Datamining of Imaging Mass Spectrometry Data for Biomedical Tissue Exploration

Nico VERBEECK

Examination committee:
Prof. dr. ir. H. Hens, chair
Prof. dr. ir. B. De Moor, supervisor
Prof. dr. E. Waelkens, co-supervisor
Prof. dr. ir. J. Suykens
Prof. dr. ir. J. Vanderleyden
Prof. dr. R. Heeren
(Maastricht University, The Netherlands)
Prof. dr. ing. R. Van de Plas
(TU Delft, The Netherlands)
(Vanderbilt University, U.S.A.)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering Science.

December 2014

© 2014 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Nico Verbeeck, Kasteelpark Arenberg 10, box 2446, B-3001 Heverlee (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.

ISBN 978-94-6018-937-1

D/2014/7515/160

Dankwoord

Inmiddels vijf jaar geleden begon ik aan mijn doctoraat, gebeten door het topic bio-informatica en de mogelijkheid om mijn ingenieursopleiding op een biologisch onderwerp los te laten. Een doctoraat in de bio-informatica is een geweldige reis, waarbij je een heleboel mensen met allerlei verschillende achtergronden ontmoet en waarbij je immens veel bijleert zowel op wetenschappelijk als op persoonlijk vlak. Een doctoraat is een soort van emotionele rollercoaster met veel ups, maar ook veel downs, want onderzoek doen betekent soms dagen aan een stuk zwoegen en verschillende methoden uitproberen om een probleem op te lossen, zonder enig resultaat of zicht op een oplossing. Maar, wanneer je dan uiteindelijk de oplossing vindt, is het een enorm vervullende ervaring en is het het zwoegen helemaal waard.

Ik wil hieronder graag de mensen bedanken die deze reis hebben mogelijk gemaakt en die erbij waren om samen met mij de ups te vieren, maar er ook waren om mij door de downs te loodsen. Zonder hen had ik hier vandaag niet gestaan. Ik heb geprobeerd iedereen te vermelden, maar vergeef me moest ik toch nog iemand vergeten zijn.

Om te beginnen wil ik mijn promotor, professor Bart De Moor, bedanken voor de geweldige opportuniteit om mijn doctoraat onder hem te af te leggen. Barts gedrevenheid en ondernemingsgezindheid zijn werkelijk bewonderenswaardig en een bron van inspiratie. Ik herinner me een bepaalde meeting om 9u 's morgens waarbij hij net van de luchthaven kwam van een congres om dan vervolgens na onze meeting naar een volgende in Brussel te vertrekken, waarna ik me afvroeg waar hij die schijnbaar onuitputbare energie vandaan bleef halen. Ik kan alleen maar hopen dat er een beetje van die drive en dat enthousiasme naar mij zijn overgevloeid tijdens mijn doctoraat. Bart, bedankt om me steeds de vrijheid te geven om die onderwerpen te onderzoeken die mij fascineerden en om me steeds te motiveren met de gevleugelde woorden "Go for it!".

Ik wil ook heel speciaal professor Raf Van de Plas bedanken. Nog voor alles begon was Raf al volledig in de weer om me op weg te helpen met mijn doctoraat. Hij hielp me met het schrijven van mijn IWT voorstel en reed speciaal met mij naar Brussel om alles op tijd ingediend te krijgen. Toen al stond Raf voor mij in de bres en gedurende onze hele samenwerking heeft hij steeds het onderste uit de kan gehaald om mij te ondersteunen wanneer het nodig was. Hij is een enorme bron van kennis en creatieve inspiratie geweest voor het werk in deze thesis en ik heb heel veel van hem bijgeleerd op persoonlijk vlak. Raf, ik ben je immens dankbaar voor alles wat je al voor mij hebt gedaan en voor de geweldige tijden die we tot nu toe al gehad hebben tijdens onze samenwerking. Bedankt om van mijn verblijf in Nashville een fantastische en leerrijke ervaring te maken waar ik nog vele jaren met plezier aan zal terugdenken.

Een tweede persoon die ik speciaal wil bedanken is professor Etienne Waelkens. Etienne is werkelijk een fantastisch mens, steeds bereid om te helpen en nooit is er iets te veel gevraagd, zelfs wanneer sommige verzoeken mogelijk wat last-minute komen (ik probeer er aan te werken). Hij heeft mij enorm veel bijgeleerd over IMS zowel in het labo als gezellig op een terrasje op congres. Etienne, bedankt voor al je hulp doorheen mijn doctoraat en voor alle tijd die je in mij hebt geïnvesteerd. Bedankt voor je luisterend oor en de pep-talks wanneer ik er even door zat, en voor de vele wijze raad die je mij doorheen de jaren hebt gegeven.

Daarnaast wens ik de verschillende leden van de jury te bedanken voor hun waardevolle bijdragen aan deze thesis. Professor Johan Suykens, hartelijk dank voor de feedback inzake de methodologie van de thesis en de fijne ontmoetingen doorheen mijn doctoraat. Professor Jos Vanderleyden, dank u voor het grondige nalezen van het manuscript en de vele suggesties ter verbetering die u heeft aangebracht. Ik wil zeker ook professor Ron Heeren bedanken om als extern jurylid te willen optreden voor deze thesis. Het is een eer om een boegbeeld van IMS in mijn jury te mogen hebben, hartelijk dank voor de waardevolle input en voor de steeds fijne ontmoetingen op conferenties. Dank ook aan professor Hugo Hens om als voorzitter van deze doctoraatsjury te willen optreden.

Dit doctoraat zou niet hetzelfde geweest zijn zonder de geweldige set collega's die ik op Esat heb rondlopen en waarvan er verschillende doorheen de jaren van dit doctoraat goede vrienden zijn geworden. Alejandro, bedankt voor alle fijne filosofische en werk-gerelateerd gesprekken, de lunches en onze gezamenlijke "rise early"-actie (die weliswaar van korte duur was). Dusan, thanks for the valuable discussions and the much needed coffee-breaks from time to time. Who would have known we'd be here when we started our master in AI together. Yousef, thank you for the deep and sometimes very animated discussions on IMS, and for being such great company at conventions. Verder wil ik mijn eiland-buddy Babis, Raf W., Ryo, Marc, Peter K., Leo, Toni en Arnaud bedanken voor de

steeds inspirerende gesprekken tijdens de lunch. Ook een extra dikke merci aan Arnaud voor alle hulp bij het regelen van de praktische zaken op het einde van dit doctoraat. Ook wil ik Inge bedanken voor de fijne samenwerking tijdens het proposal-schrijven, en Olivier Gevaert om mij te introduceren in de wondere wereld van de bio-informatica. Voorts wil ik professor Yves Moreau en professor Jan Aerts en alle andere collega's van de Bioi-groep bedanken, Peter R., Amin, Lieven T., Pooya, Gorana, Jaak, Griet, Sarah, Tunde, Minta, Maira, Adam en Xian, jullie zijn stuk voor stuk geweldige mensen.

De collega's van het vierde verdiep, mijn "bakermat", en het vijfde verdiep, kan ik hier natuurlijk niet weglaten. Bedankt, professor Lieven De Lathauwer voor de input inzake matrix decompositions en Kim, Tom, Maarten B., Mauricio, Philippe, Antoine, Kris, Carlos, Fabian, Marco, Rocco en Niels voor de fijne werksfeer. Maarten en Liesbeth, bedankt voor de IT-ondersteuning en de fijne eiland-tijd in de begindagen van mijn doctoraat. Ook de mensen van BIOMED, waaronder Devy (hierbij nog eens een officiële sorry voor de barbecue – wink,wink), Bori, Kirsten, Laure, Alexander, Katrien, Ivan, Ninah, Carolina, Steven en professor Sabine Van Huffel wens ik te bedanken, alsoof professor Joos Vandewalle, Pieter en Bruno voor de steeds aangename samenwerking voor de oefenzittingen van het vak Systeemtheorie. Ik wil ook Ida, John, Wim, Elsy, Ilse en Mimi danken voor de geweldige hulp die ze steeds bieden om de administratieve last tot een minimum te beperken.

I would also like to thank professor Richard Caprioli for giving me the opportunity to visit the MSRC lab at Vanderbilt University. I learned a lot during this time, and it was great to spend time in such an inspiring research environment. Also thank you Dave, Junhai and professor Kevin Schey for providing us with the superb IMS data used in this thesis. An extra thank you goes to Dave for the delicious English-style Thanksgiving dinner and to him and Jeff for the great times we had during the stay and at conferences.

Ik ben ook het IWT immens dankbaar voor het financieren van mijn onderzoek gedurende de voorbije 4 jaar en voor de steeds vlotte communicatie die ik met hen heb gehad. Ook wil ik het FWO bedanken om mijn onderzoeksverblijf aan de universiteit van Vanderbilt te financieren, zonder hen had dit niet mogelijk geweest. Ook wil ik graag de mensen op Gasthuisberg bedanken, met name professor Thomas D'Hooghe, Amelie, Isabelle, Alexandra, professor Rita Derua en Antoine voor de fijne samenwerking en het aanleveren van de data gebruikt doorheen mijn doctoraat.

Jeroen, Bart, Wim, Kim, Leo en Sharon, Thomas en Joke, Dimitri, Katrien, Thijs, bedankt aan jullie allen voor de amusante afleiding en het luisterend oor wanneer ik het nodig had. Speciaal ook bedankt aan Steve, die de laatste dagen

van het schrijven van deze thesis van dichtbij heeft mogen meemaken – sorry, slaaptekort doet rare dingen met een mens.

Ik wil ook mijn familie en schoonfamilie bedanken voor hun steun en interesse, hopelijk wordt het vandaag min of meer duidelijk waar ik de voorbije 5 jaar mee bezig ben geweest. Speciaal wil ik hier mijn peter Alain, tante Cocole en mijn neef Chris bedanken voor al hun steun en de belangrijke rol die zij al spelen in mijn leven sinds mijn jonge jaren. Ook wil ik mijn lieve oma en metie danken voor hun steun en de positieve invloed die zij altijd hebben gehad op mijn leven. Natuurlijk kunnen ook mijn zussie, Molly en mijn schoonbroer Roeland niet ontbreken, bedankt voor al jullie hulp en immense gastvrijheid. Ook wil ik de jongste aanwinst van onze familie, mijn nichtje Ella-Marie, danken voor haar bijdrage in onnoemelijke schattigheid.

Heel in het bijzonder wil ik mijn ouders bedanken. Mams, paps, bedankt, door jullie onvoorwaardelijke steun en liefde hebben jullie ervoor gezorgd dat ik hier vandaag sta. Bedankt om mij steeds de vrijheid te geven om te zijn wie ik wou zijn, zonder mij ergens naartoe te dwingen. Ondanks de tegenslagen die jullie zelf hebben gekend, hebben jullie steeds gezorgd voor een geweldige thuis waarvoor ik jullie voor altijd dankbaar ben. Jullie hebben me steeds gemotiveerd om door te zetten, ook als het moeilijk was, en ik had me geen betere ouders kunnen wensen.

Als allerlaatste is er een speciaal plaatsje gereserveerd om mijn vriendinnetje, Gerlinde, te bedanken. Gerlindeke, bedankt dat jij steeds in mij bent blijven geloven, zelfs wanneer ik het zelf niet meer zag zitten. Bedankt voor alle keren dat je mij er bovenop hebt gepraat, en me terug een boost hebt gegeven om er volledig voor te gaan. Bedankt om mij steeds te doen lachen met je geweldige humor en voor de verrassingspakketjes die je, elke keer ik op conferentie vertrek, in mijn valies verstopt. Bedankt voor alles.

Nico Verbeeck
Leuven, december 2014.

Abstract

Imaging Mass Spectrometry (IMS) is a powerful molecular imaging technology that enables the direct analysis of the spatial distribution of biomolecules in a tissue section. This technology allows for the monitoring of thousands of molecules throughout the tissue in a single experiment, and enables comparison of the biomolecular content between various areas in the tissue, ranging from low-mass metabolites and lipids to high-mass biomacromolecules such as proteins. IMS requires no hypothesis on possible target molecules beforehand as no labeling of target molecules is involved at any point, making IMS an immensely valuable technology for explorative research. An IMS analysis from a single examined tissue slice, however, also leads to file sizes of several gigabytes and more recently even terabytes of complex and high dimensional data. Manual exploration of these data is becoming increasingly infeasible and in order to harness the full potential of these data, there is a strong need for computational methods that can extract the valuable information that is captured in these mountains of data. It is the development of these computational methods that is the focus of this work.

In this thesis we have concentrated on three computational challenges in particular: feature selection and dimensionality reduction for IMS data, differential analysis of IMS data from multiple experiments, and data fusion of IMS data with additional data sources.

In the feature selection and dimensionality reduction for IMS data, we use the Discrete Wavelet Transform (DWT) as a dimensionality reduction tool that can achieve considerable reduction of the original data size, while keeping all the information in the IMS spectra intact. We accomplished this by discarding only those variables that contain noise, and which take up unnecessary storage space and hamper computational analysis. We were able to further improve on previous work in this field by incorporating spatial information, a key feature of IMS data, into the selection procedure that differentiates between noise and true biological signal. We demonstrate two different ways of incorporating this

spatial information, namely global and local spatial information, and show that both methods can improve feature selection over traditional methods. The new methods show a better retention of features that contain true signal and can provide improved data compression and dimensionality reduction by discarding more features that are due to noise.

In the second part of this work we focus on the challenge of computationally comparing IMS data from multiple tissue slices. Performing such a comparison is far from trivial due to the large heterogeneity and the large size of IMS datasets. We therefore introduce a method called Group Independent Component Analysis (GICA) to perform this comparison. This method allows for an in-depth comparison of a large number of IMS datasets without the need for peak picking, extracting features from the full profile IMS data at the level of an individual tissue sample. By performing the pattern extraction at the level of a single tissue, GICA emphasizes small, tissue-specific features which could be lost in a simultaneous analysis over all tissues. Correct operation of the algorithm is demonstrated using an artificially generated dataset containing 3 different disease classes. This case study shows that the algorithm allows for near-perfect retrieval of the differential components between tissues in the correct locations. Subsequently, this technique is applied to two real-life IMS datasets, a case study of amyotrophic lateral sclerosis (ALS) in mouse spine and one of retinal degeneration disease in mouse retina. Application of the GICA algorithm to these datasets results in retrieval of multiple differential biomolecular patterns between the different tissue classes, which are expressed in biologically relevant areas for the diseases under study.

The third and final topic is the data fusion of IMS data with additional data sources. Combining IMS data with external data sources can create richer datasets and enable deeper insights into the IMS data. We could demonstrate this by linking IMS data directly to an anatomical atlas, namely the Allen Mouse Brain Atlas (AMBA). This integrates anatomical insights, an important resource that many human interpretations will naturally rely upon, with the IMS data. We spatially map IMS data to the atlas using non-rigid image registration techniques and use the established link to investigate spatial correlations between ion images and the anatomical structures of the atlas, gaining basic insights into the relationships between the two. Subsequently we move beyond simple correlation and use the atlas information to provide automated anatomical interpretation of the ion images in the IMS data. This challenge is solved as a convex optimization problem that deconstructs ion distributions as combinations of known anatomical structures. We demonstrate that establishing a link between an IMS experiment and an anatomical atlas can serve as an important accelerator both for human and machine-guided exploration of IMS experiments.

Beknopte samenvatting

Massaspectrale beeldvorming of Imaging Mass Spectrometry (IMS) is een krachtige moleculaire beeldvormingstechnologie die het mogelijk maakt om de spatiale verdeling van biomoleculen in een weefselsnede te analyseren. Met deze technologie kunnen duizenden moleculen gevolgd worden doorheen het weefsel in één enkel experiment, en kan de biomoleculaire inhoud van verschillende zones in het weefsel vergeleken worden op het niveau van metabolieten, lipiden en proteïnen. IMS vereist geen voorafgaande hypothese over een mogelijk doelwitmolecule, wat dit tot een immens waardevolle techniek maakt voor exploratief onderzoek. Een IMS analyse leidt echter tot grote hoeveelheden complexe en hoog-dimensionele data voor de analyse van een enkele weefselsnede, met bestandgroottes van gigabytes en recentelijk zelfs terabytes aan data. Dit maakt dat handmatig bestuderen van deze data zo goed als onuitvoerbaar wordt, zeker in de context van een exploratieve studie. Daarom is er een grote nood aan rekenkundige methoden om de kostbare biomoleculaire informatie, die vervat zit in deze gigantische hoeveelheid data, te ontginnen.

Omwille van de complexiteit van IMS data zijn er verscheidene rekenkundige problemen die moeten opgelost worden, en in dit werk zullen we ons richten op drie hiervan in het bijzonder: kenmerk-selectie en dimensionaliteitsreductie, differentiële analyse van IMS data afkomstig van verschillende experimenten en data fusie van IMS data met bijkomende databronnen.

In ons werk over kenmerk-selectie en dimensionaliteitsreductie wordt de toepassing van de Discrete Wavelet Transformatie (DWT) bestudeerd voor dimensionaliteits- en datagrootte-reductie van IMS data. Door de DWT te gebruiken kunnen we alle informatie in het massaspectrum behouden, en toch een sterke vermindering van de oorspronkelijke data-grootte verkrijgen. Dit doen we door alleen die delen van de data te verwijderen die ruis bevatten en daardoor onnodige opslagruimte innemen en die bovendien de rekenkundige analyse belemmeren. Om de juist selectie te maken tussen welke delen van de data veroorzaakt worden door ruis en welke door reëel biologisch signaal, wordt

een essentieel element van IMS data, namelijk spatiale informatie, betrokken in het selectieproces. Twee manieren om deze spatiale informatie te incorporeren worden onderzocht, namelijk het gebruik van globale en locale spatiale informatie. In vergelijking met traditionele methoden, zijn de nieuw-ontworpen methoden beter in staat om die delen van de data te behouden die echt biologisch signaal bevatten, en leveren ze een betere datacompressie en dimensionaliteitsreductie door meer delen van de data te verwijderen.

In het tweede deel van dit werk richten we ons op de rekenkundige vergelijking van IMS data afkomstig van meerdere weefselsnedes. Het uitvoeren van zulke vergelijking is verre van triviaal omwille van de grote heterogeniteit en de grootte van de IMS datasets. We introduceren daarom een methode, genaamd Group Independent Component Analysis (GICA), die een gedetailleerde vergelijking van een groot aantal IMS datasets mogelijk maakt zonder dat "peak picken" noodzakelijk wordt. De correcte werking van het algoritme wordt aangetoond door middel van een artificiële dataset die drie verschillende weefselklasses bevat. Hierbij zien we dat het algoritme bijna perfect de patronen kan extraheren die de drie klasses van elkaar onderscheiden. Vervolgens passen we deze techniek toe op twee gemeten IMS datasets, namelijk een casus van Amyotrofe Laterale Sclerose (ALS) in muis-ruggengraat en een casus van retinale degeneratie in muis-retina. Toepassing van het GICA algoritme op deze datasets maakt het mogelijk om meerdere biomoleculaire patronen te extraheren die de verschillende weefselklasses van elkaar onderscheiden, en die tot uitdrukking komen in de biologisch-relevante regio's van het weefsel voor de bestudeerde ziektes.

Het derde en laatste topic dat we verkennen is de datafusie van IMS data met bijkomende databronnen. Door IMS data te combineren met externe databronnen wordt het mogelijk een dieper inzicht te verkrijgen in de IMS data. In dit werk tonen we dit aan door IMS data direct te koppelen met een anatomische atlas, namelijk de Allen Mouse Brain Atlas (AMBA). Dit maakt een directe, computationele toegang mogelijk tot een belangrijke informatiebron waar de interpretatie van IMS data vaak op zal berusten, namelijk anatomische informatie. We leggen de voorgestelde link door de IMS data spatiaal met de atlas te registreren met behulp van niet-rigiede beeldregistratietechnieken, en gebruiken de gelegde link vervolgens om spatiale correlaties tussen de ionbeelden, afkomstig uit IMS, en de anatomische structuren, afkomstig uit de atlas, te onderzoeken. Vervolgens gebruiken we de informatie in de altas om de ionbeelden van de IMS data automatisch anatomisch te interpreteren. We lossen deze uitdaging op door deze interpretatie te beschrijven als een convex optimalisatieprobleem dat de spatiale ion verdelingen vertaalt naar combinaties van de gekende anatomische structuren. We tonen aan dat een link leggen tussen een IMS experiment en een anatomische atlas kan dienen als een belangrijke katalysator voor de verkenning van IMS experimenten.

Abbreviations

Ac	Approximation Coefficients
ALS	Amyotrophic Lateral Sclerosis
ANOVA	Analysis of Variance
AMBA	Allen Mouse Brain Atlas
amu	Atomic mass unit
CWT	Continuous Wavelet Transform
alpha-CHCA	alpha-cyano-4-hydroxycinnamic acid
Da	Dalton
Dc	Detail Coefficient
DWT	Discrete Wavelet Transform
DESI	Desorption Electrospray Ionization
ESI	Electrospray Ionization
FFD	Free Form Deformation
fMRI	functional Magnetic Resonance Imaging
FT-ICR	Fourier Transform Ion Cyclotron Resonance
IC	Independent Component
ICA	Independent Component Analysis
IDWT	Inverse Discrete Wavelet Transform
IMS	Imaging Mass Spectrometry
ITO	Indium Tin Oxide
TOF	Time-of-Flight
GICA	Group Independent Component Analysis
HGP	Human Genome Project
LA-ICP	Laser Ablation - Inductively Coupled Plasma
LAESI	Laser Ablation Electrospray Ionization
LESA	Liquid Extraction Surface Analysis
MALDI	Matrix Assisted Laser Desorption/Ionization
MAD	Median of the Absolute Deviations
MCP	Microchannel Plates
MDL	Minimum Description Length

MIRT	Medical Image Registration Toolbox
MRS	Magnetic Resonance Spectroscopy
MS	Mass Spectrometry
MSI	Mass Spectrometry Imaging
m/z	Mass over charge
PC	Principal Component
PCA	Principal Component Analysis
PET	Positron Emission Tomography
RAM	Random-access Memory
RPE	Retinal Pigment Epithelium
SELDI	Surface Enhanced Laser Desorption/Ionization
SIMS	Secondary Ion Mass Spectrometry
SVD	Singular Value Decomposition
SWT	Stationary Wavelet Transform
TIC	Total Ion Current
WT	Wild Type

Contents

Abstract v

Abbreviations ix

Contents xi

List of Figures xv

List of Tables xix

1 Introduction 1

1.1 Mass spectrometry 2

1.2 Imaging Mass Spectrometry 2

1.3 Advantages of IMS 4

1.4 Types of IMS 5

1.5 Computational analysis of IMS data 8

2 Feature selection and dimensionality reduction for IMS data 13

2.1 Introduction 13

2.2 Continuous Wavelet Transform 16

2.2.1 Application of CWT to IMS Data 17

2.2.2 Redundancy in CWT 18
2.3 Discrete Wavelet Transform 20
2.3.1 Wavelet compression 23
2.4 Coefficient selection incorporating global spatial information . . 25
2.4.1 Data reconstruction 29
2.4.2 Coefficient selection using universal threshold 35
2.4.3 Incorporation of global spatial information: conclusions 43
2.5 Wavelet coefficient selection using localized spatial information 44
2.5.1 Workflow 45
2.5.2 Artificial dataset 47
2.5.3 Results 48
2.5.4 Incorporation of localized spatial information: conclusions 51
2.6 Conclusions & Future perspectives 52

3 Differential Analysis of IMS Experiments 55

3.1 Introduction 55
3.2 Methods 57
3.2.1 Sample-specific PCA 58
3.2.2 Group-wide PCA 61
3.2.3 Estimation of number of components 62
3.2.4 Independent component analysis (ICA) 64
3.2.5 Back-reconstruction 65
3.2.6 Discerning common from differential components 65
3.2.7 IMS Data 66
3.3 Results and Discussion 69
3.3.1 Results artificial dataset 69
3.3.2 Results mouse spine dataset 75

3.3.3 Results retina 82
3.4 Conclusions & Future work 90

4 Automated Anatomical Interpretation of Imaging Mass Spectrometry Data using Curated Atlases 93

4.1 Introduction 95
4.2 Methods 96
4.2.1 Anatomy Data 97
4.2.2 IMS Data 98
4.2.3 Registration of Data Sources 99
4.2.4 Correlation-based Querying 100
4.2.5 Automated Anatomical Interpretation 101
4.3 Results & Discussion 103
4.3.1 Results on Registration of the Data Sources 103
4.3.2 Results on Querying the Correlation between Anatomy and Ions 104
4.3.3 Results on Automated Anatomical Interpretation 107
4.4 Conclusions 110
4.5 Supplementary Information 111
4.5.1 IMS Data Acquisition and Microscopy Staining Details . 111
4.5.2 Registration Process Details 112
4.5.3 Additional Correlation-based Querying Examples 116
4.5.4 Additional Automated Anatomical Interpretation Examples 119
4.5.5 Sparsity in Automated Anatomical Interpretation 119
4.5.6 Automated Anatomical Interpretation and the Structure of Annotation Labels 124
4.5.7 Automated Anatomical Interpretation of Non-Conforming Ion Images 127

5 Conclusions & Future perspectives **129**

5.1 Conclusions 129

5.2 Future perspectives 131

5.2.1 Data size and dimensionality of IMS data 131

5.2.2 Data integration 132

Bibliography **135**

5.3 International Journal Publications 149

5.4 Oral Talks at National and International Conferences or Workshops 149

5.5 Posters at National and International Conferences or Workshops 150

List of Figures

1.1 Workflow of a typical MALDI-based IMS experiment. 4

1.2 Demonstration of the ion images acquired in a coronal section of mouse brain. 5

1.3 The data originating from an IMS experiment can be represented as a 3-Dimensional tensor, with 2 spatial dimensions and one m/z dimension. 10

2.1 The Mexican hat wavelet $\psi_{Mexicanhat}(t)$. 17

2.2 Example of CWT analysis on a section of a mass spectrum. . . 19

2.3 Results of the CWT analysis on a mass spectrum, using 10 levels of scaling. 20

2.4 Illustration of the DWT applied to a section of a mass spectrum. 21

2.5 DWT analysis on a single mass spectrum can be represented as a cascade of filters. 22

2.6 Decomposition of a single mass spectrum using 3 levels of DWT with the Sym8 wavelet. 24

2.7 DWT decomposition of multiple mass spectra. 27

2.8 PCA analysis results of the DWT level 1 detail coefficients. . . 30

2.9 PCA analysis results of the DWT level 2 detail coefficients. . . . 31

2.10 PCA analysis results of the DWT level 3 detail coefficients. . . 32

2.11 PCA analysis results of the DWT level 4 detail coefficients. . . 33

2.12 Comparison of reconstructed mass spectra after wavelet selection. 34

2.13 Comparison of reconstructed mass spectra after wavelet selection using the universal threshold. 38

2.14 PCA analysis results of the DWT level 2 detail coefficients after applying the universal threshold. 40

2.15 PCA analysis results of the DWT level 3 detail coefficients after applying the universal threshold. 41

2.16 PCA analysis results of the DWT level 4 detail coefficients after applying the universal threshold. 42

2.17 Overlay image of a PCA analysis which groups together pixels with similar biochemical content 44

2.18 Workflow of the wavelet coefficient selection using localized spatial information. 48

2.19 Construction of artificial dataset. 49

2.20 Comparison between the noisy spectrum, reconstructed spectrum after wavelet selection using local spatial information and ground truth spectrum without noise. 51

3.1 Workflow of the GICA algorithm. 59

3.2 Construction basis artificial dataset. 67

3.3 Construction artificial datasets. 68

3.4 Summary results GICA on artificial dataset. 72

3.5 Comparison of the components retrieved by GICA with the ground truth. 73

3.6 Comparison results GICA and PCA on artificial dataset. 74

3.7 Comparison of component selection for the mouse spine data . 77

3.8 Data reconstruction for component selection in the mouse spine samples. 78

3.9 Summary GICA results on the mouse spine samples 79

3.10 In-depth look at component IC1 in the mouse spine samples. . 80

3.11 In-depth look at component IC4 in the mouse spine samples. . . 81

3.12 Results of the PCA analysis on the mouse spine data. 83
3.13 Data reconstruction results for the mouse retina samples. . . . 84
3.14 Spatial expression of IC 34 in the retina samples. 86
3.15 Expression of IC9 in the retina samples. 87
3.16 Overview GICA results retina subset. 87
3.17 Expression of IC2 in the retina samples. 88
3.18 Expression of IC3 in the retina samples. 89

4.1 Registered microscopy and anatomy in the Allen Mouse Brain Atlas. 98
4.2 Workflow of the registration process. 100
4.3 Overview of registration results. 105
4.4 Example of an anatomical query and an ion query 106
4.5 Examples of automated anatomical interpretation. 109
4.6 Affine registration of m/z 18411 to the experimental histology. 113
4.7 The experiment histology has a relatively high deformation compared to the reference histology in the Allen Brain Atlas. The need for non-rigid registration is apparent. 115
4.8 Free Form Deformation mesh used for non-rigid registration . . 117
4.9 Results of the non-rigid registration process. 117
4.10 The correlation table displaying spatial correlation between the anatomical structure images and ion image of the peak picked IMS protein data. 118
4.11 Example of an anatomical and an ion query 120
4.12 Examples of positively correlating anatomical structure/ion image combinations. 121
4.13 Correlation table resulting from the spatial correlation analysis between the anatomical structure images and the ion images of the peak-picked IMS lipid data. 122
4.14 Examples of several positively correlating lipid ion images. . . . 122

4.15 Examples of automated anatomical interpretation applied to the protein-focused IMS case study. 123

4.16 Primary somatosensory area 125

4.17 Primary somatosensory area, nose (a radial sub-area of the "Primary somatosensory area" parent shown in the previous figure) 126

4.18 Primary somatosensory area, nose, layer 2/3 (a concentric sub-area within a radial sub-area of the "Primary somatosensory area") 126

List of Tables

2.1 Comparison of the compression factors and signal reconstruction. 35

2.2 Comparison of the compression factors and signal reconstruction using all three DWT approaches. 36

2.3 Comparison results peak picking. 50

Chapter 1

Introduction

In this first chapter we start by situating the biological context of this work. We then introduce imaging mass spectrometry (IMS) as a powerful molecular imaging technology for studying biological processes in tissue sections, and provide a short overview of the different technologies used in IMS. We go on to explain the strong need for computational methods to analyze the rich and complex data resulting from an IMS experiment. It is the development of these computational methods that will be the main focus of this work. We conclude the chapter by outlining the remainder of the thesis.

