnijenhuis, A., McDonnell, L., and Mize, T. A mini-review of mass spectrometry using high-performance fticr-ms methods. *Analytical and bioanalytical chemistry 378*, 4 (2004), 1048–1058.

[57] Heeren, R. M., Smith, D. F., Stauber, J., Kükrer-Kaletas, B., and MacAleese, L. Imaging mass spectrometry: hype or hope? *Journal of the American Society for Mass Spectrometry 20*, 6 (2011), 1006–1014.

[58] Hinz, C., Liggi, S., and Griffin, J. L. The potential of ion mobility mass spectrometry for high-throughput and high-resolution lipidomics. *Current opinion in chemical biology 42* (2018), 42–50.

[59] Hoffman, E. d., and Stroobant, V. Mass spectrometry: principles and applications. *West Sussex: John Wiley & Sons, Bruxellas, Bélgica 1*, 2 (2007), 85.

[60] Houjou, T., Yamatani, K., Nakanishi, H., Imagawa, M., Shimizu, T., and Taguchi, R. Rapid and selective identification of molecular species in phosphatidylcholine and sphingomyelin by conditional neutral loss scanning and ms3. *Rapid Communications in Mass Spectrometry 18*, 24 (2004), 3123–3130.

[61] Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M., and Graham Cooks, R. The orbitrap: a new mass spectrometer. *Journal of mass spectrometry 40*, 4 (2005), 430–443.

[62] Jones E, Oliphant E, P. P. Scipy: Open source scientific tools for python, 2001.

[63] Kaddi, C. D., and Wang, M. D. *Computational Methods for Mass Spectrometry Imaging: Challenges, Progress, and Opportunities*. Springer International Publishing, Cham, 2017, pp. 37–49.

[64] Kaddurah-Daouk, R., Kristal, B. S., Weinshilboum, R. M., et al. Metabolomics: a global biochemical approach to drug response and disease. *Annual review of pharmacology and toxicology 48*, 1 (2008), 653–683.

[65] Kertesz, V., and Van Berkel, G. J. Improved imaging resolution in desorption electrospray ionization mass spectrometry. *Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry 22*, 17 (2008), 2639–2644.

[66] Kobak, D., and Berens, P. The art of using t-sne for single-cell transcriptomics. *Nature Communications 10*, 1 (2019), 5416.

[67] Kohonen, T. The self-organizing map. *Proceedings of the IEEE 78*, 9 (1990), 1464–1480.

[68] Kölling, J., Langenkämper, D., Abouna, S., Khan, M., and Nattkemper, T. W. Whide - a web tool for visual data mining colocation patterns in multivariate bioimages. *Bioinformatics 28*, 8 (2012), 1143–1150.

[69] Kompauer, M., Heiles, S., and Spengler, B. Atmospheric pressure maldi mass spectrometry imaging of tissues and cells at 1.4-μm lateral resolution. *Nature methods 14*, 1 (2017), 90–96.

[70] Konicek, A. R., Lefman, J., and Szakal, C. Automated correlation and classification of secondary ion mass spectrometry images using ak-means cluster method. *Analyst 137*, 15 (2012), 3479–3487.

[71] Lefevre, R. Microanalyse par emission ionique secondaire d'un pyroxene jadeitique d'age alpin de la bande d'acceglio (alpes cottiennes italiennes). *Bulletin de la Société Géologique de France 7*, 3 (1974), 248–254.

[72] Levy, A. J., Oranzi, N. R., Ahmadireskety, A., Kemperman, R. H., Wei, M. S., and Yost, R. A. Recent progress in metabolomics using ion mobility-mass spectrometry. *TrAC Trends in Analytical Chemistry 116* (2019), 274–281.

[73] Lu, C., and Thompson, C. B. Metabolic regulation of epigenetics. *Cell metabolism 16*, 1 (2012), 9–17.

[74] Lu, W., Su, X., Klein, M. S., Lewis, I. A., Fiehn, O., and Rabinowitz, J. D. Metabolite measurement: pitfalls to avoid and practices to follow. *Annual review of biochemistry 86* (2017), 277–304.

[75] Ma, S., and Dai, Y. Principal component analysis based methods in bioinformatics studies. *Briefings in Bioinformatics 12*, 6 (2011), 714–722.

[76] Ma, S., and Dai, Y. Principal component analysis based methods in bioinformatics studies. *Briefings in Bioinformatics 12*, 6 (2011), 714–722.

[77] Maaten, L. v. d., and Hinton, G. Visualizing data using t-SNE. 2579–2605.

[78] Mahfouz, A., van de Giessen, M., van der Maaten, L., Huisman, S., Reinders, M., Hawrylycz, M. J., and Lelieveldt, B. P. Visualizing the spatial gene expression organization in the brain through non-linear similarity embeddings. *Methods 73* (2015), 79–89.

[79] Manicke, N. E., Nefliu, M., Wu, C., Woods, J. W., Reiser, V., Hendrickson, R. C., and Cooks, R. G. Imaging of lipids in atheroma by desorption electrospray ionization mass spectrometry. *Analytical chemistry 81*, 21 (2009), 8702–8707.

[80] Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., and Ferrari, R. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics 19*, 2 (11 2016), 286–302.

[81] McCune, B., Grace, L., and Urban, D. *Analysis of Ecological Communities*. MjM Software Design, Gleneden Beach, OR, 2002.

[82] McDonnell, L. A., and Heeren, R. M. Imaging mass spectrometry. *Mass spectrometry reviews 26*, 4 (2007), 606–643.

[83] McInnes, L., and Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints* (Feb. 2018).

[84] McInnes, L., Healy, J., and Astels, S. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software 2*, 11 (mar 2017).

[85] McInnes, L., Healy, J., and Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction.

[86] McNaught, A. D., Wilkinson, A., et al. *Compendium of chemical terminology*, vol. 1669. Blackwell Science Oxford, 1997.

[87] Minerva, L., Ceulemans, A., Baggerman, G., and Arckens, L. Maldi ms imaging as a tool for biomarker discovery: Methodological challenges in a clinical setting. *Proteomics. Clinical applications 6* (12 2012).

[88] Minerva, L., Clerens, S., Baggerman, G., and Arckens, L. Direct profiling and identification of peptide expression differences in the pancreas of control and ob/ob mice by imaging mass spectrometry. *PROTEOMICS 8*, 18 (2008), 3763–3774.

[89] Monge, M. E., Dodds, J. N., Baker, E. S., Edison, A. S., and Fernández, F. M. Challenges in identifying the dark molecules of life. *Annual Review of Analytical Chemistry 12* (2019), 177–199.

[90] Montenegro, F., and Indraccolo, S. Metabolism in the tumor microenvironment. In *Tumor Microenvironment*. Springer, 2020, pp. 1–11.

[91] Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., van den Elzen, A., Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G., and Krishnaswamy, S. Visualizing structure and transitions for biological data exploration.

[92] Nuñez, J. R., Anderton, C. R., and Renslow, R. S. Optimizing colormaps with consideration for color vision deficiency to enable accurate interpretation of scientific data. *PloS one 13*, 7 (2018), e0199239.

[93] NVIDIA. Open source gpu data science, rapids, 2019.

[94] Passarelli, M., and Winograd, N. Characterizing in situ glycerophospholipids with sims and maldi methodologies. *Surface and interface analysis 43*, 1-2 (2011), 269–271.

[95] Patti, G. J., Yanes, O., and Siuzdak, G. Metabolomics: the apogee of the omics trilogy. *Nature reviews Molecular cell biology 13*, 4 (2012), 263–269.

[96] Petras, D., Jarmusch, A. K., and Dorrestein, P. C. From single cells to our planetâĂŤrecent advances in using mass spectrometry for spatially resolved metabolomics. *Current Opinion in Chemical Biology 36* (2017), 24 – 31.

[97] Petyuk, V. A., Qian, W.-J., Hinault, C., Gritsenko, M. A., Singhal, M., Monroe, M. E., Camp, D. G., Kulkarni, R. N., and Smith, R. D. Characterization of the mouse pancreatic islet proteome and comparative analysis with other mouse tissues. *Journal of proteome research 7*, 8 (2008), 3114–3126.