Once upon a time in the 1960's, the number of protein-encoding genes in humans was estimated to be around 2 million [1]. By the time the Human Genome Project (HGP) started in 1990, which set out to identify and map all genes in the human genome, this number was estimated to be around 100.000. At the end of the Human Genome Project, the number of genes was reassessed to be somewhere between 20.000 and 25.000 [2] and a recent study suggests we could only have as few as 19,000 protein-encoding genes [3], fewer than nematode worms. Meanwhile, the estimated total number of human proteins has trended upwards and is now generally believed to be somewhere between 250.000 and 1 million, with recent studies discovering new proteins to date [4]. From the foregoing numbers we can conlcude that, although the HGP and genomics in general are essential tools for gaining insight into biological processes, they only provide one piece of the big biological puzzle. A single gene can code for more than one and in some cases for up to hundreds of different proteins due to alternative splicings and post-translational modifications [5]. Multiple neurodegenerative diseases (Alzheimer's [6], Kennedy's disease [7]) and various cancers (prostate, breast cancer [7]) have been shown to be due to

post-translational modifications or other molecular processes that cannot be traced by studying the genome alone. For this reason, proteomics, and more recently metabolomics and lipidomics, have emerged as new and promising fields of study, revealing an even larger biological complexity than previously thought.

1.1 Mass spectrometry

One of the prime technologies to study the emerging fields of proteomics and lipidomics is mass spectrometry (MS). The introduction of soft ionization techniques such as matrix assisted laser desorption ionization (MALDI) [8] [9] and electrospray ionization (ESI) [10] around 20 years ago enabled MS to perform fast, sensitive and very specific analysis of larger biomolecules including (intact) proteins, peptides and lipids. This triggered an immense increase in relevance of MS in biological and medical research [11], and has made it a key technology in many laboratories.

In a typical MALDI TOF MS analysis, one of the most common forms of MS to date, the sample under study is mixed with an enery-absorbing chemical matrix which is typically a small organic compound that crystalizes on drying, hence the name matrix assisted laser desorption ionization. The mixed sample is then irradiated using a laser, generating ions from the sample, which are typically singly charged in MALDI. The matrix ensures that laser energy is gradually transferred to large proteins, enabling ionization without fragmenting the molecule. In TOF (Time-of-flight) MS, the formed ions are then accelerated using an electric field of known strength and are directed through a TOF-tube before hitting a detector. The time it takes for the ion to travel the distance of the TOF tube will be determined by its mass (small mass = faster) and its charge state (more charge = faster). Hence the term mass-to-charge ratio (m/z-ratio) is introduced. The generated mass spectrum plots the ion intensity (the number of ions hitting the detector) versus the m/z-ratio. In MALDI, the generated ions are often singly charged (charge $z = 1$), making their m/z-ration equal to their mass, which is expressed in atomic mass unit (amu) or Dalton (Da).

1.2 Imaging Mass Spectrometry

The importance of understanding the spatial context in which molecular changes take place is increasing in the study of biological and pathological processes in

living organisms [12]. The spatial distribution of biomolecules, such as proteins and lipids, throughout tissue often hold crucial clues towards determining the biological functions of these biomolecules. In a standard MS experiment the tissue under study is homogenized before analysis and thus this spatial context of the measured biomolecules is lost.

An emerging technology that can deliver such spatial information is MALDI imaging mass spectrometry (IMS), also called mass spectrometry imaging (MSI), which was introduced by Caprioli *et al.* in 1997 [13, 14]. IMS allows the direct mass spectrometric analysis of a tissue section and therefore allows the study of the spatial distribution and relative abundace of a whole range of molecules directly in the tissue. IMS has gained considerable momentum in recent years, primarily in the field of tissue biomarkers [15, 16] and drug delivery [17, 18], and has been successfully applied to tissues of various origin, including insect [19], mammalian [20], and human tissue [21–23].

Figure 1.1 shows an overview of a typical MALDI-based IMS experiment. First the tissue sample under study, in this case a mouse brain, is cut into sections (10 μm thick), which are mounted on a conducting, indium tin oxide (ITO) covered glass microscope slide. The energy-absorbing matrix solution is then deposited on the mounted tissue section, either by spray coating [24], use of a robotic spotter [24] or sublimation [25]. Then a virtual grid is defined over the tissue section. At each grid point, the tissue is irradiated with a MALDI laser, generating ions from the molecular content of the tissue at that location. The distance between the grid points determines the spatial resolution of the experiment. The ions are directed into the mass spectrometer for analysis, resulting in a mass spectrum reporting the molecular content at each grid point. By visualizing the ion intensity for a particular mass-over-charge value across the tissue, one generates a so-called ion image corresponding to that m/z value. This ion image can be overlayed over a standard microscopic image of the tissue, visualizing the position and the relative abundance of that particular molecule in the tissue. Figure 1.2 demonstrates this for a coronal section of mouse brain. (Bottom) We show the mean spectrum over all acquired mass spectra throughout the tissue. Each peak in the mass spectrum represents a biomolecular ion that is present in the tissue. The multitude of peaks observed immediately shows the richness in biomolecular species we can measure. (Top) Above the mass spectrum we show examples of ion images for several manually selected ions. These ion images were obtained by first generating ion images for the highest-abundant ions in the dataset, based on the mean spectrum. From this list of ion images we then selected several of them with distinct spatial distributions for illustration purporses. The presented ion images clearly show that some of these biomolecular ions are exclusively expressed in certain anatomical structures.

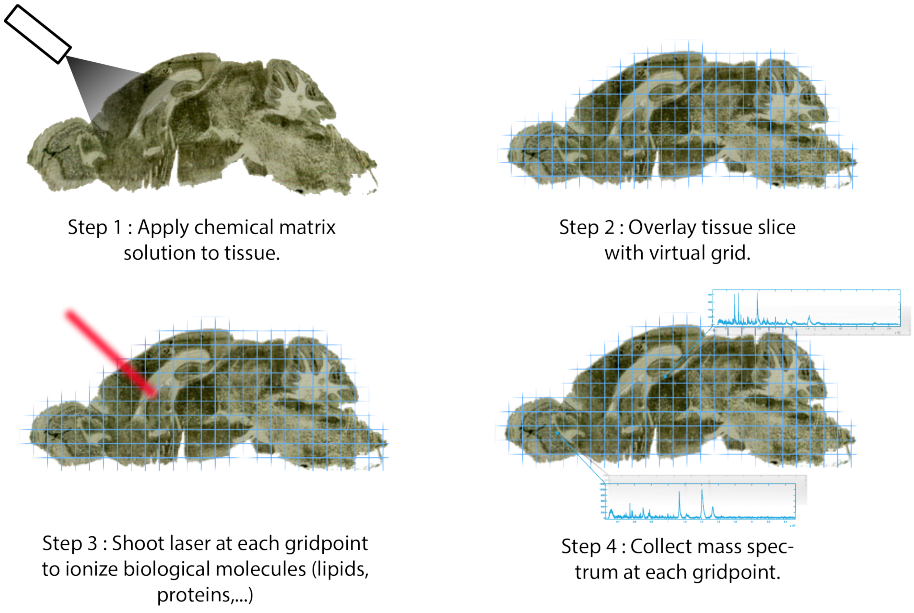

Figure 1.1: Workflow of a typical MALDI-based IMS experiment. Step 1: Using a microtome, the tissue sample is cut into sections (10 μm thick), which are mounted on a conducting glass microscope slide. A chemical matrix solution is deposited on the mounted tissue section. Step 2: A virtual grid is defined over the tissue section. Step 3: at each grid point the tissue is irradiated with a MALDI laser, generating ions from the (molecular) content of the tissue at that location. Step 4: The ions are directed into the mass spectrometer for analysis, resulting in a mass spectrum reporting the molecular content at each grid point.

1.3 Advantages of IMS

The main advantage of IMS is that in a single analysis, a broad range of molecules can be registered. The data allows for comparison of proteomic, peptidomic and metabolomic content between various areas in the tissue, and this without the need to hypothesize on possible target molecules beforehand as no labelling of target molecules is involved at any point. This characteristic makes IMS suitable for the traditional confirmatory role for which molecular imaging modalities such as immunostaining are used, but at the same time makes this an imaging technique that is of prime interest for exploratory experiments. The spatio-molecular information collected via IMS can then be linked to biological functions in the underlying tissue. This could enable us for example to discover proteins that are present in the core of a tumour versus those that

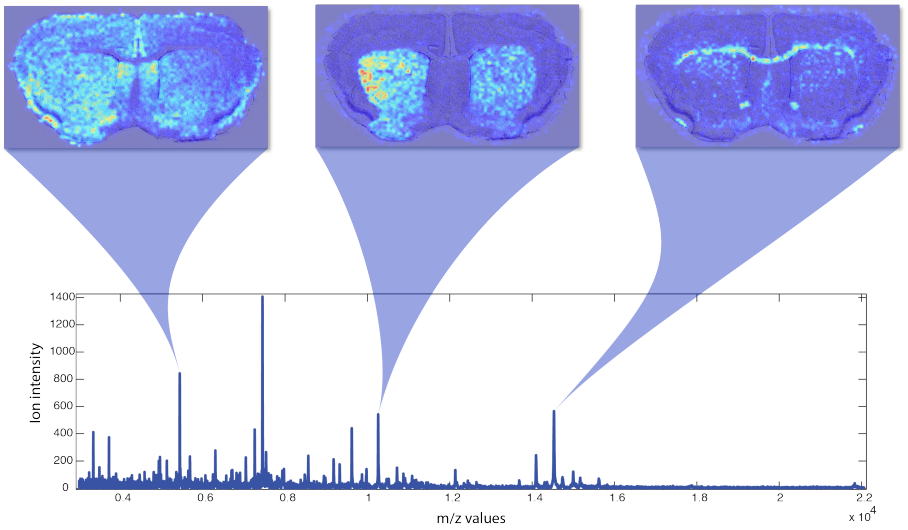

Figure 1.2: Demonstration of the ion images acquired in a coronal section of mouse brain. (Bottom) The mean spectrum over all acquired mass spectra throughout the tissue. Each peak in the mass spectrum represents a biomolecular ion that is present in the tissue. (Top) Ion images which show the spatial distribution and relative abundance for a particular ion.

are present at the growth front of a tumour and to compare to healthy tissue altogether.

A second advantage IMS offers, is the ability to monitor hundreds to thousands of biomolecules at every location in the tissue simultaneously in a single experiment. This property strongly differentiates IMS from immunohistochemical tags and radiolabels, which can only monitor a few molecular species at the same time in the same experiment. Also, these techniques require prior knowledge of the molecule of interest, which pre-empts the possibility of discovering new biomolecules.

1.4 Types of IMS

While so far we have focused on MALDI-based TOF IMS, many different types of IMS are possible. A mass spectrometer, which of course forms the basis of IMS, consists of three main components: the ion source, the mass analyzer and

the ion detector. For each of these components several alternatives exist, which can be combined together to perform the IMS experiment.

Ion source The ion source is the component responsible for ionizing the molecules in the sample. Once the molecules are ionized they are guided towards the mass analyzer. There are a variety of techniques that can be used to create the ions, and the choice of the ionization technique has a large impact on the spatial resolution of the experiment and the biomolecular species that can be measured. We will therefore shortly discuss the three most commonly used techniques for IMS.

- **MALDI** Matrix Assisted Laser Desorption/Ionization (MALDI) is the most used ionization technique in the analysis of tissue sections to date. As already mentioned, a chemical matrix is deposited on the sample. Most matrices are small organic acids with the essential property to absorb at the wavelength of the ionizing laser. Upon evaporation of the solvent, the matrix forms co-crystals with the analyte molecules, extracting the analyte molecules from the complex sample and isolating them in a sea of matrix [14]. These crystals are then irradiated using UV laser pulses, causing ionization and desorption. MALDI allows for the analysis of a very wide range of biomolecules such as metabolites, lipids, peptides and proteins in a mass range exceeding 100 kDa. MALDI currently remains the best method for studying intact proteins. Spatial resolutions depend on the analyte under study and the matrix deposition method, with common resolutions around 50-100 μm for proteins and 5-20 μm for lipids. However, new advances are making even better spatial resolutions possible, going down to sub-micron pixel sizes [26]. Despite its great merits, there are several disadvantages to using a matrix to enable ionization in IMS. The matrix deposition greatly influences the ionization of the analyte molecules, and an uneven matrix deposition may give rise to differences in measured ion intensities, despite the underlying tissue having the same biomolecular content. Additionally, the matrix deposition holds the risk that the analyte molecules migrate while the matrix is still liquid, causing spatial displacement of the analyte. These issues make a correct sample preparation a crucial step in the MALDI IMS experiment. The recent development of sublimation approaches to deposit the matrix have greatly alleviated several of these problems, as this approach enables a very even matrix deposition and does not hold the risk of analyte migration since the matrix immediately crystallizes [25].
- **SIMS** In Secondary ion mass spectrometry imaging (SIMS), ions are generated by bombarding the surface of a specimen with a focused primary

ion beam, which causes the release of secondary ions from the sample [27]. SIMS was the first technique to be used for IMS, both for imaging of tissue sections as well as a wide variety of other samples in materials science. SIMS for biological applications has very high spatial resolutions typically down to a few hundreds of nanometers [28], however, compared to MALDI, SIMS suffers from extensive fragmentation of secondary ions and a smaller mass range up to 1500 Da, making it suitable to monitor the distribution of smaller molecules such as phospholipids, primary metabolites and pharmaceuticals in tissue sections. In recent years, cluster ion sources have been introduced for use in biological SIMS [29,30]. These cluster ion sources, which use for example C_{60} ions, reduce source fragmentation by providing softer ionization and therefore allow for an increased sensitivity and a wider mass range. SIMS has the advantage that it requires little to no sample preparation.

- **DESI** Desorption Electrospray Ionization (DESI) is a relatively new ionization technique which was introduced in 2004 by Takats *et al.* [31] and consists of spraying the sample with an electrically charged mist of solvent droplet. The impact of the charged particles on the surface causes the molecules at the sample surface to be ionized and ejected from the sample. DESI has the advantage that it can operate at ambient pressures and does not need any sample preparation. DESI has been used for imaging of biological samples targeting phospholipids [32], drugs and metabolites [33] and more recently tryptic digests of proteins [34]. Resolutions for DESI range around 200 μm. However, Laskin *et al.* [35] have demonstrated that resolutions down to 12 μm are attainable using nanospray DESI (nano-DESI) for analysis of lipids.

Several other ionization techniques such as laser ablation inductively coupled plasma (LA-ICP) [36], laser ablation electrospray ionization (LAESI) [37] and liquid extraction surface analysis (LESA) [34] mass spectrometry have recently been introduced as new techniques for IMS and are gaining increased popularity due to the fact that they require minimal to no sample preparation and can be used at ambient pressures.

Mass analyzer The mass analyzer separates the ions collected from the ion source according to their mass-to-charge ratio. A variety of different mass analyzers exist which differ in sensitivity, mass resolution and mass accuracy. TOF mass analyzers, which we previously described, were the first to be used for MALDI IMS, and are still very popular for protein-focused IMS today [38]. Next to TOF mass analyzers, ion traps [39] [40], QqTOF's [21], trap-TOF [41] and triple quadrupoles [42] have previously been used in IMS. More recently,

Fourier transform ion cyclotron resonance (FT-ICR) [43] and Orbitrap [44] mass spectrometers have made their way into the IMS field. Though these devices typically have a smaller mass range, their extremely high mass resolving power enables the user to make very sensitive and accurate measurements of small molecules such as lipids, drugs and metabolites.

Ion detector The ion detector enables detection of the ions that are sent through the mass analyzer. Ion detection in TOF-mass spectrometry is usually done using microchannel plates (MCPs) [45]. A noteworthy and relatively new development in tissue-focused IMS is the introduction of position-sensitive detectors. Traditionally, IMS is done in 'microprobe mode', where each individual pixel of the image is measured separately, and the ions of each pixel are measured by the detector sequentially. Alternatively one can perform IMS in 'microscopy mode', where a larger area of the tissue is irradiated using a relatively wide laser spot (typically 200-300 μm), and the resulting ions are then projected onto a position-sensitive detector using ion-optical microscope elements [14]. This has several advantages, firstly that the spatial resolution is not determined by the spot size of the laser, enabling high spatial resolutions (down to 6 μm resolution in MALDI IMS), and secondly that analysis speed can be greatly increased by analyzing large areas of the tissue in a single measurment (40.000 laser shots in a microprobe mode in a single laser shot in microscopy mode) [46].

The different types of IMS all have their own advantages and the correct type of IMS to use depends on the goal of the IMS experiment, the required spatial and mass resolution, and the target analyte. The different IMS types are often complementary and therefore several studies have previously combined different IMS types to maximize the amount of biomolecular information that can be extracted from the sample [47] [48].

1.5 Computational analysis of IMS data

As described above, IMS enables us to monitor several thousands of molecules at many different locations in the tissue simultaneously, ranging from low-mass metabolites and lipids to high-mass biomacromolecules such as proteins at an unmatched precision. While this makes IMS an immensely valuable technology for explorative research, it also leads to file sizes of several gigabytes and more recently even terabytes of complex and high dimensional data from a single examined tissue slice. As shown in figure 1.3, the data structure of an IMS dataset is a 3-Dimensional tensor or data cube, with two spatial modes (x and

y) and one mass-to-charge mode (m/z). The size of an IMS dataset is primarily influenced by two parameters: the number of measured locations or pixels, and the number of scanned m/z-bins. The pixel number increases as the tissue surface area that needs to be covered grows larger or the spatial resolution becomes finer. The number of m/z-bins on the other hand is proportional to the extent of the mass range and the mass resolution of the ion detector ($= m/z$ step size). In recent years both the spatial resolution and the mass resolution have been pushed upwards by technological advances, with spatial resolutions going to sub-micrometer levels and the number of m/z-bins going into the millions for high-mass resolution instruments such as FT-ICRs. An ion image can be generated for each measured m/z-bin (see Figure 1.2), leading to potentially millions of ion images per IMS experiment. Typically, however, only several hundreds to thousands of different biomolecular species are present in the mass range measured in an IMS experiment. Consequently, a large number of the measured mass bins will contain only noise. Retrieving the relevant biomolecular information from the millions of ion images, and discovering relationships between the retrieved biomolecular ions through manual examination and interpretation of the data is becoming increasingly infeasible. This is especially the case in an explorative setting where no prior targets of analysis are known. In order to harness the full potential of these data, there is a strong need for computational methods that can reduce the size of these data to manageable proportions, that can automatically extract features of interest and report them back to the user and that can guide the user towards relevant findings and aid in interpreting them. **It is the development of these computational methods that will be the focus of this thesis.**

Computational analysis of IMS data as a field only started to develop relatively recently, but has gained significant momentum over the last several years [49] [50]. Due to the complexity of IMS data there are a variety of computational challenges to be met and in this thesis we will focus on three of these in particular, with each having its own chapter dedicated to it:

- In Chapter 2 we focus on dimensionality reduction and feature selection for IMS data using the Discrete Wavelet Transform. More specifically, we improve on previous work done in this area by incorporating the spatial information available in IMS to extract key features and allow a more specific DWT analysis.
- Chapter 3 deals with the challenge of computationally comparing IMS data from multiple tissue slices. Performing such a comparison is far from trivial due to the large heterogeneity and the large size of IMS datasets. We therefore introduce a technique called Group Independent Component Analysis (GICA) to do this comparison.

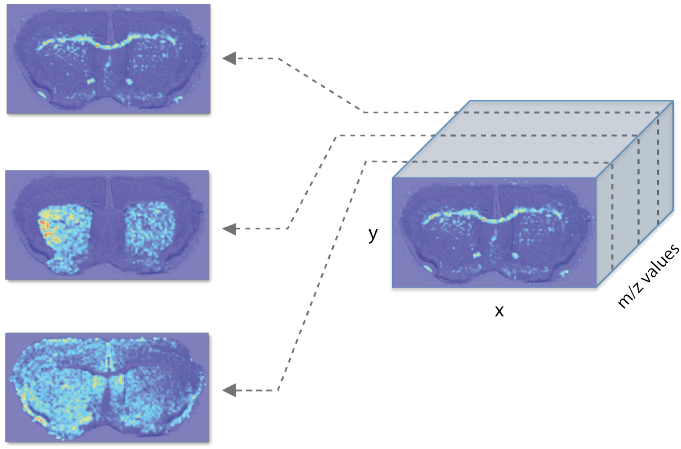

Figure 1.3: The data originating from an IMS experiment can be represented as a 3-Dimensional tensor, with 2 spatial dimensions and one m/z dimension.

- In Chapter 4 we explore the potential of integrating IMS data with information from additional data sources and, by doing so, creating richer datasets and enabling deeper insights into the IMS data. We integrate IMS data with a curated anatomical atlas and use these data to provide automated anatomical interpretation of IMS datasets.

We conclude the thesis with Chapter 5 where we give our concluding remarks and future perspectives.

If not stated otherwise, the IMS data analyzed in this thesis have been acquired and kindly made available by our collaborative partners:

- Prof. Dr. Etienne Waelkens, Laboratory of Protein Phosphorylation and Proteomics, KU Leuven, Leuven, Belgium (Chapters 2 and 3)
- Prof. Dr. Richard M. Caprioli's group, Mass Spectrometry Research Center and Departments of Biochemistry, Chemistry, Pharmacology, and Medicine, Vanderbilt University, Nashville, TN, USA (Chapters 1, 3, 4)

The main results of this work overall are the following:

- In our work on feature selection and dimensionality reduction, we demonstrate that using the DWT, IMS data can be considerably reduced in data size and dimensionality while still allowing for very good

reconstruction of the original mass spectra. We improve on previous work in this field by incorporating spatial information into the analysis and do this in two different ways, namely using global and local spatial information. We show that these methods can provide improved feature selection over traditional methods that do not incorporate this information. The developed new methods show a better retention of features that contain true signal and can provide improved data compression and dimensionality reduction by discarding more features that are due to noise and thus do not need to be retained.

- In our work on differential analysis of IMS data, we first demonstrate correct operation of the GICA algorithm using an artificially generated dataset containing three different disease classes, and show that the algorithm allows for near-perfect retrieval of the differential components between tissues in the correct locations. We then continue by applying this technique to two real-life IMS datasets, a case study of amyotrophic lateral sclerosis (ALS) mouse spine and one of retinal degeneration disease in mouse retina. Application of the GICA algorithm to these datasets results in retrieval of multiple differential biomolecular patterns between the different tissue classes, which are expressed in biologically relevant areas for the diseases under study.