[98] Plas, R. V. d., Yang, J., Spraggins, J., and Caprioli, R. M. Image fusion of mass spectrometry and microscopy: a multimodality paradigm for molecular tissue mapping. 366–372.

[99] Prideaux, B., Staab, D., and Stoeckli, M. Applications of maldi-msi to pharmaceutical research. In *Mass Spectrometry Imaging*. Springer, 2010, pp. 405–413.

[100] Prieto, N., Pawluczyk, O., Dugan, M. E. R., and Aalhus, J. L. A review of the principles and applications of near-infrared spectroscopy to characterize meat, fat, and meat products. *Applied spectroscopy 71*, 7 (2017), 1403–1426.

[101] Qi, Y., and O'Connor, P. B. Data processing in fourier transform ion cyclotron resonance mass spectrometry. *Mass spectrometry reviews 33*, 5 (2014), 333–352.

[102] Race, A. M., Rae, A., Vorng, J.-L., Havelund, R., Dexter, A., Kumar, N., Steven, R. T., Passarelli, M. K., Tyler, B. J., Bunch, J., et al. Correlative hyperspectral imaging using a dimensionality-reduction-based image fusion method. *Analytical Chemistry 92*, 16 (2020), 10979–10988.

[103] Race, A. M., Steven, R. T., Palmer, A. D., Styles, I. B., and Bunch, J. Memory efficient principal component analysis for the dimensionality reduction of large mass spectrometry imaging data sets. *Analytical Chemistry 85*, 6 (2013), 3071–3078. PMID: 23394348.

[104] Reich, G. Near-infrared spectroscopy and imaging: basic principles and pharmaceutical applications. *Advanced drug delivery reviews 57*, 8 (2005), 1109–1143.

[105] Römpp, A., and Spengler, B. Mass spectrometry imaging with high resolution in mass and space. *Histochemistry and Cell Biology 139*, 6 (Jun 2013), 759–783.

[106] Rubakhin, S. S., and Sweedler, J. V. A mass spectrometry primer for mass spectrometry imaging. In *Mass Spectrometry Imaging*. Springer, 2010, pp. 21–49.

[107] RuÌĹbel, O., Greiner, A., Cholia, S., Louie, K., Bethel, E. W., Northen, T. R., and Bowen, B. P. Openmsi: a high-performance web-based platform for mass spectrometry imaging. *Analytical chemistry 85*, 21 (2013), 10354–10361.

[108] Sans, M., Feider, C. L., and Eberlin, L. S. Advances in mass spectrometry imaging coupled to ion mobility spectrometry for enhanced imaging of biological tissues. *Current opinion in chemical biology 42* (2018), 138–146.

[109] Schramm, T., Hester, Z., Klinkert, I., Both, J.-P., Heeren, R. M., Brunelle, A., Laprévote, O., Desbenoit, N., Robbe, M.-F., Stoeckli, M., et al. imzmlâĂŤa common data format for the flexible exchange and processing of mass spectrometry imaging data. *Journal of proteomics 75*, 16 (2012), 5106–5110.

[110] Ščupáková, K., Dewez, F., Walch, A. K., Heeren, R. M., and Balluff, B. Morphometric cell classification for single-cell maldi-msi. *Angewandte Chemie*.

[111] Shimma, S., and Sugiura, Y. Effective sample preparations in imaging mass spectrometry. *Mass Spectrometry 3*, Special_Issue (2014), S0029–S0029.

[112] Siegel, T. P., Hamm, G., Bunch, J., Cappell, J., Fletcher, J. S., and Schwamborn, K. Mass spectrometry imaging and integration with other imaging modalities for greater molecular understanding of biological tissues. *Molecular Imaging and Biology 20*, 6 (2018), 888–901.

[113] 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. In *Proceedings of the 8th IEEE International Conference on Bioinformatics and Bioengineering, BIBE 2008, October 8-10, 2008, Athens, Greece* (2008), pp. 1–6.

[114] Smets, T., De Keyser, T., Tousseyn, T., Waelkens, E., and De Moor, B. Correspondence-aware manifold learning for microscopic and spatial omics imaging: A novel data fusion method bringing mass spectrometry imaging to a cellular resolution. *Analytical Chemistry 93*, 7 (2021), 3452–3460.