- In our work on data fusion we successfully link IMS data directly to an anatomical atlas, namely the Allen Mouse Brain Atlas (AMBA) [107, 117]. This allows for direct, computational access to an important resource that many human interpretations of IMS data will naturally rely upon, namely anatomical insights. We spatially map IMS data to the atlas using non-rigid image registration techniques and use the established link to investigate spatial correlations between ion images and the anatomical structures of the atlas, gaining basic insights into the relationships between the two. We finally move beyond simple correlation and use the atlas information to provide automated anatomical interpretation of the ion images in the IMS data. This challenge is solved as a convex optimization problem that deconstructs ion distributions as combinations of known anatomical structures. We demonstrate that establishing a link between an IMS experiment and an anatomical atlas can serve as an important accelerator both for human and machine-guided exploration of IMS experiments.

  This work was published in the following paper:
  N. Verbeeck, J. Yang, B. De Moor, R.M. Caprioli, E. Waelkens, R. Van de Plas, "Towards Automated Anatomical Interpretation of Ion Distributions in Tissue: Linking Imaging Mass Spectrometry to Curated Atlases", *Analytical Chemistry*, vol. 86, no. 18, pp. 8974-8982, 2014.

Additional work on SELDI-TOF MS (surface enhanced laser desorption/ionization time-of-flight mass spectrometry) was done over the course of this PhD. This work has not been included in the remainder of this thesis, but has been published in the following journal publications:

A. Fassbender, N. Verbeeck, D. Boernigen, C.M. Kyama, A. Bokor, A. Vodolazkaia, K. Peeraer, C. Tomassetti, C. Meuleman, O. Gevaert, R. Van de Plas, B. De Moor, Y. Moreau, E. Waelkens, T. D'Hooghe, "Combined Microarray and Proteomic analysis of eutopic endometrium of women with and without endometriosis.", *Human Reproduction*, vol. 27, no. 7, pp. 2020-2029, 2011.

In this publication we analyzed plasma samples from infertile women with or without pelvic pain with SELDI-TOF MS to predict the presence endometriosis. Least Squares Support Vector Machines (LS-SVMs) were used to successfully classify SELDI-TOF peptide and protein profiles from different disease groups with high sensitivity and specificity.

A. Fassbender, E. Waelkens, N. Verbeeck, C.M. Kyama, A. Bokor, A. Vodolazkaia, R. Van de Plas, C. Meuleman, K. Peeraer, C. Tomassetti, O. Gevaert, F. Ojeda, B. De Moor, T. D'Hooghe, "Proteomics Analysis of Plasma for Early Diagnosis of Endometriosis.", *Obstetrics & Gynecology*, vol. 119, no. 2, pp. 276–285, Feb. 2012.

In this study the aim was to perform a combined mRNA microarray and proteomic analysis on the same eutopic endometrium sample obtained from patients with and without endometriosis. SELDI-MS data were successfully classified using LS-SVMs, and allowed for diagnosis of endometriosis in the luteal phase with high sensitivity and specificity. mRNA is differentially expressed in women with and without endometriosis during the menstrual phase when compared with the early luteal phase.

Chapter 2

Feature selection and dimensionality reduction for IMS data

In this chapter we focus on the dimensionality reduction and feature selection of IMS data. We propose the Discrete Wavelet Transform (DWT) as a dimensionality reduction tool that can achieve considerable reduction of the original data size, while keeping all the useful information in the IMS spectra intact. We accomplished this by discarding only those features that contain noise, and which take up unnecessary storage space and hamper computational analysis. We differentiated between variables that contain noise and those that are the result of true biological signal by incorporating a key feature of IMS data, spatial information, into the selection procedure.

2.1 Introduction

IMS enables the simultaneous detection of thousands of biomolecules throughout tissue, ranging from low-mass metabolites and lipids to high-mass proteins. This commonly leads to several gigabytes of complex and high-dimensional data from a single tissue experiment. The size of an IMS dataset is primarily influenced by (1) the number of measurement locations (pixels) and (2) the number of scanned m/z-bins. Both these parameters are constantly pushed upwards with ever-advancing instrumentation, making the computational analysis of IMS data

in raw form increasingly difficult. The two data aspects that hamper analysis the most are:

Data size As an example, a relatively small IMS experiment of only 100 x 100 locations, results in 10 000 pixels. If each pixel provides a mass spectrum with a size of 1 MB, this amounts to a total dataset size of 10 GB for a single tissue slice. This size exceeds the RAM capacity of most present-day desktop computers (8 GB RAM). This makes it impossible to load the data into memory in one go and greatly hampers computational processing of the IMS data. Recent datasets that were measured by our collaborators using an FTICR analyzer range in their raw form up to 1.5TB of data for a single tissue slice, making processing of these datasets problematic even on high-performance clusters.

High dimensionality The second issue is the high dimensionality of raw mass spectrometry signals, combined with a relatively low number of samples or pixels in most IMS scenarios. In many IMS experiments, the number of features (scanned m/z-bins) M, is in the order of 100.000 to 1.000.000, whereas the number of samples (measured pixels) P [1], is in the order of 10.000 to 100.000 pixels. Thus, $M \gg P$, which means that the number of features is much larger than the number of measurements. Ideally, we would want the reverse, namely $P \gg M$, such that the data could be used directly in so-called well-posed problem formulations. However, the situation is even somewhat worse since the number of true independent samples is typically less than the total number of samples acquired. The reason is that spectra obtained from neighboring pixels will often have a strong correlation with each other, and therefore cannot be considered truly independent from each other. The high dimensionality of full profile mass spectra makes operating directly on the full IMS dataset computationally heavy. It can also compromise the quality of data analysis methods through accumulation of noise over these high numbers of features, and the collinearity between IMS data used predictors can steer many datamining algorithms towards erroneous conclusions [51].

As a result, there is a growing need for thorough preprocessing steps, which can reduce both the dimensionality and data size through proper feature selection. Currently, the most-commonly used form of feature selection in IMS is peak picking. In this approach, local maxima in the spectrum are identified as m/z-bins of interest and the mass spectrum is converted to a discrete set of peaks. Although a common step in mass spectral data analysis, peak picking for IMS data is far from a standardized method. Most frequently, peak picking

[1] In IMS literature, a pixel commonly refers to a single gridpoint of the measurement grid defined in the IMS experiment.

is performed on the mean or median spectrum over all pixels. This method is very fast and will extract the most abundant features from the IMS dataset, but carries the risk of discarding peaks (ions) that are only present in a small area of the tissue (i.e. a limited number of pixels) or that are only present in low quantities (i.e. have a low ion intensity). When taking the average over the entire dataset, the signal of such ions will be diluted over the large number of pixels. This leads to a low signal in the average spectrum, which can in turn result in the omission of these features. Alexandrov *et al.* [52] have circumvented the averaging issue by performing peak picking on a limited and randomly selected subset (10%) of all spectra. However, this approach once again risks discarding peaks that are only present in a limited area of the tissue. McDonnell *et al.* [53] have developed a custom peak-picking method for IMS data, which specifically attempts to capture localized features. Eijkel *et al.* have designed an efficient peak-picking algorithm that uses the derivative of the IMS spectra to identify peaks in the data. The general approach of peak picking typically offers a strong reduction in both size and dimensionality of the dataset, often reducing the number of m/z-bins from 100.000's to 100's and reducing the data size by a factor 100 to 1000. However, it is a relatively harsh approach to dimensionality reduction that commonly throws out a lot of potentially relevant information. A good example of such information is the peak shape. Peak shape can be used to discern whether a single peak might be reporting two distinct but (almost) isobaric ion species, i.e. ions that have nearly the same mass. Such information is commonly lost when peak picking.

Ideally one would want to reduce the size of the dataset to a practical level, while still retaining the maximum amount of biochemical information contained within the signal. This chapter discusses the Discrete Wavelet Transform (DWT) as a means of accomplishing just that. In our group the DWT has previously been successfully applied to IMS data by Van de Plas *et al.* [54] [55] In this chapter we improve on that approach by incorporating one of the key aspects of IMS, spatial information, into the feature selection process. By using this information, we can selectively remove features that do not exhibit a spatial structure and are thus presumably caused by noise while retaining as much of the useful information as possible. Wavelet analysis has been previously used in the context of mass spectrometry [56] [57], but applications in IMS data remain scarse.

This chapter first introduces the concept of wavelet transformation through the so-called Continuous Wavelet Transform (CWT). We then progress to the DWT and the concept of wavelet thresholding that determines which features are retained after the analysis. We demonstrate two different methods that incorporate the spatial information of IMS data into this thresholding process in order to improve feature selection. One method uses global spatial information

from the IMS data, and one uses localized spatial information. We demonstrate the performance of these new feature selection methods on a case study of a 10 μm thick sagittal section of the brain of a BL57/6 mouse and on an artificial dataset.

2.2 Continuous Wavelet Transform

Wavelet analysis [58] has a huge number of applications in science, engineering, and mathematics. It is used for the analysis of a broad variety of data, including image and audio data [59]. As a mathematical tool, wavelet analysis is closely related to the well-known Fourier transform. Both methods use a particular signal as a probing function and assess how strong the correspondence is between this pre-defined probing function and the signal one wants to analyze. Since the Fourier transform analyzes the frequency content of a signal and ignores signal events that are localized to only a specific time window, it uses a probing function that does not have such localized features, namely a perpetually repeating sine wave at a particular frequency. This means that through a Fourier transform essentially all information on localized and non-repeating peaks in the signal is removed and lost. Wavelet analysis, on the other hand, concentrates on localized events, even in non-repeating signals. This is why the wavelet transform uses a probing function that is explicitly local, the wavelet. In a wavelet transform, scaled and shifted versions of a "mother wavelet" are compared to the signal and will yield both localization and frequency information on the measured signal. The Continuous Wavelet Transform (CWT) can be written as follows:

$$C(a,b) = \frac{1}{|a|} \int_{-\infty}^{\infty} x(t) \overline{\psi} \left(\frac{t-b}{a} \right) dt \tag{2.1}$$

where $\bar{\psi}$ represents the complex conjugate of the mother wavelet ψ, $x(t)$ represents the signal under analysis, a is the scale function ($a > 0$), and b represents the position along the signal. By probing the signal of interest with wavelet functions at continuously varying and different values for the scale parameter, a, and the position parameter, b, we obtain a set of coefficients $C(a,b)$ as a function of a and b.

Many different types of mother wavelet shapes exist and can be used in CWT analysis. This has the advantage that a wavelet can be specifically selected for the application at hand, by selecting a wavelet that closely resembles the specific feauture one is looking for. An example of one of the most commonly used wavelets, the Mexican hat wavelet, is shown in Figure 2.1. The Mexican

hat wavelet is defined as [60, 61]

$$\psi_{Mexicanhat}(t) = (1 - t^2)e^{-0.5t^2} \tag{2.2}$$

which is the second derivative of a Gaussian waveform. This is a useful wavelet for analysis of mass spectrometry data, due to its resemblance to a typical peak shape as observed in mass spectra. For this reason, Du *et al.* used this wavelet for peak detection in mass spectrometry data [62].

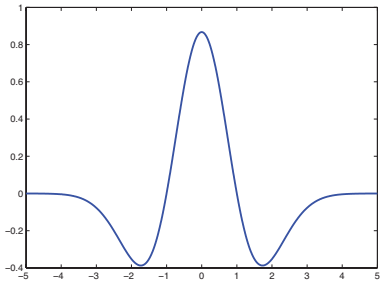

Figure 2.1: The Mexican hat wavelet $\psi_{Mexicanhat}(t)$.

2.2.1 Application of CWT to IMS Data

The easiest way to understand the CWT is by demonstrating it on an example signal. In Figure 2.2 we apply CWT analysis to our signal of interest, which is (part of) a mass spectrum. We use the Mexican hat wavelet for the analysis. In panel 1 we take the wavelet and compare it to the section at the beginning of the signal. Using Equation 2.1, we calculate the CWT coefficient C for this scale and location of the wavelet, which represents how similar this section of the signal is to the scaled wavelet. The larger the value of C, the more similar the two are, and in this particular situation the coefficient is relatively high. Once C is calculated, in panel 2 the wavelet is shifted to the right and a new coefficient C is calculated for this new position. We repeat this procedure until the entire mass spectrum is probed. In panels 3 and 4, the scale (or stretch) of the wavelet is changed and the probing procedure is repeated from the beginning of the signal. By stretching the wavelet, it will now have a larger similarity to features with a lower frequency and a lower similarity with high-frequency features. In the context of mass spectrometry, features with a low frequency are broad peaks, however, even broad peaks have a relatively sharp peak top. For this reason, in panel 1, the high frequent wavelet shows high similarity to the small, high frequent peaks and, in panel 2, also shows high similarity to the top

of the peak. The smooth sides of the broad peak do not contain high frequency features, however, and will therefore show low similarity to this high frequent wavelet. In panel 3, on the other hand, the small, higher frequent peaks exhibit low similarity to the low-frequent wavelet, while in panel 4 the similarity of the stretched wavelet with to larger peaks results in relative larger coefficient values. In summary, in the CWT analysis of a mass spectrum, broad peaks will be present in many different scale levels of the CWT analysis due to their sharp top and broad base. Thin and small peaks, on the other hand, will only be present in the lower scale levels of the CWT analysis.

When we repeat these steps for the whole spectrum and a range of different scale levels (e.g. from 1 to 10), we obtain the table of CWT coefficients in Figure 2.3. The CWT analysis plot contains the wavelet coefficient values C for all positions over the spectrum and for the 10 different scale levels. This plot nicely illustrates the power of wavelet analysis: we can analyze the frequency content of a signal, while still retaining the spatial locations of these frequencies. We can observe (1) high similarity of the higher-scaled wavelets (low frequencies) in the locations of the mass spectrum with broader peaks and (2) high similarity of the low-scaled wavelets (high frequencies) in locations with smaller, thinner peaks and around the tops of the larger peaks. As an example, the red frame in Figure 2.3 focuses on a broad peak in the spectrum. At the first CWT level, we see high similarity of the wavelets in the location of the spectrum where the sharp top of the peak is located (sharp = high frequency), and low similarity of the wavelets in the regions left and right of the top which contain the smooth edges of the peak (smooth = low frequency). The higher-scaled wavelets continue to show high similarity with the broad peak. The blue frame focuses on an area of the spectrum which contains only high-frequent peaks, which only have high similarities with the lower-scaled wavelets (high frequencies) and show no similarity with the higher-scaled wavelets (low frequencies).

2.2.2 Redundancy in CWT

When using the CWT, the wavelet can be shifted continuously over the entire mass spectrum, which allows for an easy interpretation of the analysis results as wavelet locations can be directly linked to peak locations, as we saw in Figure 2.3. Moreover, any desirable scale can be selected for the wavelets, from the smallest scale in the original signal up to any arbitrary scale determined by the user. Thus, performing the analysis in each location with multiple scaled wavelets makes it possible to perform a very detailed, high resolution analysis of the mass spectrum at hand. It also, however, makes the analysis computationally demanding and highly redundant (as each level of CWT coefficients has the same data size as the original spectrum). Given our initial standpoint on the

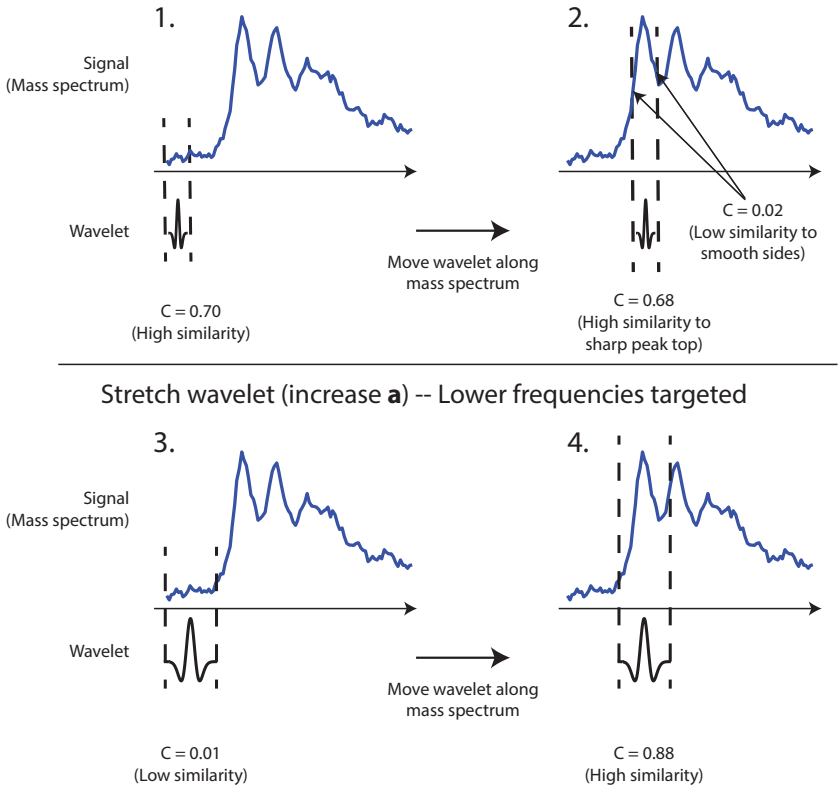

Figure 2.2: Example of CWT analysis on a section of a mass spectrum. We observe high similarity of the higher-scaled wavelets in the locations of the mass spectrum with broader peaks and high similarity of the low-scaled wavelets in locations with smaller, thinner peaks and around the tops of the broad peaks. As an example, the red frame focuses on a broad peak in the spectrum. At the first CWT level, we see high similarity of the wavelets in the location of the spectrum where the sharp top of the peak is located (sharp = high frequency), and low similarity of the wavelets in the regions left and right of the top which contain the smooth edges of the peak (smooth = low frequency). The higher-scaled wavelets continue to show high similarity with the broad peak. The blue frame focuses on an area of the spectrum which contains only high-frequent peaks, which only have high similarities with the lower-scaled wavelets (high frequencies) and show no similarity with the higher-scaled wavelets (low frequencies).

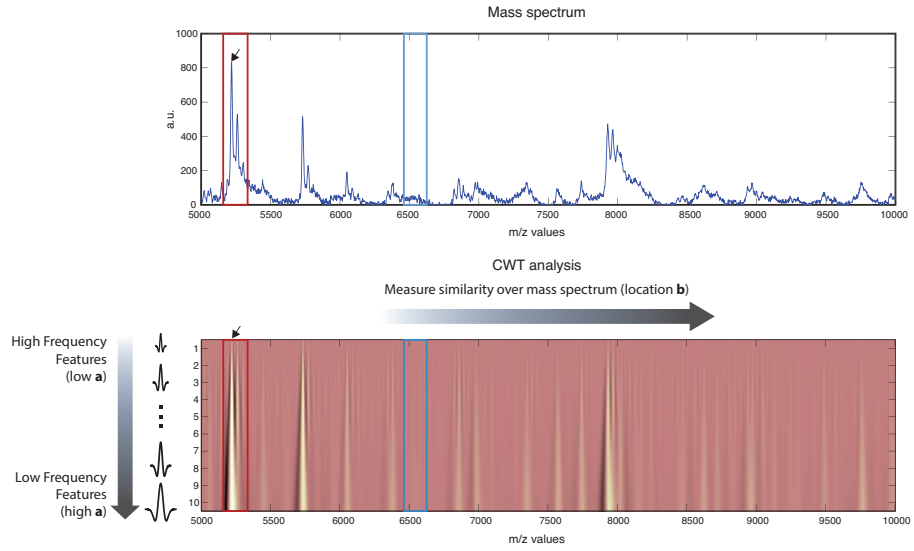

Figure 2.3: Results of the CWT analysis on a mass spectrum, using 10 levels of scaling.

large data size of IMS datasets, and our initial goal of representing the data in a more compact form, this might seem counter-intuitive. This is where the discrete wavelet transform comes in, as we will see in the following Section.

2.3 Discrete Wavelet Transform

One of the main differences between the CWT and the discrete wavelet transform (DWT) is the dyadic sampling scheme used by the DWT, which means that the shifting and scaling of the wavelets is based on a power of 2. We illustrate this in Figure 2.4, where we revisit our example of the mass spectrum used in the CWT analysis. Rather than smoothly and continuously moving the wavelet over the entire spectrum as we did in the CWT, the DWT breaks the signal up into sections, which are each individually compared to a wavelet of choice to calculate the wavelet coefficient for that section. At each consecutive (scale) level, the wavelet is scaled (stretched) by a factor of 2, and the signal sections into which the signal is broken up become twice as long.

The DWT can be represented as a tree of low and high pass filters, with each consecutive (scale) level of the DWT analysis operating on the output of the

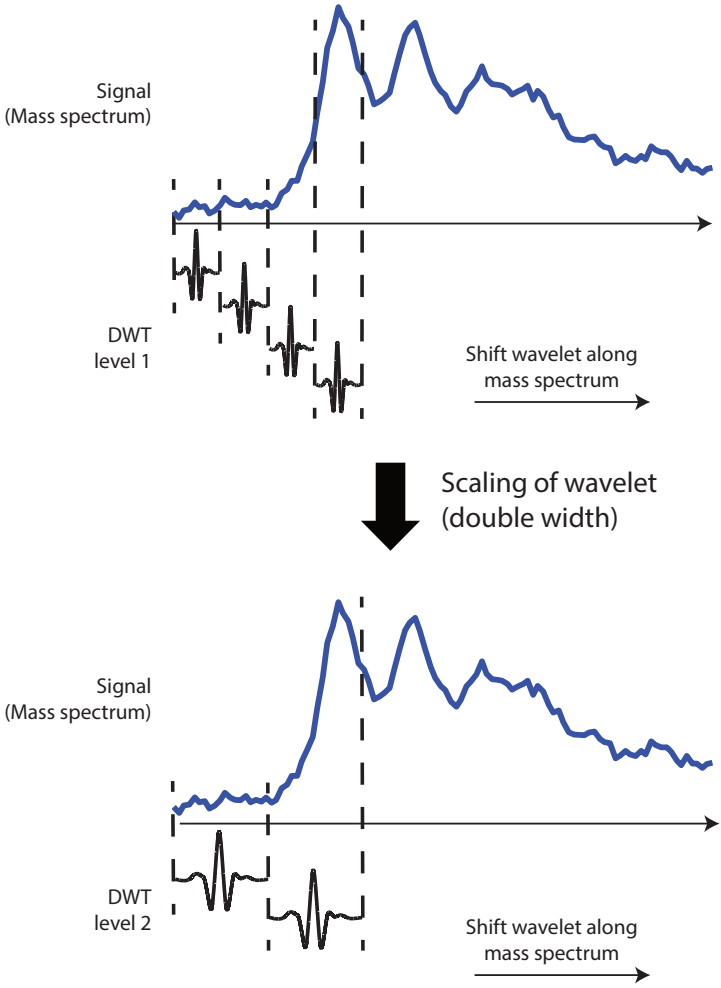

Figure 2.4: Illustration of the DWT applied to a section of a mass spectrum. The DWT uses a dyadic sampling scheme, where the scaling and shifting of the wavelets is based on a power of 2. This means that at each consecutive (scale) level, the wavelet is stretched by a factor 2.

low pass filter. As shown in Figure 2.5, by performing the DWT analysis, the original signal of length N is split into (1) an approximate signal A_1, which carries the low frequency components of the signal and is represented by the approximation (wavelet) coefficients Ac_1 and (2) a detail signal D_1 which carries the high frequency components and is represented by the detail (wavelet) coefficients Dc_1. Both the approximate signal and the detail signal initially have the same length N as the original signal, however, the number of coefficients can be downsampled by two for both the detail signal and the approximate signal [63]. In this way, we can write the approximate signal and the details signal with only $N/2$ wavelet coefficients, Ac_1 and Dc_1 respectively. The next DWT level then takes the approximate signal as an input and again splits an approximate signal A_2 and a detail signal D_2, which can be represented by $N/4$ wavelet coefficients, respectively Ac_2 and Dc_2. This process can be repeated until only one coefficient remains, however in general the number of levels is user-specified according to the goal of the application. When using k levels of scaling, we decompose the original signal into one approximate signal A_k with wavelet coefficients Ac_k and k detail signals $D_1, \cdots, D_k$ represented by wavelet coefficients $Dc_1, \cdots, Dc_k$ respectively.

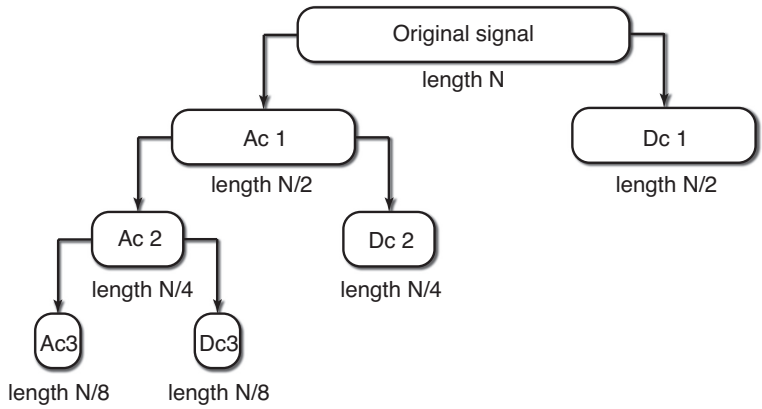

Figure 2.5: At each level of the DWT analysis, the data are split into 2 signals, an approximate signal (A), which is represented by the approximate wavelet coefficient (Ac) and a detail signal (D), which is represented by the detail wavelet coefficients (Dc). The number of wavelet coefficients necessary to represent these signals is halved at each consecutive level.

Performing wavelet analysis using the DWT avoids the redundancy encountered in the CWT. When the total number of wavelet coefficients from the approximate signal Ac_k and all detail signals Dc_1 to Dc_k are summed up it amounts to a total of N coefficients. This means that the result from a DWT-based wavelet

decomposition of a signal takes up the same space as the original signal. The advantage of the DWT lies in the extra information provided, namely the decomposition of the signal into parts that are 'details' and parts that are 'approximate'. It is this added distinction between essential and detail signal parts that allows the DWT to be used for one of its most popular applications: data compression.

2.3.1 Wavelet compression

In theory, the DWT is a lossless transformation: using an operation called the inverse discrete wavelet transform (IDWT), the original signal can be completely reconstructed by taking in all N coefficients from the Dc's and the Ac [64]. Keeping all these coefficients is often not necessary though, especially in the context of IMS. Mass spectra are often very sparse, and many of the measured m/z-bins only contain noise and no real signal. These noisy features take up unnecessary data storage space and hamper computational analysis by the added dimensions they represent. The power of the DWT lies in the fact that many of these noisy features are high in frequency and are therefore contained in the lower-level detail coefficients Dc, whereas true peaks have lower frequencies and are contained in the approximate coefficients Ac. If, for a moment, we presume that all the detail coefficients represent noise, we should be able to replace them by zeros without hurting the true content of the signal. Replacing the detail coefficients by zeros is essentially the same as reconstructing the signal solely on the basis of the approximate coefficients Ac. As an example, Figure 2.6 shows the decomposition of a single mass spectrum using 3 levels of DWT with the near-symmetrical Sym8 wavelet. On the right we show the detail signals and the number of detail coefficients necessary to represent this signal for each level, on the left we show the approximate signal for each level, which is the signal without the detail coefficients, and the number of approximation coefficients necessary to represent this signal. At the 3rd level of the DWT, we only need 824 approximation coefficients to represent the approximate signal, which is only 1/8th of the original number of datapoints in the mass spectrum, effectively reducing the data by a factor 8 in size. Interestingly, this eight-fold reduction in data size still maintains a good approximation of the original signal, and one could even argue that the signal contains much less noise. The wavelet approximation coefficients Ac_3 can thus be seen as a very efficient way of representing the original mass spectrum.

Van de Plas *et al.* [54, 55] have previously demonstrated that this concise representation of the data (using only 819 approximation coefficients) can be used instead of the original data to perform pattern extraction in IMS data using principal component analysis (PCA). Performing this analysis on the reduced

data gave PCA analysis results very similar to using PCA on the original data, reduced total processing time by 44% and led to a more than 4-fold reduction in memory requirements to conduct the analysis.

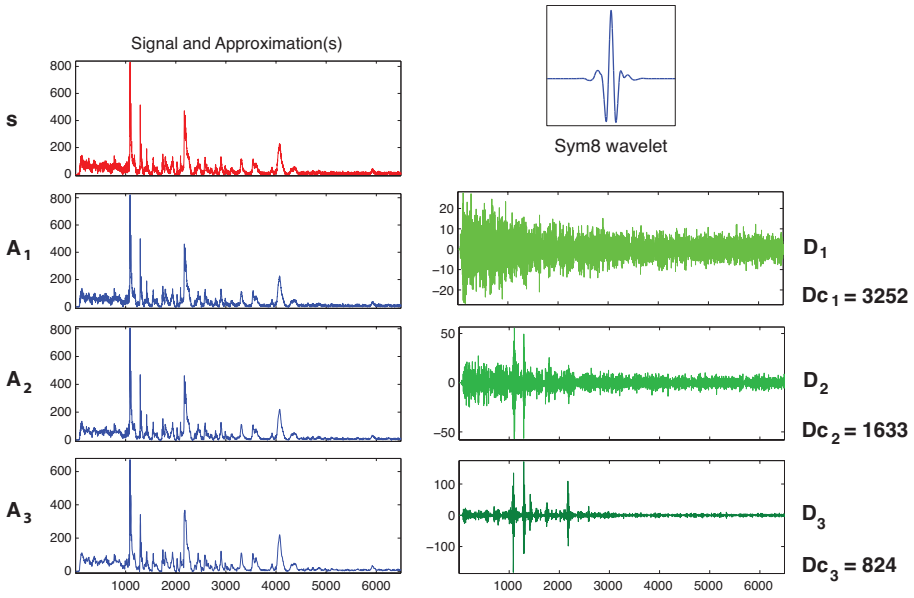

Figure 2.6: Decomposition of a single mass spectrum using 3 levels of DWT with the Sym8 wavelet. On the right we show for each level the detail signals and the number of detail coefficients necessary to represent these signals. On the left we show the approximate signal for each level, which is the signal without the detail coefficients, and the number of approximation coefficients necessary to represent this signal.

While keeping the number of wavelet coefficients retained to a minimum is key in reducing the data size and reducing computational complexity, there is a caveat. The DWT cannot distinguish between biologically relevant details and noise. This means that, although the detail coefficients of the DWT contain a substantial amount of noise as we have so far presumed, they also contain real signal. If left unchecked real tissue signal might be discarded along with the noise details. The higher the level of DWT we use, the lower the data size required to represent the original data, but also the higher the chance that genuine signal is removed. This is because the peak variation that gets removed grows broader by a factor of two each level we go deeper, and broader peaks tend to represent more genuine tissue signal. In short, with regards to data size reduction one wants as many DWT levels as possible, keeping as few wavelet

coefficients as possible, but with regards to signal fidelity one wants as little DWT levels as practical, keeping as many wavelet coefficients as practical.

Throwing away too many detail coefficients can thus result in loss of important data and the research presented in the remainder of this chapter focuses specifically on the issue of the detail coefficients by providing the tools to tell apart the tissue signal-carrying detail coefficients from the noise-carrying ones. It hence provides a means for a more nuanced feature reduction in which only the noise-carrying detail coefficients are removed, and the others are maintained. We have developed two different approaches to perform this sub-selection of detail coefficients. Both approaches incorporate one of the key aspects of IMS, spatial information, into the selection procedure. By using this information, we can selectively remove only those details that do not exhibit a spatial structure. The assumption is that genuine tissue signals will tend to describe non-random spatial patterns, while randomly distributed details are more likely to be noise. The first method is described in Section 2.4 and uses global experiment-wide spatial information to advise on which details to keep. The second method is described in Section 2.5 and uses localized spatial information to improve its feature selection.

2.4 Coefficient selection incorporating global spatial information

As we discuss above, our hypothesis is that detail coefficients that exhibit spatial structure are likely to describe true tissue signal, and should therefore be retained. This is based on the common observation that the biological ions measured in IMS will show a clear relationship with the underlying tissue, and will either be present in distinct regions of the tissue or will be distributed throughout the entire tissue. The underlying assumption is that a biological molecule serves a specific purpose in the tissue, and that it can be linked to the tissue's specific biological function, or that it is necessary for basic cellular functioning (molecules present in all cells throughout tissue). A biological molecule that is randomly distributed throughout the tissue makes little sense from a biological point of view, and therefore detail coefficients that show a random spatial distribution in the tissue are most likely due to noise, and can therefore be safely discarded. To identify coefficients that exhibit spatial structure, we use Principal Component Analysis (PCA), a method commonly used for pattern extraction in IMS [65, 66]. PCA is a well-known multivariate dimensionality reduction technique [67] that uses orthogonal transformation to transform a set of potentially correlating variables into a set of uncorrelated variables called the principal components (PCs). PCA is defined such that

the first PC explains the largest possible variance in the original data. Each subsequent PC is orthogonal to the previous PCs and explains the largest possible variance that remains in the data after the preceding PCs have been removed from the original data. In chapter 3 we will study PCA in greater detail.

We demonstrate this new feature selection method on a case study of a 10 µm thick sagittal section of the brain of a BL57/6 mouse (a common inbred line of laboratory mice). The IMS measurement was done on an Applied Biosystems 4800 MALDI TOF/TOF mass spectrometer with the recorded mass range set from m/z 2800 to 25000, divided into 6490 m/z-bins. The grid size was 51 $\times$ 34 pixels with an interspot distance of 300 µm in both the x and y directions. After removing pixels located outside of the tissue, a total of 1381 mass spectra remained [2]. The IMS data were normalized using total ion current (TIC) and baseline was subtracted using MATLAB's standard baseline correction function (MATLAB 2010a, The Mathworks Inc., Natick, MA) using a window width of 150 Da and a step size of 150 Da. DWT analysis of the IMS data was performed using the Sym8 wavelet. Symlets, of which the Sym8 wavelet is a member, are a family of near-symmetric wavelets suitable for DWT analysis. Since the features of interest that we want to extract form the signal (ion peaks) are also near-symmetrical in shape, this choice of mother wavelet (as defined in section 2.2) is a natural fit. We use 4 levels of DWT analysis on all mass spectra (and thus pixels) separately, which splits the data up into a set of Ac_4 approximation coefficients, and 4 sets of detail coefficients, Dc_1 to Dc_4. As shown in Figure 2.7, each of these sets is stored in a matrix of size $P \times Dc_k$ where P is the total number of pixels or spectra.

The goal is now to find the detail coefficients in Dc_1 to Dc_4 that contain true tissue signal contrary to noise signal. We apply PCA to each of the matrices containing Dc_1 to Dc_4 in the same way as we would apply it to a full IMS dataset, and employ the decomposition into principal components as a means of finding which detail coefficients seem to correlate across multiple mass spectra. If details correlate across multiple measurement locations, it means that those details are not random occurrences and that they have a base of support at multiple tissue locations. Such a situation would indicate that the details in question are describing some structural pattern along the spatial mode of the data, and thus are likely to be a genuine sample observation. Alternatively, one can argue that details that do not correlate with observations at other tissue locations are more likely to be noise, and can be discarded. Figures 2.8

[2]The work presented in this chapter was done in the early stages of the PhD, and the numbers of m/z-bins (6490) of the presented datasets are modest in comparison to some current day datasets (100.000 m/z-bins), showing the speed at which IMS is evolving. Nonetheless the principles demonstrated in this chapter are equally applicable to present day datasets.

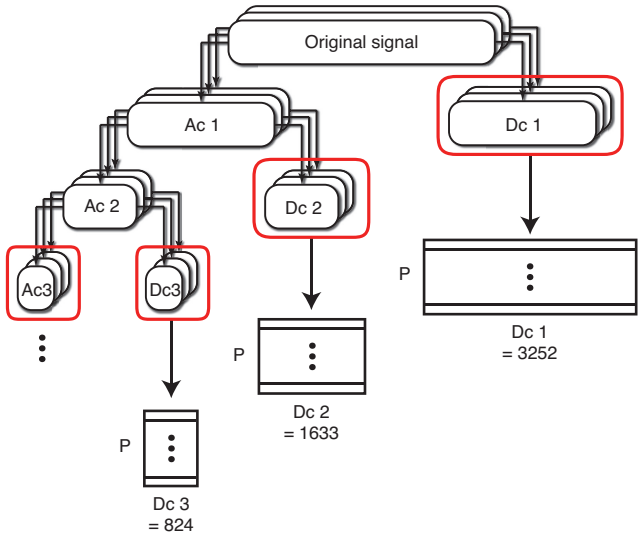

Figure 2.7: The DWT decomposition is repeated for all mass spectra in the IMS dataset. This results in a set of detail coefficients Dc for each level of the DWT per mass spectrum (and thus per pixel). The resulting detail coefficients are stored in a matrix of size $P \times Dc_k$. We show this for the first three levels of the DWT decomposition.

to 2.11 show the PCA results for Dc_1 to Dc_4 respectively. In each of these figures, the top panel shows the spatial expression (loadings) of PC1 to PC6 extracted from the Dc in question. The spatial signatures of most of these PCs are far from random and, although these data pertain to signal details, there seems to be a nontrivial amount of spatial (and even anatomical) structures present in the detail signals for each DWT level. Even coefficients describing very narrow variation at the least reducing level of DWT, Dc_1, seem to report some spatial structure, albeit at a less prominent amplitude than the deeper more reducing DWT levels. As expected, Dc_4, capturing details at a level that already represents relatively low-frequent and thus broader spectrum features, shows the largest amount and clearest anatomical structures. This observation is in line with our earlier statement that deeper DWT levels increasingly start capturing genuine signal rather than just noise. The bottom panels of Figures 2.8 to 2.11 show Pareto plots. A Pareto plot visualizes the cumulative variances accounted for by each successive PC. In all cases, the Pareto plots further indicate that the detail coefficients in question still contain a nontrivial amount of structured variation. To be precise, for DWT levels 1 through 4, the amount of variance explained by the first 6 components is 10%, 72%, 86%, and 87%. If

the entire Dc were noise, one would expect much lower amounts of variance to be explained by only six components and the monotonically decreasing slope of variance explained would be much less steep (each component should explain a similar amount of variance in non-structured noise). The 10% variance explained for level 1 does show that the structure contained in the coefficients remains limited, but the 72% variance explained at level 2 is surprisingly high, given that these are features that still have a very high frequency and were therefore expected to contain mostly noise. One of the reasons for true mass spectral signal to be captured at level 2 already is that the mass resolution of this example is relatively limited. In cases of modest mass resolution measurements a pattern spanning four m/z bins or more (the scale of variation captured by DWT level 2) is rarely a random intensity spike and therefore often describing genuine tissue signals. The DWT level depth at which the detail coefficients pivot from describing mostly noise to describing mostly genuine signal is thus dependent on the mass resolution at which the data set was required.

The center panels of Figures 2.8 to 2.11 show the scores resulting from the PCA analysis. These scores show along the spectral mode of the data the correlating detail coefficients that exhibit the spatial patterns in the top panels. From these spectral signatures, we determine which detail variables are part of the nonrandom spatial pattern and thus need to be retained. This is accomplished using a threshold (the red lines in the score plots) based on the standard deviation of the score values. Basing this threshold on the standard deviation has the advantage that no absolute threshold needs to be defined, but rather that this threshold adapts to the signal under analysis. This is analogous to more traditionally used wavelet thresholding techniques as we will see in Section 2.4.2. In this example, we take thresholds of 3, 3, 1.5, and 0.5 times the standard deviation for levels 1, 2, 3, and 4 respectively. Detail coefficients with a score that exceeds the threshold are retained, the remainder is discarded. The logic behind the decreasing threshold for higher DWT levels is based on two pieces of information. The first is based on the previously mentioned fact that low DWT levels have high frequent features, less likely to be due to true signal, and thus a feature must have strong evidence in the data to be selected for retention. For levels 1 and 2 we set this to 3 times the standard deviation: we assume that these wavelet coefficients are noise unless they exhibit extraordinary behaviour (3 sigma event), in this case have a very high score. For higher levels, we gradually decrease this threshold. The second reason is the fact that the first few levels of the DWT filter out the largest amount of coefficients. In our example, level 1 of the DWT contains the most detail coefficients (3252 detail coefficients, see Figure 2.7), and thus setting the threshold too low would incur a large impact on the number of additional features that are to be retained and thus on the total final data size. As the first six principal components seemed to capture most of the spatial structure that remains in the detail coefficients, we

apply these threshold to only these components and discard any details selected by the seventh PC and onwards. While the current threshold sub-selection method shows good results, as we shall see, one of the future challenges would be to develop a more rigorous and nonparametric thresholding method for the detail selection.

2.4.1 Data reconstruction

When applying the thresholds and discarding all detail coefficients that do not express spatial structure, we can use the retained coefficients to reconstruct the mass spectra. Figure 2.12 shows three versions of a randomly selected mass spectrum from the IMS dataset. The first is the original normalized and basline corrected mass spectrum before DWT analysis, the second is the mass spectrum reconstructed after DWT using only the approximation coefficients and discarding all detail coefficients, and the last spectrum is the mass spectrum reconstructed after DWT using our new method, retaining detail coefficients that exhibit spatial structure. The general shape of the spectrum seems well maintained by both DWT approaches. However, when we zoom in and look at a section of the mass spectrum, the strength of the new method is demonstrated. We observe a much more refined and better conservation of features that one would identify as peaks, and a better conservation of the original peak height. Regions of the spectrum that do not contain structured information are smoothed, and since for these regions the detail coefficients have been discarded, they no longer take up dimensions and memory space.

We summarize and compare the results obtained through both compression approaches in Table 2.1. The mean squared error and max error are calculated by comparing the compressed spectra with the original spectra. The mean squared error measures the average of the squares of the differences (the errors) between the original and the reconstructed spectrum. The max error, as the name suggests, measures the largest difference between the original and the reconstructed spectrum after DWT compression. The mean squared error and max error are strongly reduced by retaining the detail coefficients that show spatial structure. We can see that there is a noticeable decrease in the compression factor when we keep the details with spatial structure. However, from the previous analysis it should be clear that the higher compression in throwing away all detail coefficients comes at the cost of removing potentially biologically relevant signal variation. Depending on the application this may or may not be problematic. If, for example, the goal is to get a first impression of the content of an IMS dataset and to retrieve the general trends in spatial expression of the data, then not retaining the detail coefficients is usually sufficient and will greatly increase analysis speed and decrease memory requirements; if on

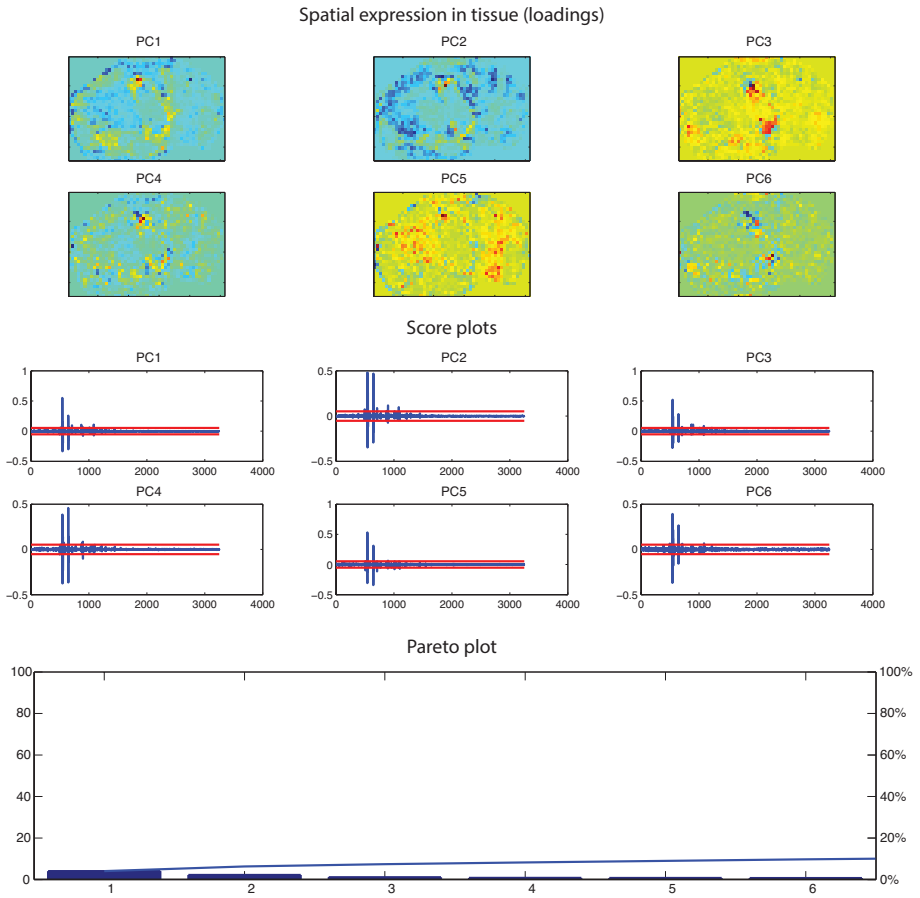

Figure 2.8: PCA analysis results of the DWT level 1 detail coefficients. (Top) Spatial expression of the first six detail coefficient principal components. The spatial expression of most of the PCs are far from random and there seems to be a nontrivial amount of spatial (and even anatomical) structures present in the detail signals, even at the first level of the detail coefficients. (Middle) Score plots corresponding to the principal components shown above, showing the correlating features. The red lines indicate the threshold employed for coefficient retention. (Bottom) Pareto plot showing the amount variance in Dc_1 explained by its first six principal components. The amount of variance explained by the first 6 components is much higher than what one would expect if no structure was present in the data. In the case of unstructured noise, each component of the PCA should explain approximately the same amount of variance.

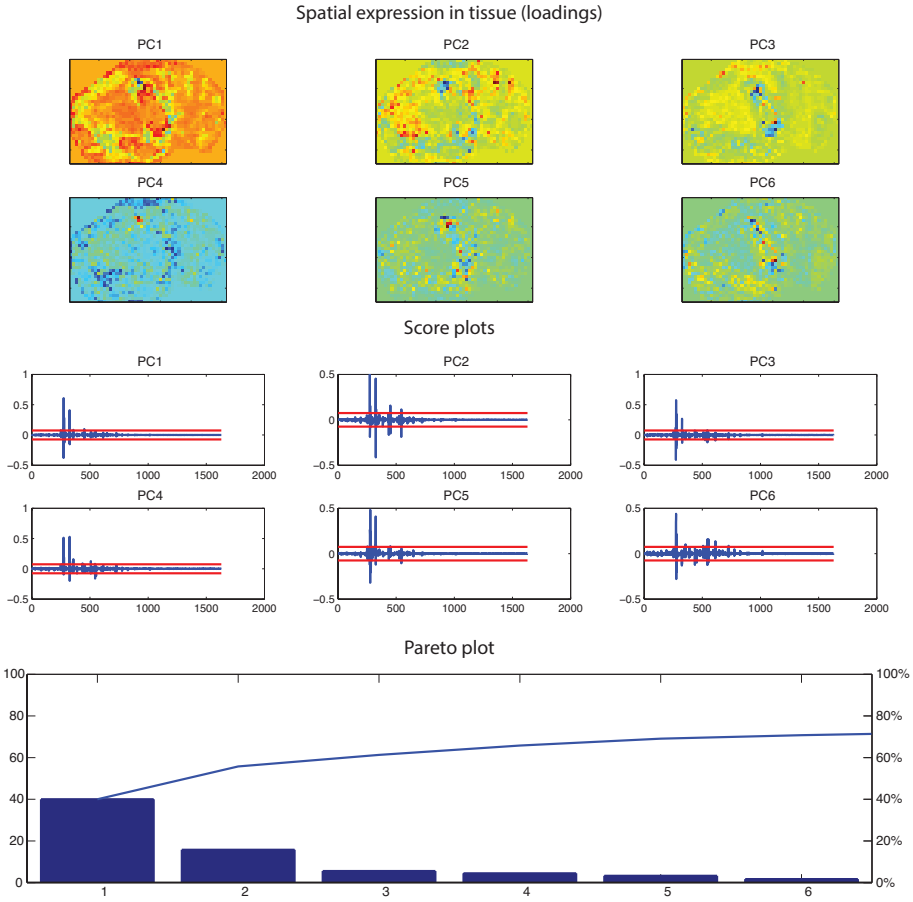

Figure 2.9: PCA analysis results of the DWT level 2 detail coefficients. (Top) Spatial expression of the first six detail coefficient principal components. The spatial expression of most of the PCs are far from random and there seems to be a nontrivial amount of spatial (and even anatomical) structures present in the detail signals, even at the first level of the detail coefficients. (Middle) Score plots corresponding to the principal components shown above, showing the correlating features. The red lines indicate the threshold employed for coefficient retention. (Bottom) Pareto plot showing the amount variance in Dc_2 explained by its first six principal components. The amount of variance explained by the first 6 components is again much higher than what one would expect if no structure was present in the data.

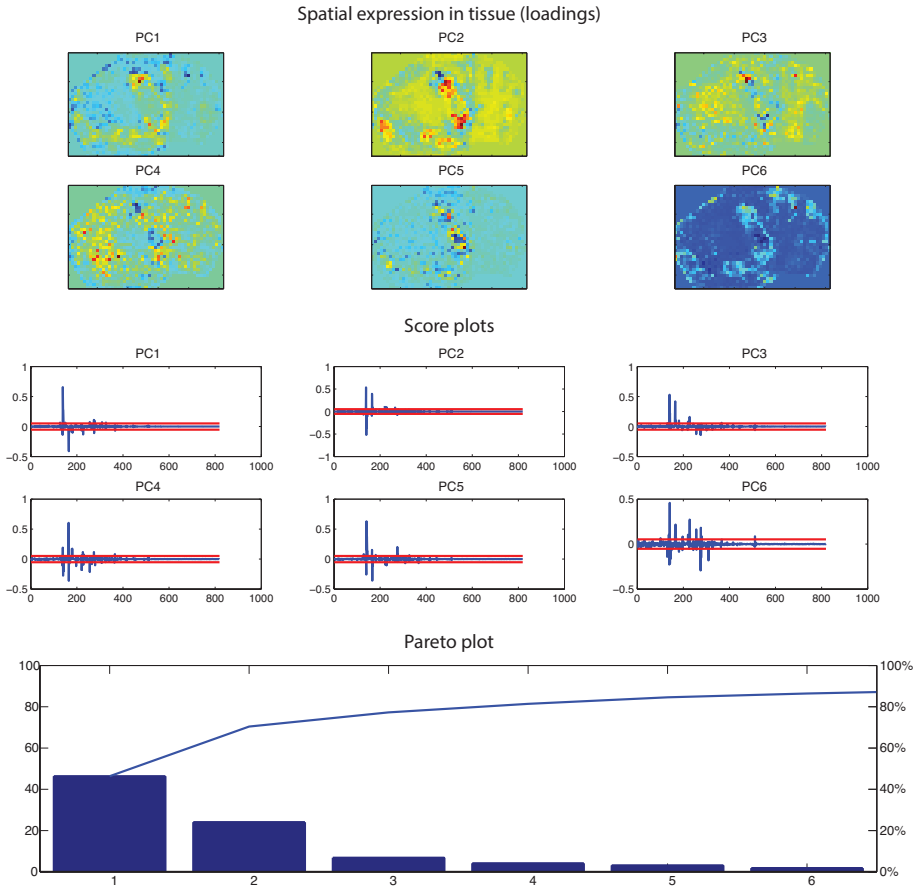

Figure 2.10: PCA analysis results of the DWT level 3 detail coefficients. (Top) Spatial expression of the first six detail coefficient principal components. The spatial expression of most of the PCs are far from random and there seems to be a nontrivial amount of spatial (and even anatomical) structures present in the detail signals, even at the first level of the detail coefficients. (Middle) Score plots corresponding to the principal components shown above, showing the correlating features. The red lines indicate the threshold employed for coefficient retention. (Bottom) Pareto plot showing the amount variance in Dc_3 explained by its first six principal components. The amount of variance explained by the first 6 components is again much higher than what one would expect if no structure was present in the data.