[115] 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* (2019).

[116] 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 (2020), 5240–5248.

[117] Spratlin, J. L., Serkova, N. J., and Eckhardt, S. G. Clinical applications of metabolomics in oncology: a review. *Clinical cancer research 15*, 2 (2009), 431–440.

[118] Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J., and Frisén, J. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science 353*, 6294 (2016), 78–82.

[119] Stewart, S., Priore, R. J., Nelson, M. P., and Treado, P. J. Raman imaging. *Annual Review of Analytical Chemistry 5* (2012), 337–360.

[120] Takáts, Z., Wiseman, J. M., Gologan, B., and Cooks, R. G. Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. *Science 306*, 5695 (2004), 471–473.

[121] Takats, Z., Wiseman, J. M., Ifa, D. R., and Cooks, R. G. Desorption electrospray ionization (desi) analysis of intact proteins/oligopeptides. *Cold Spring Harbor Protocols 2008*, 4 (2008), pdb–prot4992.

[122] Takats, Z., Wiseman, J. M., Ifa, D. R., and Cooks, R. G. Desorption electrospray ionization (desi) analysis of tryptic digests/peptides. *Cold Spring Harbor Protocols 2008*, 4 (2008), pdb–prot4993.

[123] 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 2016, Athens, Greece, December 6-9, 2016* (2016), pp. 1–7.

[124] Tilg, H., Zmora, N., Adolph, T. E., and Elinav, E. The intestinal microbiota fuelling metabolic inflammation. *Nature Reviews Immunology 20*, 1 (2020), 40–54.

[125] Trindade, G., Abel, M.-L., and Watts, J. Non-negative matrix factorisation of large mass spectrometry datasets. *Chemometrics and Intelligent Laboratory Systems 163* (March 2017), 76–85.

[126] Turner, N. C., and Reis-Filho, J. S. Genetic heterogeneity and cancer drug resistance. *The lancet oncology 13*, 4 (2012), e178–e185.

[127] van der Maaten, L. Barnes-hut-sne. *CoRR abs/1301.3342* (2013).

[128] van der Maaten, L., and Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research 9* (2008), 2579–2605.

[129] van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T., and the scikit-image contributors. scikit-image: image processing in Python. *PeerJ 2* (6 2014), e453.

[130] Venna, J., and Kaski, S. Neighborhood preservation in nonlinear projection methods: An experimental study. In *Artificial Neural Networks — ICANN 2001* (Berlin, Heidelberg, 2001), G. Dorffner, H. Bischof, and K. Hornik, Eds., Springer Berlin Heidelberg, pp. 485–491.

[131] 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.

[132] Vogt, S., Löffler, K., Dinkelacker, A. G., Bader, B., Autenrieth, I., Peter, S. M., and Liese, J. Fourier-transform infrared (ftir) spectroscopy for typing of clinical enterobacter cloacae complex isolates. *Frontiers in microbiology 10* (2019), 2582.

[133] 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 (2011), 209–222.

[134] Winderbaum, L. J., Koch, I., Gustafsson, O. J. R., Meding, S., and Hoffmann, P. Feature extraction for proteomics imaging mass spectrometry data. *Ann. Appl. Stat. 9*, 4 (12 2015), 1973–1996.

[135] Wu, J. M., Halushka, M. K., and Argani, P. Intratumoral heterogeneity of her-2 gene amplification and protein overexpression in breast cancer. *Human pathology 41*, 6 (2010), 914–917.

[136] Yang, J., and Caprioli, R. M. Matrix sublimation/recrystallization for imaging proteins by mass spectrometry at high spatial resolution. *Analytical chemistry 83*, 14 (2011), 5728–5734.

[137] Zeggini, E., Gloyn, A. L., Barton, A. C., and Wain, L. V. Translational genomics and precision medicine: Moving from the lab to the clinic. *Science 365*, 6460 (2019), 1409–1413.

[138] Zhang, Y., Buchberger, A., Muthuvel, G., and Li, L. Expression and distribution of neuropeptides in the nervous system of the crab carcinus maenas and their roles in environmental stress. *Proteomics 15*, 23-24 (2015), 3969–3979.