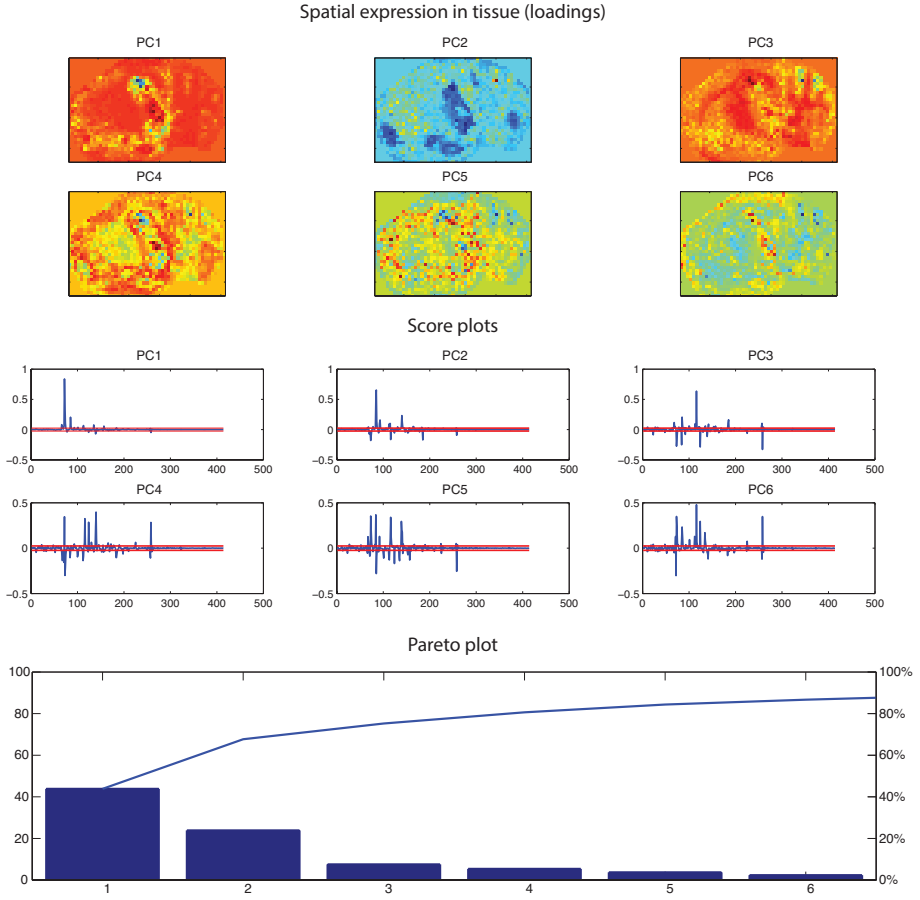

Figure 2.11: PCA analysis results of the DWT level 4 detail coefficients. (Top) Spatial expression of the first six detail coefficient principal components. The spatial expression of most of the PCs are far from random and there seems to be a nontrivial amount of spatial (and even anatomical) structures present in the detail signals, even at the first level of the detail coefficients. (Middle) Score plots corresponding to the principal components shown above, showing the correlating features. The red lines indicate the threshold employed for coefficient retention. (Bottom) Pareto plot showing the amount variance in Dc_4 explained by its first six principal components. The amount of variance explained by the first 6 components in much higher than what one would expect if no structure was present in the data.

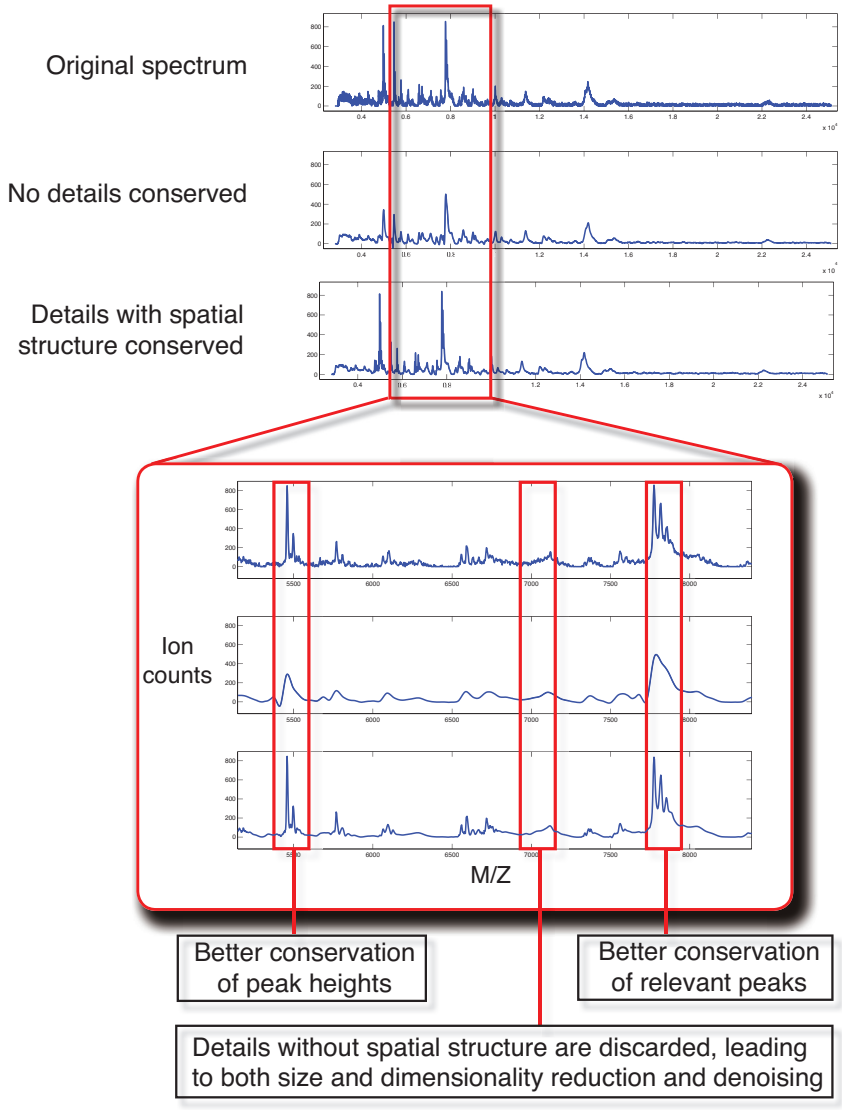

Figure 2.12: Comparison of the original spectrum (top), reconstructed spectrum without detail coefficients (middle) and reconstructed spectrum using spatially conserved features (bottom).

	Details with spatial structure conserved	No details conserved
Compression factor	10.7	15.6
Mean squared error	300.2	2004
Max error	139.6	1014

Table 2.1: By retaining a small number of detail coefficients that express spatial structure we can strongly improve reconstruction of the mass spectra while still achieving considerable compression. The mean and max errors are expressed as differences in ion intensity after normalization (arbitrary units).

the other hand the goal is to efficiently store IMS data for future use, aiming for a near-lossless compression, or to do a detailed analysis of the data that takes small features into account, conservation of spatially expressed detail coefficients is the recommended approach.

One remark concerning signal reconstruction in IMS using the IDWT, which has previously also been made by Van de Plas *et al.* [55], is the following: in theory the DWT is a completely reversible operation, and in principle the original signal can be perfectly reconstructed when all the detail coefficients Dc and the approximate coefficients Ac are used in their original form. However, when part of the wavelet coefficients are discarded or when the wavelet coefficients are quantized, this perfect reconstruction property no longer holds, and the reconstruction becomes shift-variant. This means that it is possible that shifts along the m/z-axis are introduced when reconstructing the data, and this due to aliasing effects [64]. In a mass spectrometry setting, such shifts are undesirable as they can change the observed m/z-values. In our reconstruction results we did not observe such shifts, however it is good to be aware that this phenomenon exists, and can be countered by using one of the variants on the DWT transform, such as the over-complete DWT [64] or the stationary wavelet transform (SWT) [68]. These algorithms eliminate such possible shifts by increasing the redundancy of the retained wavelet coefficients. In other words, these algorithms improve data reconstruction at the cost of a lower data compression rate.

2.4.2 Coefficient selection using universal threshold

In this section we compare our method, custom designed for IMS data, to a standard coefficient selection approach utilized in wavelet theory, namely

	Details with spatial structure conserved	Universal threshold	No details conserved
Compression factor	10.7	9.7	15.6
Mean squared error (in a.u.)	300.2	230.7	2004
Max error (in a.u.)	139.6	72.1	1014

Table 2.2: Comparison of the compression factors and signal reconstruction using all three DWT approaches. The mean and max errors are expressed as differences in ion intensity after normalization (arbitrary units).

thresholding of coefficients using the 'universal threshold'. The universal threshold is developed by Donoho and Johnstone [69] and achieves good reduction of noise while keeping the risk of discarding useful features at an acceptable level for most applications, hence the name 'universal' threshold. The universal threshold is given by:

$$\lambda_{UNIV} = \sqrt{2logN} \cdot \text{MAD} \tag{2.3}$$

where N is the length of the signal and MAD is the median of the absolute deviations of the signal.

Analogous to our previous analysis, we perform the DWT analysis of our IMS data using the Sym8 wavelet, decomposing the data over the first 4 levels. Once we have calculated the wavelet coefficients Dc_1 to Dc_4 we can now directly apply the universal threshold to the retrieved detail coefficients. The threshold is calculated for each mass spectrum separately. The wavelet coefficients that exceed the value of λ_{UNIV} have a value that exceeds the noise level far enough to be deemed true signal and are therefore retained, while coefficients below this threshold are considered noise and are discarded.

Table 2.2 shows the compression and reconstruction errors over all coefficient selection methods. The universal threshold shows a lower level of compression than that obtained with our method using spatial information, meaning that more detail coefficients are retained. The mean squared error and max error are also lower, which is to be expected, since with more detail coefficients the reconstructed signal will approach the original signal more. Figure 2.13 shows a comparison of the original spectrum, the reconstructed signal without details conserved, the reconstructed signal retaining detail coefficients that exhibit spatial structure and finally the reconstructed signal after thresholding using the universal threshold. We see a good reconstruction of the original signal

using λ_{UNIV}, capturing many of the high peaks and various features that seem to capture true signal. We observe a higher degree of details being retained, compared to the thresholding approach using spatial information. Overall, the universal threshold seems more conservative, resulting in less compression capability but higher signal fidelity. We make two observations:

(1) On the one hand we see some features in the spectrum that resemble real peaks that are not retained with the use of global spatial information. Either these features are due to true signal and have been wrongfully discarded, meaning that (a) we need to make the spatial thresholding more stringent by lowering the selection thresholds or (b) these features did not exhibit enough spatial structure to be picked up by the spatial thresholding method. The second option is that these features are in fact noise and have been wrongfully retained by the universal threshold.

(2) On the other hand we see a higher retention of very short and narrow peaks. These seem to be noisy features that do not appear to be due to a real biological signal and thus the universal threshold may be too conservative. The unnecessary retention of these features contributes to the lower observed compression rates, which could be improved using spatial information.

Unfortunately it is difficult to test which of the foregoing hypotheses is correct, and how much data is lost or retained since we have no ground truth in these IMS data. The ground truth is a set of objective measurements that are assumed to be true and can thus be used to verify the correct operation of an algorithm. This is similar to the concept of the gold standard often used in medicine and statistics. Future work should therefore include a comparison of these methods on an artificial dataset that simulates IMS data (and which can be considered the ground truth as we generate the data ourselves), in order to test the true level of compression, and exactly how many peaks are lost in the analysis.

Residual spatial structure Given the higher conservation exhibited by the universal threshold and the slightly lower compression rates compared to using spatial information to retain detail coefficients, we were interested to see whether the reduced compression would allow for quasi-complete extraction of the biological information from the detail coefficients. We therefore performed PCA analysis on the residual detail coefficients that were not selected for retention, and where thus discarded as noise, to see whether these residual detail coefficients still exhibited spatial structure. Figures 2.14 to 2.16 show the results of the PCA analysis for DWT levels 2 to 4 respectively, where we can clearly see that there are still features left in the discarded data that clearly

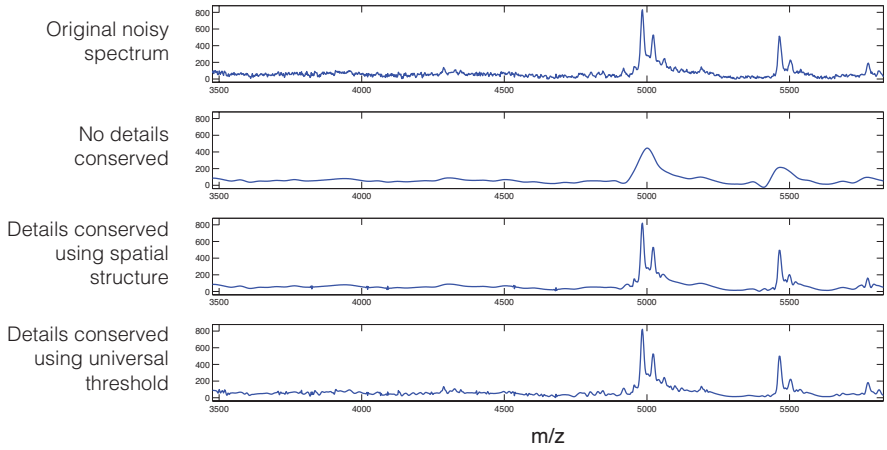

Figure 2.13: Comparison of the original spectrum, reconstructed spectrum without detail coefficients, reconstructed spectrum using spatially conserved features and finally the reconstructed spectrum using the universal threshold. We see a good reconstruction of the original signal using λ_{UNIV}, capturing many of the high peaks and various features that seem to capture true signal. We observe a higher degree of details being retained, compared to the thresholding approach using spatial information. Overall, the universal threshold seems more conservative, resulting in less compression capability but higher signal fidelity.

express spatial structure, meaning that not all biologically relevant features have been extracted. These features are even present in the discarded details for the higher levels of the DWT analysis, which, as we have seen, represent large and broad peaks in the original mass spectrum, not just potentially noisy features. The first level of the DWT after thresholding did not contain any visible spatial structure.

When we look at the score plots we see that these have a very different structure compared to those in Figures 2.8 to 2.11 and seem to be much noisier. Additionally, the Pareto plots show a percentage of variation explained by the first 6 PCs around 11%, 18% and 22% for DWT levels 2 to 4 respectively. This is much lower than what we observed previously, but still much higher than what one would expect in the case of a random distribution. When considering these PCA results, we must take one important notion into consideration. PCA will detect groups of correlating wavelet coefficients, and by removing wavelet coefficients in individual spectra as we do with the universal threshold, these correlation patterns will be perturbed, making it much harder for the PCA analysis to extract this information. For example let us assume we have two

features, feature 1 and feature 2, which always occur together in the spectrum in the same relative quantities and thus have a very high correlation. Now let us assume that the intensity of feature 1 is relatively low and consequently has a low signal to noise ratio. This means that feature 1 will be discarded as a noisy feature by the universal threshold in some of the spectra, while feature 2 is not discarded anywhere. After the thresholding, we will now have spectra where feature 2 is present, but feature 1 is not. When we now calculate the correlation between these two features over all spectra, the correlation will be lower than than before applying the universal threshold. This means that the percentages provided by the Pareto plots may be misleading and that there may in fact be more information left after the thresholding than we can see from this analysis. The noisy score plots are presumably the result of this effect together with a higher amount of information already being extracted through use of the universal threshold.

From the foregoing analysis we can conclude that, despite the larger number of features retained with the universal threshold, this method is not able to extract the full biological information. How much of the true signal is lost, remains unclear due to the lack of a ground truth (or gold standard), and towards the future this should be investigated using an artificial dataset. The universal thresholding is performed on individual mass spectra, which has the advantage that it is computationally 'cheap' (calculating the median of the absolute deviations for a single mass spectrum) and fast, but has the disadvantage that it loses access to the spatial information leading to an incomplete analysis and an overly conservative retention of some of the noisy components. The PCA analysis on the other hand can be computationally expensive at the lower levels of DWT. At the first level of the DWT, the datasize and the number of variables is only reduced by half, meaning that this would still be very computationally intensive for large datasets (100's of GBs). Additionally, the use of global information has the risk that very localized structures could be lost. Therefore perhaps a hybrid approach between the two techniques would be optimal, such that global information is incorporated at the deeper levels of the DWT, together with the universal threshold allowing retention of localized features. Depending on the computational power available (for example in the case of analyzing the data on a desktop computer), the universal threshold, or another wavelet thresholding technique such as minimax thresholding [70], could be applied at the lower DWT levels with the knowledge that much of the noise is located in these levels.

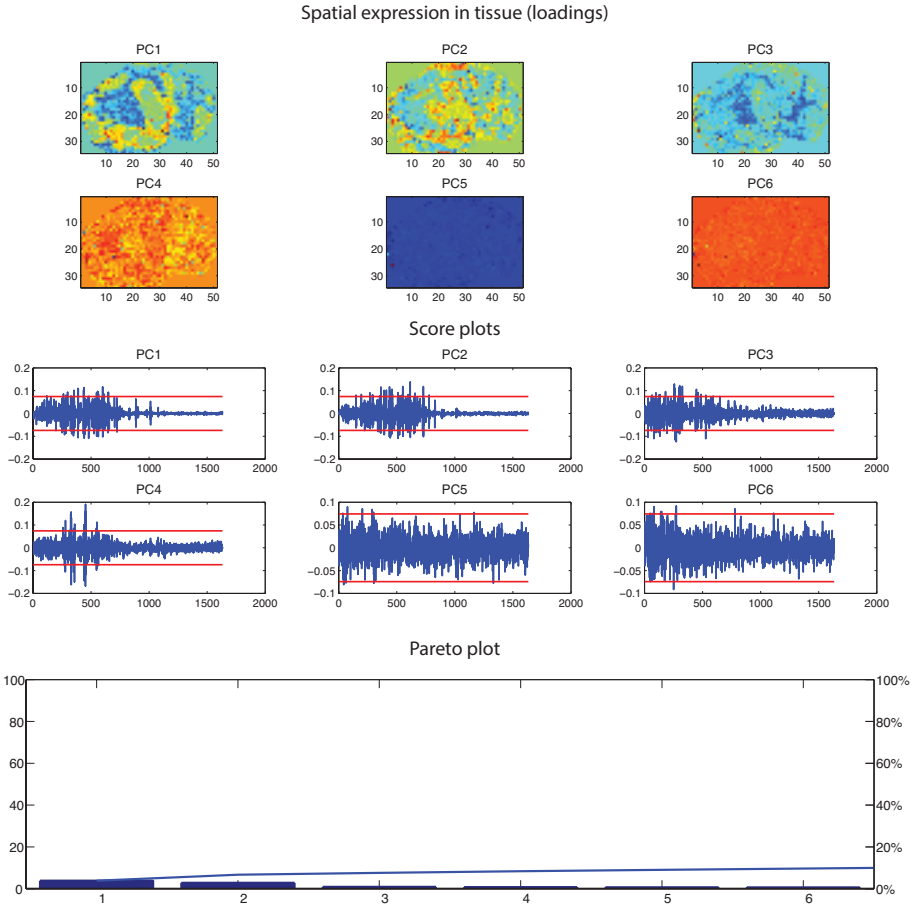

Figure 2.14: PCA analysis results of the DWT level 2 detail coefficients after applying the universal threshold. (Top) Spatial expression of the first six detail coefficient principal components. There are still clearly features left that express spatial structure, meaning that not all biologically relevant features could be extracted using the universal threshold. (Middle) Score plots corresponding to the principal components shown above, showing the correlating features. The red lines indicate the threshold employed for coefficient retention. (Bottom) Pareto plot showing the amount variance in Dc_2 explained by its first six principal components.

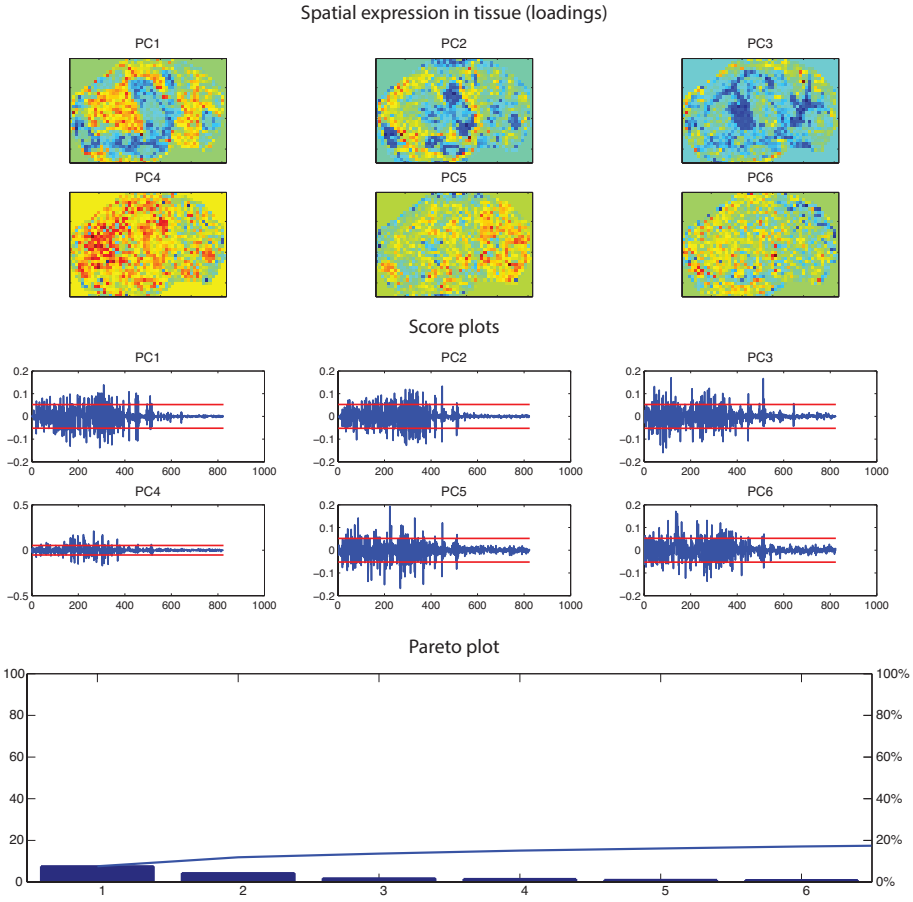

Figure 2.15: PCA analysis results of the DWT level 3 detail coefficients after applying the universal threshold. (Top) Spatial expression of the first six detail coefficient principal components. There are still clearly features left that express spatial structure, meaning that not all biologically relevant features could be extracted using the universal threshold. (Middle) Score plots corresponding to the principal components shown above, showing the correlating features. The red lines indicate the threshold employed for coefficient retention. (Bottom) Pareto plot showing the amount variance in Dc_3 explained by its first six principal components.

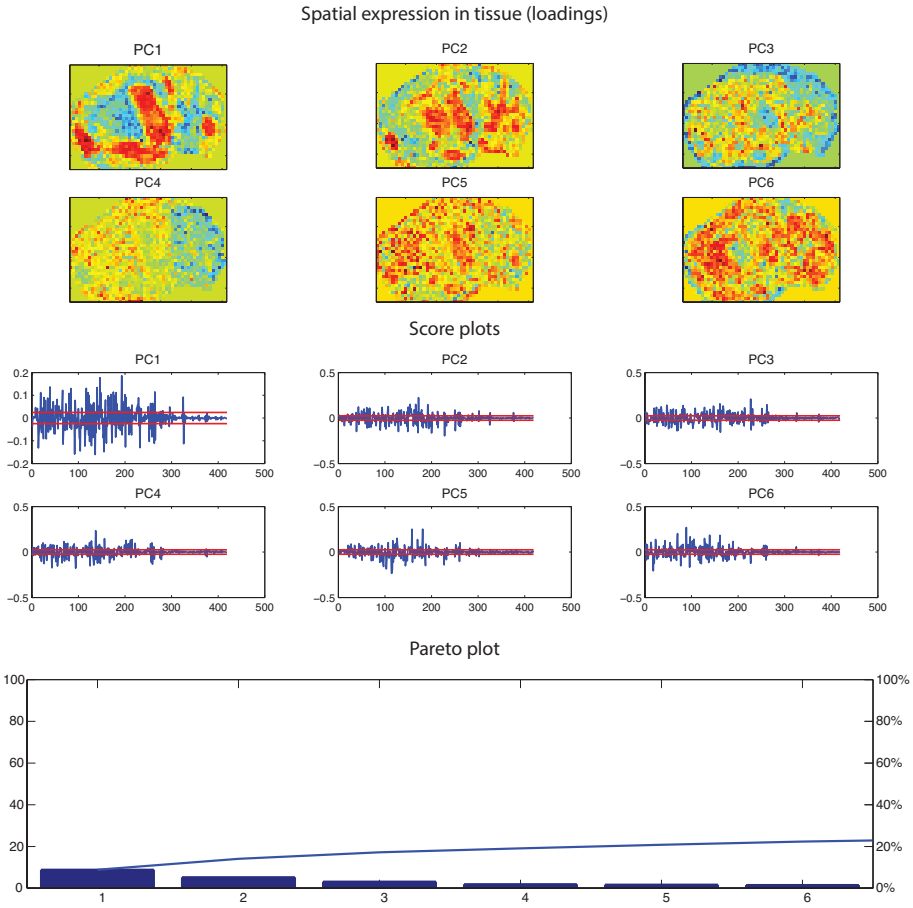

Figure 2.16: PCA analysis results of the DWT level 4 detail coefficients after applying the universal threshold. (Top) Spatial expression of the first six detail coefficient principal components. Even at the highest level of the DWT, which as we've seen represent broad peaks, there is still a lot of spatial structure left in the discarded features. (Middle) Score plots corresponding to the principal components shown above, showing the correlating features. The red lines indicate the threshold employed for coefficient retention. (Bottom) Pareto plot showing the amount variance in Dc_4 explained by its first six principal components.

2.4.3 Incorporation of global spatial information: conclusions

We have demonstrated that incorporation of global spatial information into the wavelet selection process can help to discriminate genuine tissue signal details apart from noise details. We show that by retaining a small number of detail coefficients that express spatial structure we can strongly improve reconstruction of the mass spectra while still achieving considerable compression (10-fold vs. 15-fold).

We compare our approach to a standard wavelet selection approach, namely wavelet thresholding using the universal threshold, which operates at the level of a single mass spectrum. This approach showed good reconstruction of the original mass spectrum, but retained a higher amount of wavelet coefficients, leading to lower compression rates. Despite this increase this higher retention rate, the residual wavelet coefficients still exhibited spatial structure, showing incomplete retrieval of the true biological signal, and thus that an incorporation of spatial information could further improve this standard approach.

For the future, it would be useful to compare these techniques using an artificial dataset, which can provide a ground truth. This way we can investigate how much of the signal is retained versus how much is removed and what compression rates can be achieved when the goal is a near-perfect reconstruction of the original, noiseless signal. A second vector for further research is the automated determination of the thresholds used in the selection of components showing spatial structure. Currently, these are largely determined empirically; the number of PCs included and the height of the thresholds per DWT level were selected based on manual inspection of the PCA plots. Ideally these parameters would be optimized automatically for a dataset, depending on the application goal. For example, for long-term storage we would aim for an optimal reconstruction of the original data or retention of at least $x\%$ of the information in the data. By automating this parameter setting, we can speed up the analysis process and prevent a possible user-bias in setting the thresholds.

In the next Section we present related work on the incorporation of local spatial information into the selection of wavelet coefficients, rather than global spatial information.

2.5 Wavelet coefficient selection using localized spatial information

This section discusses early work on a second method for incorporating spatial information into the DWT analysis, namely through the use of localized spatial information. A disadvantage of incorporating global spatial information in the way we demonstrated is that, especially for the shallower levels of the DWT analysis, the PCA analysis needs to be performed on a relatively large number of wavelet coefficients. For the first level of the DWT, this is half the size of the original data, which may require prohibitively large amounts of memory for large datasets. One alternative is discarding the details of the first few levels, or performing a threshold-based selection using the universal threshold per spectrum as previously discussed. Alternatively, we can use a more localized approach to incorporate the spatial information into the analysis, which we will discuss below. This approach can be performed using a limited number of spectra, and can be expanded depending on the available computational resources.

The following methodology was originally developed with the goal of incorporating localized spatial information into the peak picking process of IMS data, and will therefore focus more on application of the DWT analysis towards this goal. However, it can equally be applied for the improved selection and retention of DWT detail coefficients (for data compression, dimensionality reduction and denoising), analogous to the way we used global spatial information in our previous approach.

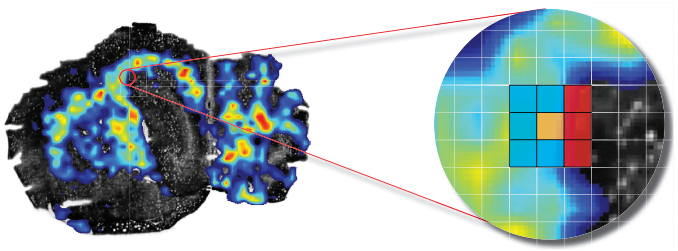

Figure 2.17: Overlay image of a principal component resulting from a PCA analysis. PCA groups together pixels with similar biochemical content. Very often a pixel will be, at least in part, surrounded by pixels with a similar mass spectrum. In the zoom-in on the right hand side we see a target pixel (orange) which is surrounded by 5 pixels which exhibit a similar biochemical content (blue) and 3 pixels which do not share this biochemical content (red). This information can be used to improve feature retrieval.

The local spatial information selection method is based on the observation that an IMS pixel will typically be, at least partly, surrounded by pixels with similar mass spectra, as shown in Figure 2.17. By comparing a target pixel with these similar neighboring pixels, we can detect which features in the spectrum are true biological signal, and which ones are due to noise. Several other studies have shown the advantages of using localized spatial information in the processing of IMS data. Alexandrov *et al.* [52] have previously proposed an edge-preserving de-noising approach based on localized spatial information using estimations of local noise levels. This edge-preserving de-noising operates on individual ion images (m/z-bins), whereas our method compares all m/z-bins of multiple pixels simultaneously. In a more recent study, Alexandrov *et al.* [71] have presented a spatially aware clustering method, that uses local spatial information of neighboring pixels to improve clustering of IMS data. Incorporation of the spatial information in this way led to more continuous regions in the spatial segmentation results.

Here we will use this localized spatial information to compare and de-noise pixels using the wavelet coefficients obtained after DWT decomposition. Performing this comparison in the wavelet space rather than in the original IMS space offers several advantages. Firstly, the DWT allows us to split the signal into several levels, with each level containing content of similar frequencies. This allows us to treat each of these levels differently, since most of the noise is located in the shallower levels of the decomposition, as we have discussed previously. Additionally, the data become sparser at higher levels of the wavelet decomposition, making it easier to compare the spectra in wavelet space. Finally, after the analysis, the resulting spectra can immediately be stored in a data compressed format, only retaining relevant wavelet coefficients.

2.5.1 Workflow

The localized wavelet de-noising algorithm for a single target pixel is shown in Figure 2.18. This operation is repeated for each pixel in the image.

1. We first select a target pixel (and implicitly its neighboring pixels). Depending on the memory requirements, a 3×3, a 5×5, or an even larger window of pixels can be chosen, with a higher number of pixels giving higher confidence in the retrieved features, but also being more computationally intensive. For the present study we will work with the minimum window of 3×3, meaning we incorporate the target pixel (T) and its 8 surrounding pixels ($P1 \cdots P8$) into the analysis.

2. We then perform a 5-level deep DWT analysis using the Sym8 wavelet on the spectrum of each of these pixels separately. (In practice we store the DWT analysis for each pixel, as otherwise this operation would need to be repeated several times for the same pixel).

3. Next we compare wavelet coefficients of the target pixel to each of its surrounding pixels and this for each DWT level separately. We compare the wavelet coefficients at two levels, global and specific:

Global. We first examine the global similarity of the target pixel and its neighboring pixels at the selected DWT level. For this we calculate the correlation coefficient over the whole set of wavelet coefficients for the target pixel and each of its neighboring pixels separately. We store the result in a correlation coefficient $CC_{L,P}$, where L is the DWT level and P is the index of the neighboring pixel (for example, at the first DWT level and calculating the similarity for neighboring pixel 8 we get $CC_{1,P8}$). This global correlation factor determines the impact of a neighboring pixel on the scores of individual features (wavelet coefficients). If a neighboring pixel has a high global correlation with the target pixel, it is very similar and there is a high chance that similar individual features will return, which should give it a bigger impact on the scoring of these features, and vice versa.

Specific. We then perform a finer comparison of the target pixel and its neighboring pixels at the selected DWT level. For this we compare wavelet coefficients between the target pixel and each of its neighboring pixels separately using a sliding window of size 5. This sliding window moves along the wavelet coefficients of both pixels simultaneously. Using this window, we calculate the correlations between sets of 5 consecutive wavelet coefficients for both pixels. For each of these sets, the score of the centermost feature or wavelet coefficient is determined by the correlation over these 5 wavelet coefficients. [3] The correlation coefficient for feature k at level L and pixel P is then calculated as

$$cc_{L,P,k} = Corr(Dc_{L,T,k-2\ldots k+2}, Dc_{L,P,k-2\ldots k+2}) \tag{2.4}$$

where $Dc_{L,P,k}$ represents the detail wavelet coefficient k at level L for pixel P. We chose to make use of a window because it offers additional confirmation on the presence of a feature if the surrounding wavelet coefficients show a similar behaviour as well. The use of correlation has the advantage that it does not take into account the absolute size of the features.

[3] This is similar to what one would do in, for example, a simple moving average (SMA) filter with a window of 5, where the $SMA_k = (p_{k-2} + \cdots + p_{k+2})/5$ with p a data point.

4. In the final step of the algorithm, we calculate the confidence score for each feature using the global similarities per level and the specific similarities per wavelet coefficient over all neighboring pixels. The confidence score cs_k for each individual wavelet coefficient k at level L is calculated as:

$$cs_k = \sum_{P=P1}^{P8} cc_{L,P,k} CC_{L,P} \tag{2.5}$$

If a feature is supported by the surrounding pixels, it will have a high score. Features with a score above a set threshold (red line in figure) are retained (marked in green in figure), the rest is discarded. The threshold can be selected freely per level of the DWT and determines how strict the selection will be. We use a threshold set at 0.40 for all levels of the DWT as this empirically showed good results.

2.5.2 Artificial dataset

We illustrate the performance of our feature selection method on a computer-generated artificial dataset. We construct this artificial dataset to approximate a real-life IMS dataset, simulating a tissue that contains several distinct "tissue regions", where each of these tissue regions has its own characteristic mass spectrum. Previous studies such as that by Hanselmann et al. [72] have previously used simulated datasets to benchmark algorithms for the computational analysis of IMS data, and our artificial dataset is generated in a similar way.

Figure 2.19 shows the spatial layout that was used to construct the artificial dataset. The dataset consists of 3 overlapping areas, representing the different tissue regions, where each colored area contains pixels with the corresponding spectrum shown on the right. The dataset simulates a 40×40 raster of which 1450 pixels contain spectra and where each spectrum has 6490 simulated m/z-bins in a mass range from m/z 2800 to 25000. Pixels that are located in regions where the areas overlap contain mixtures of the corresponding characteristic spectra. These mixtures are created by summing the characteristic spectra of the overlapping areas. Each of the characteristic spectra is constructed to contain 40 very small to large peaks, randomly distributed over the m/z-range. The peaks are modelled as Gaussian probability distribution curves. Once all these different spectra are mixed together, we add Gaussian noise to the mass spectrum of each individual pixel. Finally, the spectra are normalized using total ion current (TIC) normalization.

While this dataset is only an approximation of a real-life IMS dataset, the use of an artificial dataset has the advantage that we have a ground truth, namely

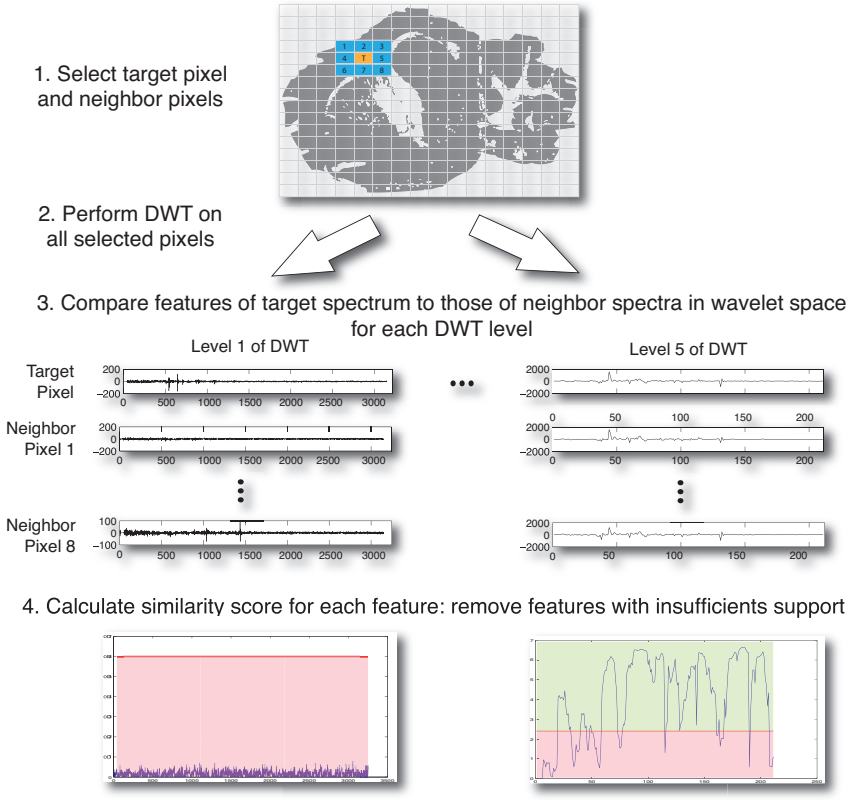

Figure 2.18: Workflow of the wavelet coefficient selection using localized spatial information.

spectra with known peaks, to compare the results to, which is not the case with real tissue measurements.

2.5.3 Results

Application of our localized wavelet de-noising strategy to the artificial dataset takes less than a minute on a standard laptop. The algorithm exhibits low memory usage as the DWT analysis can be performed on each spectrum separately, and only nine spectra need to be operated on concurrently. After the wavelet selection is complete, the spectrum is reconstructed with the retained wavelet coefficients using the IDWT. Figure 2.20 shows the comparison between

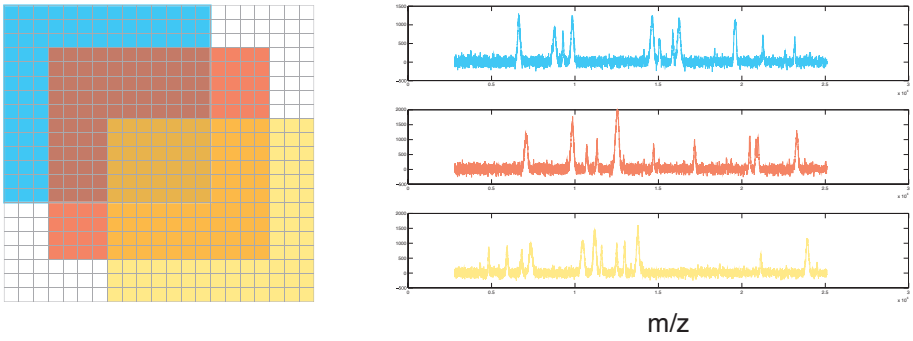

Figure 2.19: The artificial dataset used to benchmark our method. Each colored area contains pixels with the corresponding spectrum on the right. Overlapping areas contain mixes of the corresponding spectra.

the noisy spectrum present in the artificial dataset, the reconstructed spectrum after wavelet selection using the spatial information of the surrounding pixels, and the ground truth spectrum before adding noise. Despite the high levels of noise that were added to the spectrum, we can see that, except for one very small peak around 1500 m/z, we can retrieve all peaks that are present in the original, noiseless spectrum. At this point in the methodology, we have essentially reached the same type of results as those in the previous Section (Section 2.4) on the use of global spatial information: we have used the localized spatial information to discriminate between wavelet coefficients that exhibit spatial structure (and are thus presumably due to true biological variation) and those that do not (and are thus presumably due to noise). The noisy wavelet coefficients were discarded, leading to a reduction in size and dimensionality of the data. Similarly as before, these reduced data can now be used for further data analysis, long-term storage or reconstruction of denoised mass spectra. However, as we have previously noted, this methodology was originally designed to incorporate spatial information into the peak picking process. Therefore, rather than using the reconstructed and denoised mass spectrum as a way of confirming correct retention of wavelet coefficients as we did in the previous setting, we will now go one step further and use this clean and denoised spectrum for a consecutive peak picking step.

In order to benchmark the localized wavelet de-noising strategy, we perform peak picking on both the original noisy spectrum and the reconstructed spectrum. We compare the retrieved peaks with those used to construct the original data. This allows us to see how many of the peaks in the original noiseless spectrum are retained and how the noise would normally affect feature extraction using peak picking. We perform peak picking using the `mspeaks` function in MATLAB

	Standard peak picking	Localized spatial information
False negative peaks	45.4	3.6
False positive peaks	2.1	7.0
True positive peaks	145.7	187.5

Table 2.3: Comparison of the numbers of false negative, false positive and true positive peaks between standard peak picking on the noisy spectra and the reconstructed spectra after wavelet selection. All numbers in the table are the average per spectrum over all spectra.

on each individual mass spectrum. For the noisy spectra we use the default de-noising setting in the `mspeaks` function, for the reconstructed spectra we disable de-noising. This de-noising is done by wavelet thresholding, similar to the use of the universal threshold, discussed in the previous Section. The results of the peak picking are shown in table 2.3. As we can see our method greatly reduces the number of false negative peaks per spectrum, meaning that peaks that would have been lost in normal peak picking are saved in our method, which was the primary goal of using the localized spatial information. We do observe a slight increase in the number of false positives but this rate is still only 3.7% of the total number of peaks. This increase in false positives can be due to two reasons. Firstly, we use very conservative settings for the thresholding of the wavelet coefficients, in order to retain as much of the peaks as possible; this will naturally give rise to a higher number of false positives. Secondly, when we look back at Figure 2.20, we can see that reconstruction of the original noiseless spectrum is not perfect, and that some reconstruction artifacts occur around the smaller retrieved peaks, where we see small 'crevasses' on either side of the reconstructed peak. These reconstruction errors could give rise to false positives. Such reconstruction artifacts are often seen in wavelet reconstruction, and occur when a single DWT coefficient is retained and reconstructed while its adjacent coefficients are set to zero. A single wavelet always needs to have an average value of zero, and thus when a single DWT coefficient is retained, we observe these 'crevasses' around the reconstructed peak. Normally the adjacent wavelets compensate for this effect, however these have been discarded in the thresholding step. Towards the future these reconstruction errors could be resolved by using soft thresholding rather then the hard thresholding we currently use. Soft thresholding diminishes wavelet coefficients that are close to the threshold rather then completely replacing them by zeros. By doing so, such singular retained wavelet coefficients will occur much less frequently, and the surrounding wavelets will compensate the negative values, thus eliminating false positives due to these reconstruction errors.

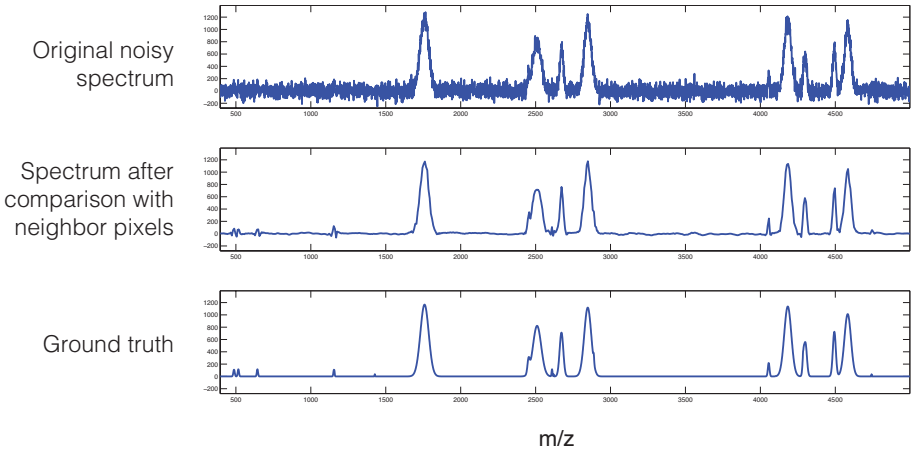

Figure 2.20: Comparison between the noisy spectrum, the reconstructed spectrum after wavelet selection using information from the surrounding pixels and the ground truth spectrum without noise. Despite the high levels of noise added to the spectrum, we can retrieve almost all small peaks originally present in the ground truth spectrum.

2.5.4 Incorporation of localized spatial information: conclusions

In this Section we have shown that we can successfully extract relevant features from IMS data by using localized spatial information. This information was used to compare and de-noise spectra using the wavelet coefficients obtained after DWT decomposition. Use of the DWT has the advantage that the original signal can be split into several frequency levels, allowing for easier comparison of the spectra. Additionally, the de-noised data can immediately be stored in a compressed data format after the analysis. We demonstrated this approach on an artificial dataset, which simulated IMS data. When comparing our novel approach to a standard peak picking algorithm, operating on individual mass spectra, a higher number of true features, present in the ground truth could be retrieved, demonstrating the advantage of using spatial information in the analysis of IMS data.

Further development of this method towards peak picking rather than the storage of reconstructed spectra, should include a comparison with peak picking methods specifically designed for IMS data such as the one developed by McDonnell *et al.* [53]. Next to this work on DWT we have initiated work on the incorporation of localized spatial information into feature selection using the

CWT. This combines previous work on peak picking by Du *et al.* [62] for regular mass spectrometry with a workflow incorporating spatial information similar to that presented above. Preliminary results were promising and showed very good retrieval of even very small peaks, with a lower number of false positives and false negatives, however more research is necessary to confirm these findings. The disadvantage of using the CWT is that it is more computationally intensive. Additionally it cannot be used for data compression and reconstruction in the way we used the DWT; rather than returning wavelet coefficients which can be used to store the signal in compressed form, the algorithm returns only the location of features of interest making this more akin to a classical peak picking approach, but one that incorporates spatial information.

2.6 Conclusions & Future perspectives

In this chapter we have introduced the discrete wavelet transform (DWT) as a tool for feature selection and data size and dimensionality reduction in IMS data. We showed that the DWT can be used to greatly compress IMS data, while retaining the original shape of the mass spectra. This constrasts to the most commonly used form of dimensionality reduction in IMS, peak picking, where all of this information is lost. The high compression rates in DWT are achieved by throwing away features in the mass spectrum that are the result of noise and which take up unnecessary data space and hamper computational analysis. Selection of which features are due to noise and which ones are due to true biological signal is not straightforward, and therefore we incorporate one of the key aspects of IMS, spatial information, into this selection procedure. By using this information, we can selectively remove only those details that do not exhibit a spatial structure, and therefore have a high likelihood of being due to noise, while retaining as much of the true biological structure as possible.

We have demonstrated two ways of incorporating the spatial information into the DWT analysis. A first method uses global spatial information, through the use of PCA, to extract features that exhibit spatial structure throughout the entire IMS dataset. A second method uses local spatial information, utilizing the common observation that neighboring pixels have similar mass spectra to a target pixel. Both methods showed a very good reconstruction of the original data, while successfully eliminating noise from the data and improved on feature retrieval using standard methods. Elimination of noisy features leads to both a dimensionality reduction and a data size reduction of the original data, where a decrease in noisy features can lead to improved, less computationally intensive data analysis, while data size reduction allows for more efficient, noiseless storage of the data. Both approaches have their advantages and disadvantages. The

incorporation of global spatial information through use of PCA makes use of the spatial information throughout the dataset but can be computationally expensive at the first levels of the DWT, and runs the risk of losing very local structures. The incorporation of localized spatial information is computationally much lighter due to the fact that a very limited number of spectra is analyzed simultaneously, but loses the global information and can therefore be misled if the target pixel is not surrounded by many similar pixels, but which could still be present throughout the data.

Future research should include a side-by-side comparison of these approaches on both artificial and genuine IMS datasets, comparing computational time, memory requirements, compression rates and reconstruction of the original data. Fusing these methods with standard wavelet selection approaches such as the universal threshold could further improve the performance of these techniques. Furthermore, future research in this domain should focus on the development of automated methods that optimize parameter setting for these algorithms. This would eliminate the need the need to manually set these parameters, thus saving time and eliminating possible user bias.

Finally, as we have previously noted, the datasets used in this work are relatively small compared to current state-of-the-art datasets due to the fact that this work was done in the early stages of the PhD. It is therefore possible that even higher compression rates could be achieved on these newer datasets, as the number of m/z-bins measured in a typical IMS experiment has increased by a factor 100-1000, whereas the number of biomolecular ions typically detected increased far less than that, making these data sparser than before. Research in other hyperspectral imaging domains have shown compression rates of 100 or more using 3D-wavelets [73].

Chapter 3

Differential Analysis of IMS Experiments

This chapter is based on the following internal report:
N. Verbeeck, R. Van de Plas, Y. El Aalamat, J. Yang, R.M. Caprioli, K. Schey, D.M. Anderson, E. Waelkens, and B. De Moor, "Capturing Inter-Subject Variability with Group Independent Component Analysis of IMS Data", *internal report*.

This chapter addresses the comparison between multiple IMS experiments acquired from different tissue sections. Performing such a comparison is nontrivial due to the large heterogeneity, dimensionality, and size of IMS datasets. To accomplish this, we introduce a novel technique that is based on the concept of Group Independent Component Analysis (GICA), a method originating from functional magnetic resonance imaging (fMRI) research. The method allows us to highlight which biomolecules at which locations differentiate one tissue section from another. The work in this chapter was awarded a "best poster award" at the NVMS/BSMS Rolduc Conference 2012, Kerkrade, The Netherlands.

3.1 Introduction

From previous chapters it is clear that IMS makes it possible to monitor the spatial tissue distributions of many hundreds of biomolecules simultaneously, making it a prime technology for exploratory research. As a result, it has

been successfully applied in biomarker discovery studies, drug distribution studies, and metabolism analysis. However, the exploratory advantage of IMS is currently hampered by the large amount of data that its experiments can deliver, making interpretation and analysis difficult. While computational methods have made it feasible to search these experiments for relevant findings [49, 50], this research has been focused primarily on exploration within a single tissue section or IMS experiment. In this work, we focus on the automated discovery of common and differential patterns across multiple tissue sections. This allows us to directly recognize which biomolecules differentiate one tissue sample from another, and at which locations this difference occurs. To accomplish this, we propose a mass spectrometry-oriented adaptation of Group Independent Component Analysis (GICA) as a technique accomplish this.

Several studies have previously undertaken the challenge of computationally comparing data from multiple IMS experiments. However, two main issues arise when trying to perform these types of studies. The first issue that makes direct comparison of samples difficult is heterogeneity, in the sense that most tissue sections contain a multitude of different cell types. Djidja *et al.* [74] have handled this difficulty by manually targeting a particular sub-area of the tissue and selecting out a limited number of spectra from the adenocarcinoma cores under investigation. These spectra were then used to generate a classification model for different tumour classes using Discriminant Analysis. A similar sub-selection approach is used by Veselkov *et al.* [75] for the classification of tissues. However, such targeted approaches are only possible in cases where it is known beforehand which tissue sub-areas, cell types, or features are of interest for the biological problem at hand. Our work is focused on providing a means of automated differential analysis for unsupervised exploration of tissue, where no such prior information is available and tissue sub-selection is thus not an option. The second issue that hampers differential analysis is the high dimensionality of mass spectral measurements and the massive amounts of data that are accumulated by IMS experiments as a result. This issue is further compounded when multiple IMS experiments need to be compared to each other. Jones *et al.* [76] have explored this problem by first performing a peak picking step to reduce the amount of data before differential analysis. While only retaining the peak information is an effective means to keep the amount of data manageable, it has multiple disadvantages. It makes any subsequent analysis fully dependent on the performance of the peak picking method, on the particular peak picking parameter settings that were employed, and on the assumptions inherent to the underlying algorithm. Besides these aspects, many peak picking algorithms have difficulty recognizing genuine features that are present only at low ion intensities or in a limited area of the tissue. The impact of such aspects is often dependent on the particularities of the dataset at hand, and for the moment peak picking largely remains an area under study.

Overall, peak picking significantly reduces dimensionality, but at the price of a nontrivial potential for missed data features. This is why in our differential analysis work we strive to develop methods that avoid the need for peak picking as a preprocessing step.

GICA is an multivariate analysis technique that has proven its use in fMRI research, where it is used to identify functionally connected neuronal networks in multi-subject studies [77]. In these studies, the brain activities of multiple subjects are measured while performing a task, with the aim of finding differences in activation of brain regions. The fMRI data share some aspects with IMS data in that both tend to consist of large datasets where a complex signal (a magnetic resonance spectrum in fMRI and a mass spectrum in IMS) is measured at multiple spatial locations (voxels in fMRI and pixels in IMS). Other aspects such as the corresponding mode between datasets (spatial mode in fMRI versus spectral mode in IMS) are fundamentally different between both data types. As a result, some adaptation of the method is essential to enable application in an IMS setting. In this work, we develop a version of GICA that is adapted to the aspects of IMS data, and we employ the method to accomplish automated differential analysis of IMS experiments.

When presented with the IMS measurements of several tissue samples (e.g. diseased versus control tissue), GICA uses a combination of principal component analysis (PCA) [67] and independent component analysis (ICA) [78] to discern the various patterns underlying the tissue data. It then assesses the extent to which these patterns are observed in each of the tissue sections under study, effectively highlighting common and differential signatures. The technique reports the patterns that are common across all tissue sections, as well as the patterns that only show up differentially in a particular tissue section or tissue class. Each of these patterns is characterized by their mass spectral signature as well as their spatial distribution throughout the tissue of all presented samples.

This chapter first describes the GICA methodology and our adaptations in detail. The algorithm is then applied to an artificially generated IMS dataset to test the correct operation of the technique, and applied to two genuine biological IMS datasets, one of which is a case study on mouse spinal cord and the other studies retinal degeneration disease in mouse retina.