[139] Zhou, Z., Shen, X., Tu, J., and Zhu, Z.-J. Large-scale prediction of collision cross-section values for metabolites in ion mobility-mass spectrometry. *Analytical chemistry 88*, 22 (2016), 11084–11091.

[140] Zhou, Z., Tu, J., Xiong, X., Shen, X., and Zhu, Z.-J. Lipidccs: prediction of collision cross-section values for lipids with high precision to support ion mobility–mass spectrometry-based lipidomics. *Analytical chemistry 89*, 17 (2017), 9559–9566.

[141] Zubarev, R. A., and Makarov, A. Orbitrap mass spectrometry, 2013.

[142] Zuiderveld, K. Graphics gems iv. Academic Press Professional, Inc., San Diego, CA, USA, 1994, ch. Contrast Limited Adaptive Histogram Equalization, pp. 474–485.

List of publications

C.1 International journal publications

Smets, Tina; Verbeeck, Nico; Claesen, Marc; Asperger, Arndt; Griffioen, Gerard; Tousseyn, Thomas; Waelput, Wim; Waelkens, Etienne; De Moor, Bart, "Evaluation of distance metrics and spatial autocorrelation in uniform manifold approximation and projection applied to mass spectrometry imaging data", *Analytical chemistry*, 91, 9, 5706-5714, 2019.

Smets, Tina; Waelkens, Etienne; De Moor, Bart, "Prioritization of m/z-values in mass spectrometry imaging profiles obtained using uniform manifold approximation and projection for dimensionality reduction", *Analytical chemistry*, 92, 7, 5240-5248, 2020.

Smets, Tina; De Keyser, Tom; Tousseyn, Thomas; Waelkens, Etienne; De Moor, Bart, "Correspondence-aware manifold learning for microscopic and spatial omics imaging: a novel data fusion method bringing MSI to a cellular resolution", *Analytical chemistry*, 93, 7, 3452-3460,2021.

Nijs, Melanie; Smets, Tina; Waelkens, Etienne; De Moor, Bart; "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*, e9181, 2021.

Sukhai, Mahadeo A.; Misyura, Maksym; Thomas, Mariam; Garg, Swati; Zhang, Tong; Stickle, Natalie; Virtanen, Carl; Bedard, Philippe L; Siu, Lillian L; Smets, Tina; Gert Thijs; Steven Van Vooren; Suzanne Kamel-Reid; Tracy L. Stockley; "Somatic tumor variant filtration strategies to optimize tumor-only molecular profiling using targeted next-generation sequencing panels", *The Journal of Molecular Diagnostics*, 21, 2, 261-273, 2019.

Luyten, Walter; Antal, Peter; Braeckman, Bart P; Bundy, Jake; Cirulli, Francesca; Fang-Yen, Christopher; Fuellen, Georg; Leroi, Armand; Liu, Qingfei; Martorell, Patricia; Andres Metspalu, Markus Perola, Michael Ristow, Nadine Saul, Liliane Schoofs, Karsten Siems, Liesbet Temmerman, Tina Smets, Alicja Wolk, Suresh IS Rattan, "Ageing with elegans: a research proposal to map healthspan pathways", *Biogerontology*, 17, 4, 771-782, 2016.

C.2 Posters and presentations at International Conferences

31 March 2019 - 2 April 2019, 4th NVMS-BSMS Conference on Mass Spectrometry, Kerkrade, The Netherlands, Smets, Tina; Verbeeck, N; Claesen, M; De Moor, B; Waelkens, E; *"Noise injection in MSI data: ranking features through the use of t-SNE"*

3 June 2018 - 7 June 2019, *ASMS Conference on Mass Spectrometry and Allied Topics*, San Diego, United States of America, Smets, Tina; Verbeeck, N; Claesen, M; De Moor, B; Waelkens, E; *"Evaluating robustness and feature extraction in t-SNE for MSI data using noise injection"*



FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
ESAT - STADIUS CENTER FOR DYNAMICAL SYSTEMS, SIGNAL PROCESSING, AND DATA ANALYTICS
Kasteelpark Arenberg 10 box 2446
B-3001 Leuven
tina.smets@esat.kuleuven.be
https://www.esat.kuleuven.be/stadius