3.2 Methods

The GICA algorithm consists of three steps: sample-specific PCA, group-wide PCA and group-wide ICA. The workflow of the GICA algorithm is shown in figure 3.1. If the goal is to compare N individual IMS experiments, the inputs

to the method consist of N individual IMS datasets, each provided in a tabular or matrix form. In such a data table, each row corresponds to a distinct tissue location or pixel and represents a full mass spectrum, while each column encodes ion intensity corresponding to a particular m/z value across all measurement locations or pixels. For the remainder of this chapter we will use the following notation to describe the IMS data: let Y_i be the $P_i \times M$ matrix representing the IMS data for a single tissue sample $i \in 1, \cdots, N$ with N the total number of samples to compare, P_i the total number of pixels for dataset i and M the same number of m/z bins provided by all N datasets, then

$$Y_i = \begin{bmatrix} S_{i,1} \\ \vdots \\ S_{i,P_i} \end{bmatrix} \tag{3.1}$$

where $S_{i,j}$ represents a mass spectrum of length M, acquired from pixel j in tissue sample i with its row-wise mean subtracted. The mean is subtracted since these data will be used in PCA analysis in the following steps. The number of m/z-bins M and measured mass range need to be the same for all provided datasets. If this is not the case, the mass spectra can be resampled to a common m/z-bin vector to accomplish this.

The following sections describe each of the steps taken in the GICA algorithm in detail.

3.2.1 Sample-specific PCA

The data of each tissue sample is first analyzed individually by applying a separate PCA on each dataset. This step extracts the tissue sample-specific principal components that can be found in the mass spectral measurements of each IMS experiment. PCA is an established multivariate dimensionality reduction technique [67] that uses orthogonal transformation to transform a set of potentially correlating variables into a set of uncorrelated variables called the principal components (PCs). PCA is defined such that the first PC explains the largest possible variance in the original data. Each subsequent PC is orthogonal to the previous PCs and explains the largest possible variance that remains in the data after the preceding PCs have been removed from the original data. Most variance in the dataset is captured by the earlier PCs and later PCs report consistently decreasing variance and impact on the data. As a result, a small subset of the initial PCs is sufficient to deliver a close approximation of the original dataset, while retaining only that small subset constitutes a substantial dimensionality reduction.

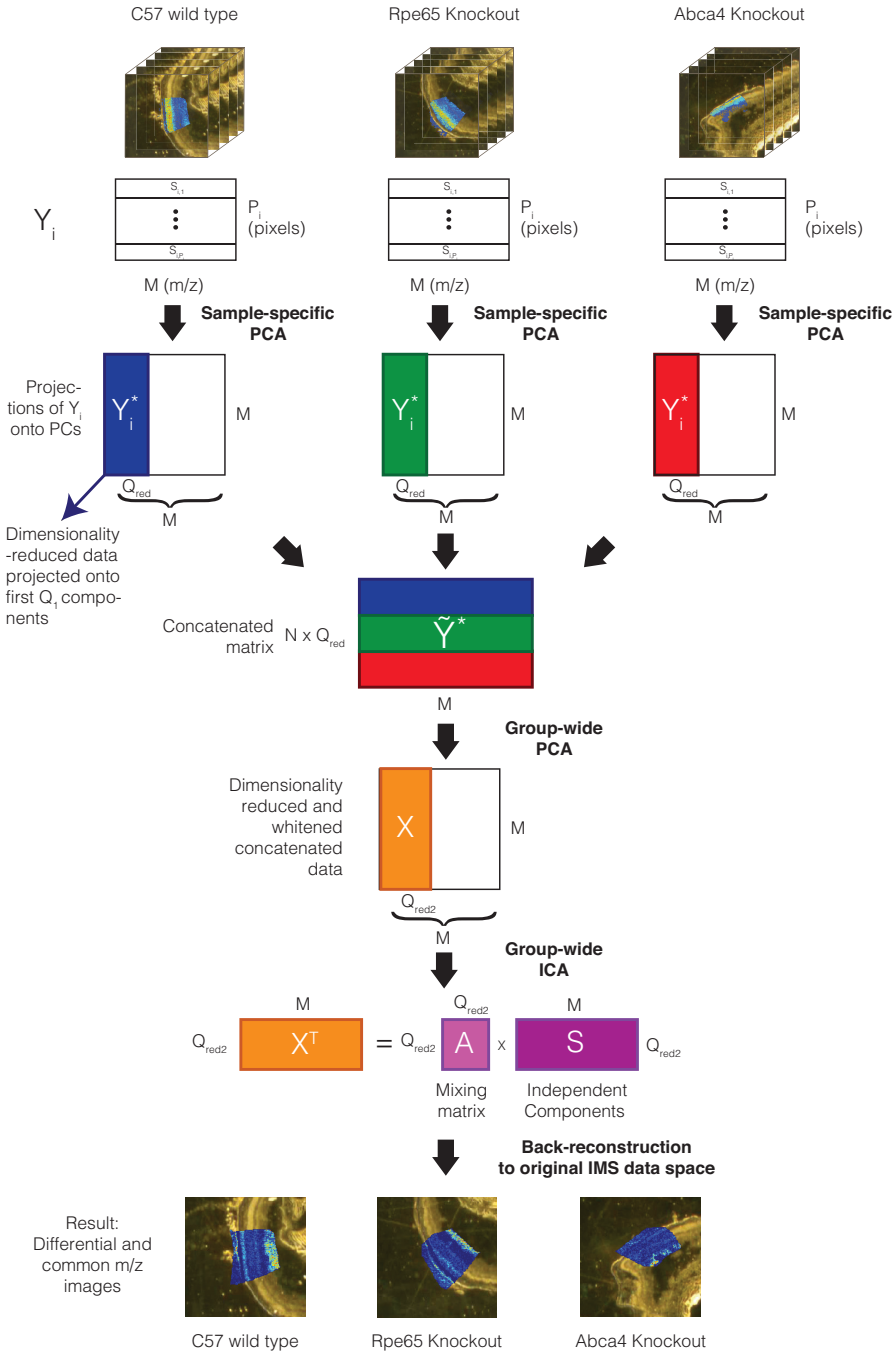

Figure 3.1: Workflow of the GICA algorithm, which consists of three main steps: a sample-specific PCA, a group-wide PCA and a group-wide ICA.

We can calculate the principal components for a single sample Y_i through its singular value decomposition (SVD), namely $Y_i = U_i \Sigma_i V_i^T$, where Σ_i is a $P_i \times M$ matrix which contains only non-zero elements on the diagonal called the singular values, U_i is a $P_i \times P_i$ orthogonal matrix containing the left singular vectors of Y_i and V_i is an $M \times M$ orthogonal matrix containing the right singular vectors of Y_i. By convention, the singular values are sorted from high-to-low, determining the order of the singular vectors. Using the SVD, we can now write the covariance matrix of the pixels of Y_i as $C = \frac{1}{1-n} Y_i Y_i^T = \frac{1}{1-n} U_i \Sigma_i^2 U_i^T$. The columns of U_i then provide us with the PCs in the pixel space of Y_i [79]. Using this result we can reduce the dimensionality of our original dataset by projecting Y_i onto the first Q_{red} principal components contained in U_i as $Y_i^* = U_{i,(Q_{red})}^T Y_i$ resulting in the $Q_{red} \times M$ reduced matrix Y_i^*. Let us define $U_i^- = U_{i,(Q_{red})}^T$ as the $Q_{red} \times P$ reducing matrix for tissue sample i for the remainder of this manuscript.

We note that in most work on GICA in fMRI research, the mean is subtracted from each column of Y_i and the decomposition of the covariance matrix of the spectra Y_i is calculated as $\frac{1}{1-n} Y_i^T Y_i = \frac{1}{n-1} V_i \Sigma_i^2 V_i^T$, in which case V_i provides the principal components in the spectral space. Projection onto the first M_{red} principal components then yields the $M_{red} \times P_i$ reduced matrix $\tilde{Y}_i^*$. However, unlike in fMRI data, the data measured in IMS are generally not spatially aligned to each other. This means that in IMS data the xth pixel in dataset A does not necessarily represent a corresponding tissue location and content as the xth pixel in dataset B. This formulation where the matrices $\tilde{Y}_i^*$ are compared across multiple datasets can therefore not be universally applied to an IMS setting. In chapter 4 we focus on the spatial registration of IMS data to a common reference space. For such registered data a comparison of the pixel content would be feasible, and towards the future it would be interesting to compare results between both formulations.

When we now return to the derivation of the reduced matrix $Y_i^* = U_{i,(Q_{red})}^T Y_i$, and fill in $Y_i = U_i \Sigma_i V_i^T$ into this equation, we get

$$\begin{aligned} Y_i^* &= U_{i,(Q_{red})}^T U_i \Sigma_i V_i^T \\ &= I_{(Q_{red})} \Sigma_i V_i^T \\ &= \Sigma_{i,(Q_{red})} V_{i,(Q_{red})}^T \end{aligned} \tag{3.2}$$

The reduced matrix Y_i^*, which we will use in the remainder of this work is thus equivalent to the multiplication of the first Q_{red} singular values $\Sigma_{i,(Q_{red})}$, which are scalar values, by the first Q_{red} principal components in spectral space, $V_{i,(Q_{red})}$.

Determining how many PCs need to be retained in order to accurately capture the original data is traditionally a difficult problem, and we further elaborate on this in section 3.2.3. Although it is possible to select a different number of PCs to retain for each tissue sample, for the present analysis we will retain the same number of components Q_{red} for all tissue samples.

PCA has been applied previously in IMS research for both pattern extraction and dimensionality reduction of both single [65,80] and multiple tissue datasets [76]. In the context of IMS, the retrieved PCs group together ions that correlate and co-localize within the tissue, and thus form a particular spectral and thus biomolecular signature within the measured spectra. Performing PCA at the sample-level encourages the extraction of small and dataset-specific components. A biomolecular pattern present only in a relatively small number of pixels might not be captured if PCA would be immediately applied to the combination of all datasets, but performing PCA at the single tissue sample level first significantly reduces the chance of this happening. Doing a sample-specific PCA per tissue sample essentially boosts the detectability of small signal patterns by the overall differential analysis, and avoids that only high-abundant signatures reported by many tissue samples are recognized.

A second important advantage of preceding any analysis across multiple tissue samples with a sample-specific analysis is that it strongly decreases the computational resources and computation time necessary to perform the overall differential analysis. The sample-specific PCA reduces the dimensionality of each individual IMS experiment, significantly reducing computational resources such as memory and processing power when analysis across datasets is pursued. Also, the cost of performing PCA on additional tissues increases linearly at the sample-specific level rather than quadratically at the group-wide level and only one dataset needs to be loaded into memory simultaneously, making it possible for sample-specific PCA analyses to be performed in parallel for different tissue samples.

3.2.2 Group-wide PCA

Once we have performed PCA on each individual dataset, we can obtain dimensionality reduced versions Y_i^* of each sample dataset Y_i. These dimensionality reduced datasets of size $Q_{red} \times M$ are then combined into a single matrix $\tilde{Y}^*$ of size $NQ_{red} \times M$,

$$\tilde{Y}^* = \begin{bmatrix} Y_1^* \\ \vdots \\ Y_N^* \end{bmatrix} \tag{3.3}$$

with M the original number of m/z bins. On the matrix $\tilde{Y}^*$ a second "group-wide" PCA is performed to further reduce the dimensionality of the aggregated data, but in this case all tissue samples are taken into account concurrently. Analogously to the tissue sample-specific PCA, we can now extract the principal components, which we store in the orthogonal matrix G. We can then project the aggregated data onto the first Q_{red2} principal components, for which we define the reducing matrix $G^- = G^T_{(Q_{red2})}$ of size $Q_{red2} \times NQ_{red}$. Projection onto the Q_{red2} principal components is then equal to multiplying the matrix $\tilde{Y}^*$ by the reducing matrix G^-, in order to receive the reduced $Q_{red2} \times M$ matrix X:

$$\begin{aligned} X = G^- \tilde{Y}^* &= \begin{bmatrix} G^T_{1,(Q_{red2})} \dots G^T_{N,(Q_{red2})} \end{bmatrix} \begin{bmatrix} Y^*_1 \\ \vdots \\ Y^*_N \end{bmatrix} \\ &= \begin{bmatrix} G^T_{1,(Q_{red2})} \dots G^T_{N,(Q_{red2})} \end{bmatrix} \begin{bmatrix} U^-_1 Y_1 \\ \vdots \\ U^-_N Y_N \end{bmatrix} \end{aligned} \tag{3.4}$$

where $G_{i,(Q_{red2})}$ is then the part of the group-wide principal components pertaining to tissue sample i.

3.2.3 Estimation of number of components

Determination of the correct number of components to retain after a PCA analysis is an important but difficult challenge to tackle. Several papers have been dedicated to the subject [81] and as mentioned by Jolliffe [67], no straightforward solution to the problem exists. However, heuristic methods can offer a solution. Previous studies that apply PCA on IMS data have recognized this challenge [82] [83]. In the GICA-algorithm, two component selection steps are necessary; one at the tissue sample-specific level and one at the group-wide level. Allen *et al.* [77] states that component selection at the subject level, or in our case the tissue sample-specific level, is less crucial than component retention at the group-wide level. In fact, an overestimation of the number of components retained at the sample-specific level can improve detection and reconstruction of sample-specific signals. To make sure no information is lost early on in the analysis, an over-estimate of the number of retained sample-specific components is favored over an underestimate. Since the retention of more sample-specific components will make subsequent computation of the group-wide PCA more challenging, a compromise needs to be made here, particularly when comparing

large numbers of tissues. At the group-wide level, correct determination of the number of retained components is very important since this will also dictate the number of components that the final ICA algorithm will attempt to discern. Allen *et al.* [77] have shown that an overestimation of the number of components in this step can decrease the quality of the final retrieved biomolecular signals due to the ICA algorithm splitting informative sources over multiple components. It is clear that determination of the optimal number of principal components to retain is an important step and for this reason, we specifically evaluate several methods that can help find this optimal component number. The first method we examine is the scree plot [67]. It is one of the easiest methods and consists of plotting the eigenvalues of the retrieved components in descending order, and selecting a cut-off at the point where there is no longer a significant change in the value of two conseqtuential eigenvalues.

The second method is the minimum description length (MDL) [84] method. This method is employed by Calhoun *et al.* [85] for estimation of the number of components to retain in fMRI data, and it has been previously used in the context of PCA [86]. MDL is an information-theoretic meausure for model selection, and is closely related to the Akaike Information Criterion which has been used previously in the context of IMS [82] for estimating the number of components for probabilistic latent semantic analysis. It is calculated [87] as

$$E_{MDL}(k) = -L(\mathbf{x}|\Theta_k) + \frac{1}{2}G(\Theta_k)\, logN \tag{3.5}$$

where $L(\mathbf{x}|\Theta_k)$ is the maximum log-likelihood of the observations, $\mathbf{x}$, based on the model parameter set Θ_k of the kth order and $G(\Theta_k)$ is the penalty for model complexity given by the total number of free parameters in Θ_k. For MDL, the penalty term is scaled by $logN$, where N is the sample size. An important caveat in the context of component selection using MDL, and PCA in general for that matter, is that these methods assume independent measurements which is not the case in fMRI data nor in IMS data as there commonly exists a nontrivial spatial correlation between neighboring fMRI voxels and IMS pixels. Failure to meet this assumption can influence the component estimation, and therefore Li *et al.* have developed an fMRI-specific technique [87] to make the samples approximately independent, which takes into account the specific properties of fMRI data. A similar framework could be developed for IMS as well, but such a development effort falls outside the scope of this work.

The third component selection method we examined is the Laplace method suggested by Minka *et al.* [88]. This well-cited method interprets PCA as a density estimation problem and uses a Bayesian model selection to determine the number of components to retain.

A final selection method uses data reconstruction to the original measurement space, making it possible to evaluate PC selection interms of domain-specific influence. This method gauges the influence of including an additional component at the group-wide PCA by reconstructing the IMS data using different numbers of components, and then plotting the change in absolute ion intensity that is contributed by each added components. Although this method is very calculation intensive, it provides the most pertinent estimate since it allows evaluation within the context of the application at hand rather than in an abstract subspace.

3.2.4 Independent component analysis (ICA)

Once we have selected a threshold for the number of principal components to retain at the group-wide level, we can use the reduced $Q_{red2} \times M$ matrix X to perform the third step of the GICA analysis, namely independent component analysis (ICA). ICA is a blind source separation technique that uses higher-order statistics to discover hidden factors (sources or features) in the observed data such that these factors are maximally statistically independent.

ICA decomposes X as follows:

$$X = \hat{A}\hat{S} \tag{3.6}$$

where $\hat{A}$ is a $Q_{red2} \times Q_{red2}$ mixing matrix and $\hat{S}$ is a $Q_{red2} \times M$ matrix which contains the sources or independent components (ICs). For IMS data, these ICs are mass spectral patterns that group co-localizing ions and underlie the measured mass spectra. The statistical independence among components imposed by ICA is much stronger than the orthogonality imposed by PCA, which only decorrelates the data and defines orthogonal directions. ICA can be seen as a higher order generalization of PCA and has been used in a wide variety of blind source separation applications [89] [90], including in IMS [82]. Many variants of the ICA algorithm exist. Here we use the infomax algorithm given its good performance on fMRI data [85] [91]. Infomax maxizes the independence of the retrieved components by minimizing the mutual information between them. In contrast to PCA, ICA commonly requires an iterative method to converge to a set of ICs and its solution is not unique. We therefore use multiple ICA runs (10) and check for consistency of the components using ICASSO [92]. ICASSO clusters ICs from multiple runs into groups based on correlation distance and the resulting cluster centroids are used as the final ICs.

3.2.5 Back-reconstruction

Once we have retrieved our biomolecular signals, captured by the ICs we want to know the tissue-specific spatial expression of these ICs. This means that the retrieved components need to be propagated back to the original IMS space, through a series of projections that invert the different decomposition steps. There exist several methods to perform this back-propagation, and here we will utilize the GICA3 method, that showed the most faithful results by Erhardt *et al.* [93].

The tissue-specific spatial expression $\tilde{R}_i$ of a retrieved biomolecular signal can be obtained through the following inverse transformations:

$$\tilde{R}_i = U_{i,(Q_{red})} G_{i,(Q_{red2})} (G^T_{i,(Q_{red2})} G_{i,(Q_{red2})})^{-1} \hat{A} \tag{3.7}$$

3.2.6 Discerning common from differential components

Unlike PCA, ICA does not provide a ranking of its components. Given the richness of IMS data, and the often large number of components present in the data, it is recommended to rank components to maintain oversight of the results and gain insight on the impact of each component. As we are interested in finding differential components between tissue classes, we use analysis of variance (ANOVA) on the spatial expression of the components as a heuristic method to rank components. ANOVA is a commonly used technique in statistics to determine whether the means of 2 or more groups or classes are significantly different [94]. Components that are highly expressed in one class or group and not in the other will be ranked higher. Although the application of ANOVA in this setting does not follow the rigorous statistical assumptions that traditionally need to be met, the method is used here as a simple ranking technique and no statistical inference is sought. This approach empirically delivered a good practical ranking of the components. An alternative approach could be to use the total sum of the spatial expression per class to rank components, giving a higher ranking to components with a large difference in spatial expression between classes. The ranking aspect does not influence the analytical results, but rather pertains to how the results are reported in a form that a human user can digest more easily. We provide some rudimentary solutions to this task, but it is clear that this aspect needs further research.

3.2.7 IMS Data

Artificial dataset

In order to evaluate the performance of the GICA algorithm, we apply it to an artificially generated dataset. Such a dataset has the advantage that we have a ground truth with known peaks to compare the results to, which is not the case in real tissue measurement. We generate a total of fifteen artificial tissue sections, simulating 3 different disease classes, namely 'Control', 'Disease 1', and 'Disease 2', providing five tissue samples for each class. Each artificial tissue sample has a size of 50x50 pixels and has a m/z range from 3500 to 20000 divided into 15691 m/z-bins. As shown in Figure 3.2, each artificial dataset shares a common base of 4 overlapping areas, each associated with its own signature generated mass spectrum containing 31 peaks, ranging from very small to very large peaks. Peaks are modelled as Gaussian probability distribution curves and are distributed at random positions across the m/z-range. In regions where these areas overlap, pixels contain a mixture of the signature mass spectra of the overlapping areas (each signature spectrum is given equal weight in this mixture). To this base mixture, common to all datasets, we add a class-specific signature (Figure 3.3), which is unique per class and is distributed randomly in groups of pixels across the tissue. Each class-specific signature consists of 23 very small to large peaks. Again, pixels located in overlapping regions contain a mixture of all overlapping mass spectra. In a final step and in order to make the retrieval task more difficult, 4 additional small, non-class-specific signals were randomly added over the different artificial tissues, in a way similar to the class-specific signal, again randomly distributed across the tissue. This brings the total number of components present in the dataset to 11. Once all these different spectra are mixed together, we add Gaussian noise to the mass spectrum of each individual pixel. Variation in the intensity per pixel was introduced by multiplying the intensities over all m/z-bins per pixel with a random number, generated from a normal distribution with mean 1 and standard deviation 0.1 .

Once we have verified correct functioning of the algorithm on the artificially generated group of data samples, we apply it to two genuine IMS datasets, each comprising multiple tissue samples.

Mouse spine data

The first IMS case study consists of four transversal sections of mouse spinal cord (lumbar and cervical region, 1 sample each), comparing an amyotrophic lateral sclerosis (ALS) mouse model to a wild type (WT) mouse both at the lumbar

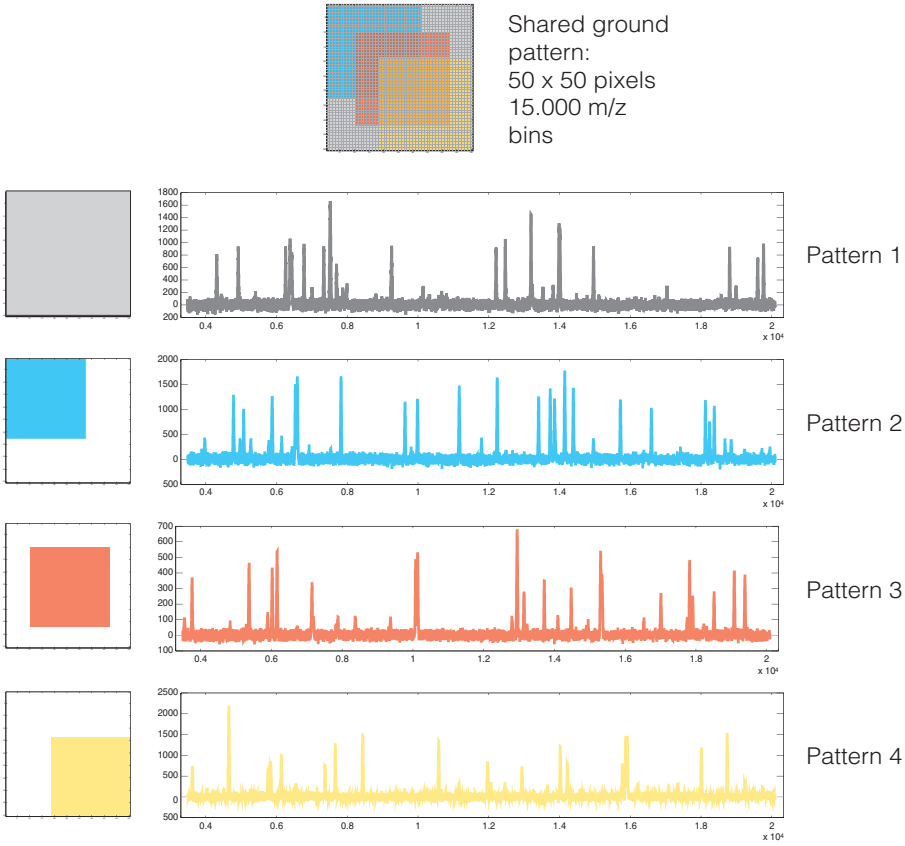

Figure 3.2: Each artificial dataset shares a common base of 4 overlapping areas, where each area has its own artificially generated 'mass spectrum' associated to it. In regions where these areas overlap, pixels contain a mixture of the signature mass spectra of the overlapping areas.

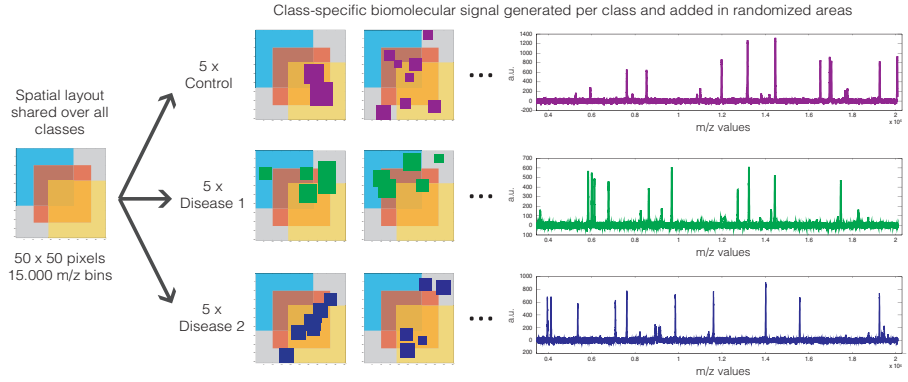

Figure 3.3: Construction artificial datasets. A total of 15 artificial tissues are generated, simulating 3 different disease classes, namely 'Control', 'Disease 1' and 'Disease 2', generating five tissues of each class. Each artificial dataset shares a common base of 4 overlapping areas to which we add a class-specific signature, which is unique per class.

and cervical region (one tissue sample for each condition). The tissue samples were frozen in liquid nitrogen and tissue sections of 12 μm thickness were cut using a cryotome. The sections were then rinsed using increasing concentrations of ethanol (30 - 70%) in order to chemically fixate the tissue. The spinal cord sections were mounted on ITO glass slides and coated with alpha-CHCA matrix (5 mg/ml, 0.1% trifluoroacetic acid) using Labcyte's Portrait P630 acoustic spotter. The IMS measurements were done on an ABI 4800 MALDI TOF/TOF mass spectrometer. The acquired mass extends from m/z 3500 to 20000 divided into 15691 m/z bins. A total of 5842 pixels were measured over the four tissues with an interspot distance of 50 μm in both the x and y directions.

Mouse retina data

The second IMS case study examines retinal degeneration disease using three different mouse models (C57 WT, Rpe65 knockout and Abcr knockout), each class represented by two tissue samples. These were measured on a Bruker UltrafleXtreme II MALDI-TOF/TOF at 10μm resolution, acquiring a total of over 22000 pixels or spectra over all tissue samples. The spectra describe the m/z range from 420 to 1800, using 68608 m/z bins. This amounted to a total data size of 12.1 GB of data when read into MATLAB 2013b (The Mathworks Inc., Natick, MA). All data in this and the previous case study are normalized

both within and across tissues using total ion current (TIC). Both real-life datasets were realigned using the msalign algorithm in MATLAB.

3.3 Results and Discussion

3.3.1 Results artificial dataset

We first describe the results of the GICA algorithm applied to the artificial dataset.

Tissue-specific PCA

We begin our analysis with the tissue-specific PCA, where we perform PCA on each of the artificial tissues separatly in order to extract the principal components (PC) for each individual tissue. The retrieved PCs in spectral space group together the ions that correlate and co-localize within the tissue. Performing this step at the individual tissue level allows retrieval of even small patterns in these data. Additionally, grouping these correlating ions together allows us to represent the original IMS data in a much more concise form for further use. This requires the selection of the number of principal components to retain. In the case of the artificial dataset, we know that the entire tissue data is constructed using only 11 base mass spectra, an thus only 11 principal components would be sufficient to capture all the patterns in the data. In the case of the artificial dataset we know the number of components in the dataset, however this is not the case in real-life IMS data. At this point in the algorithm we want to be very conservative in order not to lose any information. Allen *et al.* [77] have previously shown that retention of additional principal components improves data reconstruction after the analysis, however, the more components we retain, the more computationally intensive the group-wide PCA step will be and thus a compromise must be made. For this analysis we therefore use a standard setting that retains the first 200 principal components which should capture all components for most IMS tissue-datasets, while still being computationally feasible, however this number can be decreased or increased at will. This gives us a reduced data matrix of size 200×15691 for each artificial tissue.

Group-wide PCA

The next step in the analysis is the group-wide PCA, which retrieves the principal components over all datasets. Before performing the group-wide PCA we need to combine all the reduced datasets into a single matrix, resulting in a matrix of size 3000×15961 ($15 \times 200 = 3000$), on which PCA is then performed. Once the PCA analysis is complete, it is once again necessary to select a number of principal components to retain. As we have previously pointed out, correct selection of the number of components at this stage in the analysis is very important, as this is the number of independent components the ICA algorithm will extract from the data in the next step, regardless of whether this is the true number of components present in the data. An over-estimation of the number of components can result in a deterioration of the retrieved components by splitting informative sources over multiple components, whereas an under-estimation will result in discarding of biomolecular patterns.

We therefore applied a range of component selection methods to retrieve the number of components contained in the artificial dataset. All of the component selection methods included in this work (latent variables, MDL, the Laplace method, and data reconstruction) allowed for correct retrieval of number of components (11). Several additional component selection methods, besides the ones mentioned in the methods section were tried out, but were eliminated for further use due to either incorrect retrieval of the number of components in the artificial dataset or extremely long computation times (methods requiring hundreds of randomizations of the data), or both. Interestingly, it is worth noting that selection of the number of components with the 90% threshold, as has often been used for PCA in the context of IMS in the past, would have only retained 10 out of the, in total, 11 components. Granted, there are several differences, as (1) this is an artificial dataset, (2) these components were present in only part of the data, and (3) this selection is done at the group-wide PCA and not on the original data. Nonetheless these components are reported by the other selected methods and could be readily retrieved with GICA.

GICA results

Once the number of principal components is selected at the group-wide PCA level, the group-wide reduced data matrix X can be analyzed using ICA to retrieve the independent components (ICs) that underlie the observed data. The ICA analysis is repeated 10 times and ICASSO is used to integrate the results over the different runs. This results in the mixing matrix $\hat{A}$ and the matrix $\hat{S}$ containing the ICs. The ICs group together the co-localizing ions in a pseudo-mass spectrum, which represents the retrieved biomolecular pattern.

Using Equation 3.7 the spatial expression of these biomolecular patterns in the individual tissues can be reconstructed. Once the back-reconstruction is finished, the GICA analysis is complete.

Figure 3.4 shows the results of the GICA algorithm applied to the artificial dataset. The retrieved ICs are shown on the right hand side, while the spatial expression of the respective IC is shown on the left. Results for 6 out of the in total 15 artificial tissues are shown (2 for each class), with their respective ground truths displayed above them for easy comparison. As described in the methods section, we sort the components according to the largest differences between groups using ANOVA, as ICA does not provide a ranking of the components by itself. The results can be roughly split up into 3 sections: differential signal, non-differential signal, and common signal.

Differential signal Figure 3.4 shows IC1, IC2, and IC3, as the highest-ranking differential signals. These are indeed the class-specific signals introduced in the data, which are expressed in the correct locations as can be verified using the ground truth. Figure 3.5 shows a comparison of the components retrieved by GICA with the original spectra. As we can see, the algorithm succeeds in achieving a near-perfect and noiseless retrieval of the original signal, even for the very small peaks which originally had a very low signal to noise ratio.

Non-differential and common signal While we are most interested in the differential signal, the GICA algorithm is also able to retrieve the small signal that was randomly added to the data (non-differential) and the signals that build up the common basis across the different datasets (common). Again, the retrieved signals very closely resemble the original signals, confirming the correct operation of the algorithm.

When studying the spatial expression of the retrieved components in detail, for example in the case of the common signal, one can see variation in the expression across the found areas. This can be explained as followed: as described in the methods section, the retrieved spatial expressions are projections of the original data onto the retrieved independent components. Thus, the spectrum in the original pixel is projected onto the retrieved biomolecular pattern, and the more similar the original spectrum is to the retrieved IC, the stronger the spatial expression will be for that pixel. In the case of the common signal in IC8, for example, spatial expression is highest in the lower right corner where there is the smallest overlap with other areas. The more overlap there is with additional areas, the more mixing with other spectra occurs, and the less the spectrum in that pixel resembles the retrieved component. Even though this component is present in the same quantity throughout that entire area, it is no longer

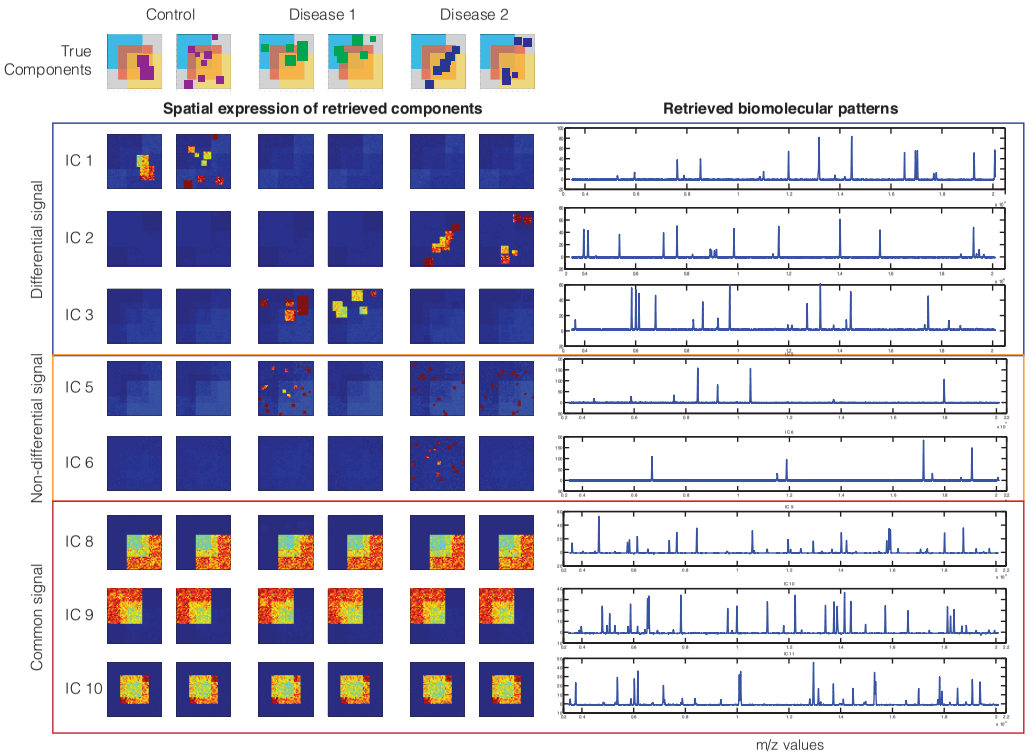

Figure 3.4: Results of the GICA analysis applied to the artificial dataset. On the right we show the retrieved biomolecular signals, which are captured by the ICs as pseudo-mass spectra containing panels of interacting ions. The spatial expression of these patterns in the artificial tissues are shown on the left. We show 2 tissues for each group, 'Control', 'Disease 1' and 'Disease 2', with the ground truth shown at the top for each tissue. Spatial expressions can be grouped into 3 groups, differential, non-differential and common signal. When comparing the expression of the class-specific signal to the ground truth, we see an exact matching of the spatial pattern and a clear absence of spatial expression in the tissue classes where this signal is not present.

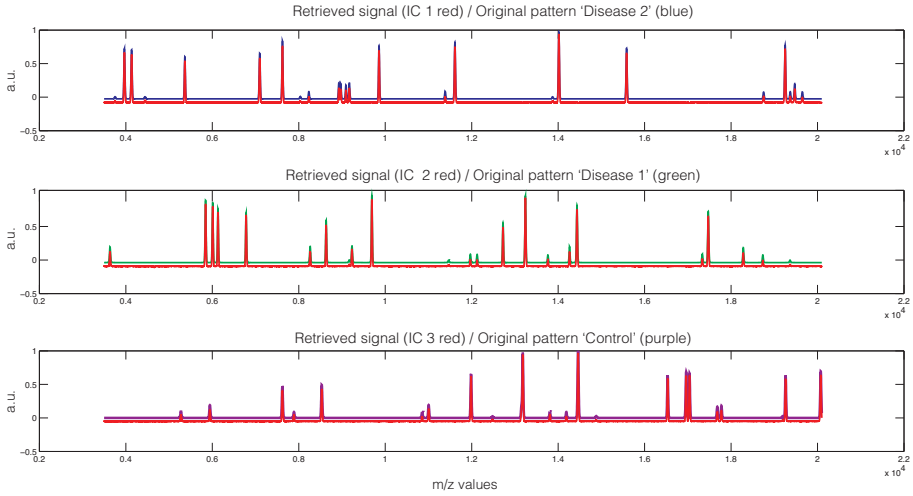

Figure 3.5: Comparison of the components retrieved by GICA with the original spectra used to construct the artificial dataset. Retrieval of the components is near-perfect, including extraction of very small peaks (retrieved components were off-set slightly as otherwise these would not be visible due to overlap).

the only constituting component decreasing similarity. This reasoning applies equally for other decomposition-based methods where the spatial expression is based on projection, including PCA and nonnegative matrix factorization.

Comparison PCA and GICA

Besides using the GICA algorithm, we also analyze the complete dataset (all 15 artificial tissues) using regular PCA, performing the PCA analysis on the complete dataset simultaneously. Figure 3.6 shows the comparison of the differential components retrieved by performing regular PCA on the entire dataset and those retrieved with GICA. The retrieved PCs and ICs are again shown on the right, while their spatial expression in the tissues is shown on the left. Performing the PCA analysis on the entire 4.6 GB dataset required 11 GB of RAM. GICA, on the other hand, only required a peak RAM of 1 GB of the memory necessary to calculate the PCA of one separate artificial tissue. The calculation resources necessary for GICA increase linearly with the number of tissues and because of this the GICA analysis can be run on a standard 8GB RAM desktop computer, whereas this is not the case for regular PCA.

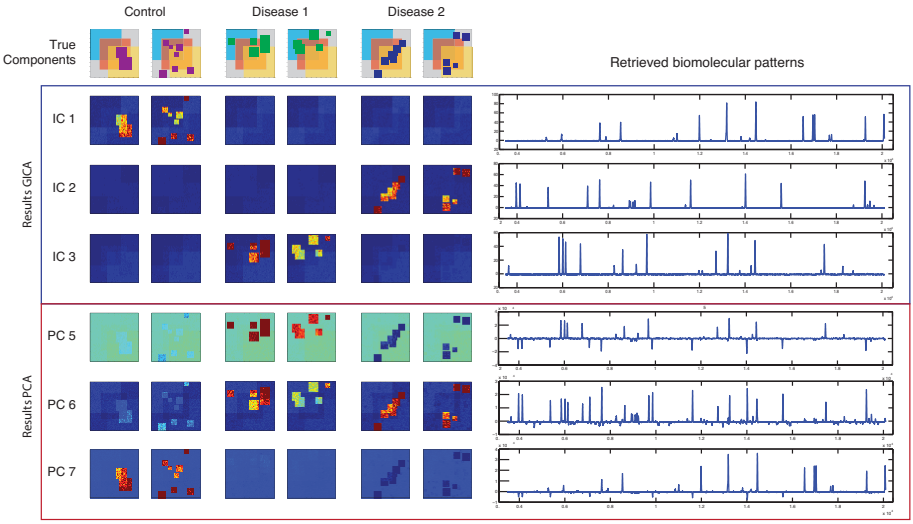

Figure 3.6: Comparison of the results obtained with GICA to those obtained with regular PCA on the full dataset. The spatial expression of PC6 shows a positively expressed component that is present in both Disease 1 and Disease 2, whereas no common peaks exist in those areas that are common to only Disease 1 and Disease 2.

Using regular PCA for the purpose of unmixing has several disadvantages, as has already been pointed out in previous studies [55]. Nonetheless, PCA is still widely used for the analysis of IMS data. This simulated dataset offers an excellent opportunity to illustrate some of the caveats in doing so. The effect is best illustrated in component PC 6, which shows the positive spatial expression of a component common to both Disease 1 and Disease 2, even though there are no common peaks between these classes exclusively located in those regions. The positive expression of this 'phantom' component in this region implies that it is not enough to simply split the expression of the component into a positive and a negative part as is routinely done in IMS research.

This relates back to our comment above, namely that the spatial expression of a component in a projection based method is related to how closely the spectrum in the original pixel resembles that in the retrieved component. This, however, offers no guarantee that peaks that are expressed in that component are also expressed in the true dataset, even if the spatial expression of this component is positive in that location. This effect is likely to be present in any projection-based method that allows negative components, however PCA suffers from this effect more than other methods. This is because the goal of PCA is

to represent the data as concise as possible, which means that it will attempt to put as much information (and thus variation) as possible into each consecutive component. It compensates for this by creating negative components such as the ones in PC5. As can be seen GICA, suffers less from this effect due to it being based on ICA, which attempts to retrieve the true sources of the observed signals (unmixing) rather than representing the data as concise as possible. Nontheless this demonstrates the importance of always going back to the original ion images, to confirm whether an ion is truly present.

3.3.2 Results mouse spine dataset

Now that we have verfied correct funtioning of the GICA algorithm, we perform the analysis on a modest-sized real-life dataset of ALS versus healthy mousespine.

Tissue-specific PCA & group-wide PCA

We perform tissue-specific PCA analysis on each of the four mousespine tissue datasets individually. Once the PCA analysis is complete and the tissue specific principal components have been extracted, we again reduce the number of components we retain using the standard setting of 200, which greatly exceeds the number one would select based using any of the selection criteria. The reduced data of each tissue are then again combined in a single matrix (size 800×15691) and group-wide PCA is performed on this concatenated matrix. After the PCA analysis we again need to select the number of components to retain for subsequent dimensionality reduction. Figure 3.7 shows the results of the different component selection techniques applied after the group-wide PCA step. Using the latent variables 12 components are selected, using the minimum description length (MDL) 22 components are selected, and using the Laplace method 107 components are selected. Using the 90% threshold, 10 components would have been selected. All of these estimates are fast criteria, and can be calculated in a matter of seconds. We can immediately see that there is large degree of variation in the number of components between the different methods. Although plotting the latent variables (scree plot method) is the recommended method by Jolliffe [67], it is very hard to determine the exact number of components from this plot, as there is no clear cut-off. Using the log2 or log10 plot as sometimes recommended [67], did not offer additional clarification.

The MDL and the Laplace method both have the advantage that there is an objective criterium on which to pick the number of components (the minimum of the graph for MDL and the maximum of the graph for the Laplace method). The

graph for the Laplace method, however, does not show a very clear maximum; just like in the scree plot method there are a large number of components incrementally adding information, leading to selection of a much larger number of components (107) than the other methods. As Minka [88] already describes in the context of dimensionality reduction on face image data, it may be that this is indeed the true number of components needed to describe the complete data without loss. This is very useful in the context of data compression, and it should be noted that this number is much higher than the 10 components suggested by the 90% threshold. However for interpretation of the data this number is too high. Additionally, Calhoun *et al.* [77] note that in the context of GICA, over-estimation of the number of components is more detrimental than under-estimation, as it can cause true components to get split up between multiple other components, distorting interpretation of the results. The MDL shows a clear minimum and returns a much lower number of components; this is due to the fact that it penalizes additional complexity of the model, which causes components that do not add sufficient extra information to be excluded from the model, making this selection criterium more useful for human interpretation.

Due to the large variation in estimates between the different methods, we use an additional method to study the impact of selecting the number of components in closer detail. This method gauges the influence of including an additional component at the group-wide PCA by reconstructing the IMS data using different numbers of components, and then plotting the change in ion intensity. Figure 3.8.A shows the impact of adding an extra component on the absolute ion counts over all tissues and variables in the mousespine dataset. We show both the maximum (absolute) change in ion count over all variables (right) as well as the 99 percentile of the change (left), which is less susceptible to outliers. From the 99 percentile graph, the 22 components selected through MDL seem reasonable as no major changes occur in the graph after this point (since the graph remains flat). Based on the graph depicting the maximum change, the 107 components selected by the Laplace method seem reasonable as well. We see that a small spike is present in the "maximum change"-graph at 107 added components, after which no major changes occur. This seems to follow the previous reasoning that the Laplace method captures all variation, whereas the MDL captures the largest variations, and discards components adding incremental improvements. Figure 3.8.B gives an overview of the amount of (absolute) change per added component, only this time it is split up per variable, which has the advantage that one can see structured changes by adding additional components at a glance. Again we show the maximum variation (right) and the 99 percentile variation (left). As expected, both share the same structure, however, the maximum per variable is more noisy and has higher intensities. One can see from this figure that structured changes are still occurring after the suggested threshold of component 22 by

MDL. To our jugdement, the final major change in ion intensities occurs around component 34 (marked with a white circle in the Figure) which was selected as an additional threshold, however, this method again lacks an objective criterion.

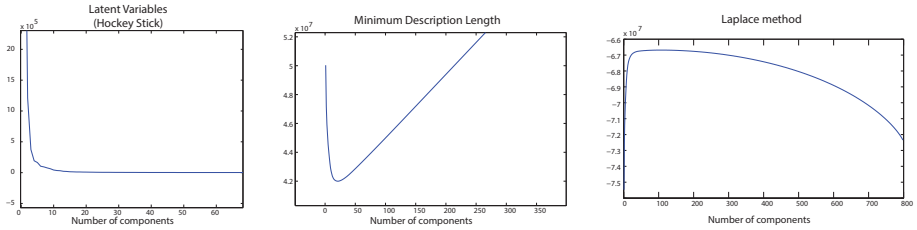

Figure 3.7: Comparison of component selection for the mouse spine data using the latent variables (12 components), minimum description length (22 components) and the Laplace method (107 components).

GICA results

Figure 3.9 shows a summary of the GICA results obtained for the mousespine dataset using the 22 components suggested by MDL. ICASSO showed a consistent retrieval of the 22 components over the 10 repetition of the ICA algorithm. Analogously to the artificial dataset, the left hand side shows the spatial expression of the retrieved components in the tissue. Healthy and diseased (ALS) tissues are grouped together for easy comparison. Again components were sorted according to differential expression between healthy and diseased tissues using ANOVA. The retrieved biomolecular patterns (ICs) are shown to the right. IC1 shows a biomolecular pattern that contains two prominent peaks which are almost exclusively expressed in the healthy tissue and are almost absent in the diseased tissue. As discussed before, we verify the expression of this signal in the actual IMS data, which is shown in Figure 3.10. Plotting the ion images for the 3 highest peaks in the retrieved biomolecular signals, we indeed confirm that the ions reported are almost exclusively expressed in the healthy tissue. A second component of interest, IC4, is expressed much more strongly in the diseased tissue than in the healthy tissue. Figure 3.11 shows the ion images associated with the three highest peaks in the retrieved biomolecular signal. We see that ion m/z 4819.9, the highest peak in IC4, and m/z 4910.1 appear almost exclusively in the diseased tissue. These ions are expressed in the anterior horn, the region in the spine containing the motor neurons and which is known to be affected in ALS. Ion m/z 5720.5 is expressed in a similar location and although it is also expressed in the cervical healthy tissue, it is almost absent in the lumbar healthy tissue. The observed expression of IC4

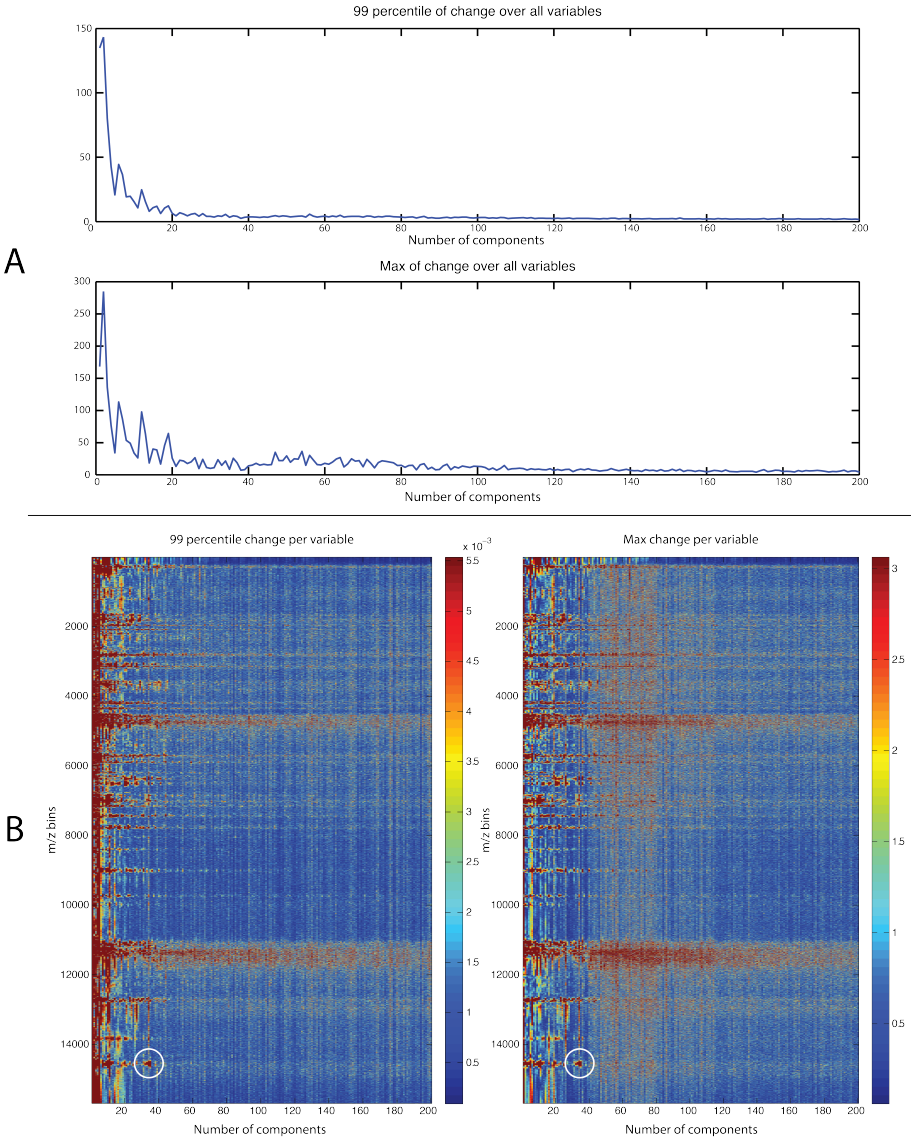

Figure 3.8: Data reconstruction for component selection in the mouse spine samples. Pane A shows the maximum and 99 percentile absolute change in ion counts per added component over all samples. Pane B shows this information split up per m/z bin which gives an immediate overview of the added structural variation with each additional component. The white circle marks the final major change in ion intensities which occurs around component 34, and which was therefore chosen as a threshold.

can be seen to be an aggregated expression of the individual ions, and does not necessarily agree with the expression of each individual ion, as mentioned in the discussion of the artificial dataset.

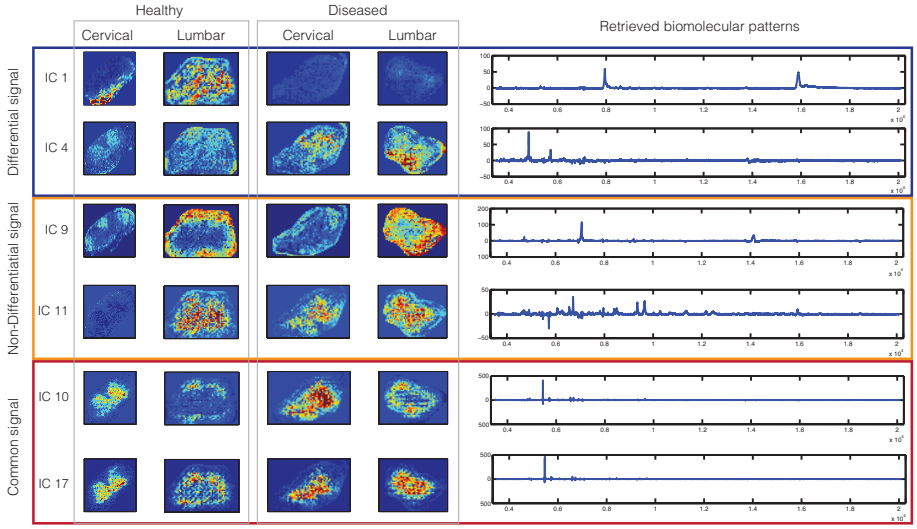

Figure 3.9: Results of the GICA algorithm applied to the mouse spine dataset. On the left the spatial expressions of the found biomolecular patterns in both the diseased and the healthy tissue are shown. Using GICA we can retrieve differential components between the healthy and the diseased tissues.

The GICA analysis was repeated using the numbers of components suggested by the other component selection methods, namely 12 (scree plot), 22 (MDL), 34 (reconstruction) and 107 (Laplace) components. Using the 12 components suggested by the scree plot, IC4 could not be retrieved, and was presumably lost at the group-wide PCA due to retention of too few components. Using 107 components, no results could be obtained as the ICA did not converge, meaning that the ICA could not reliably retrieve the 107 components, probably due to the fact that a lot of these components contained very small signals. The same differential patterns were observed using 34 and 22 components, although in the 34 components selection, IC 1 was split over several components. Retrieval of the 34 components was less consistent over the 10 runs of the ICA algorithm, with some components only retrieved in part of the runs. From this it seems that both MDL and the data reconstruction method provide good estimates for the number of components. MDL has the advantage that it needs very little computation (calculated in seconds) and is objective (no possibly biased interpretation of the graphs by the user is necessary). The data reconstruction

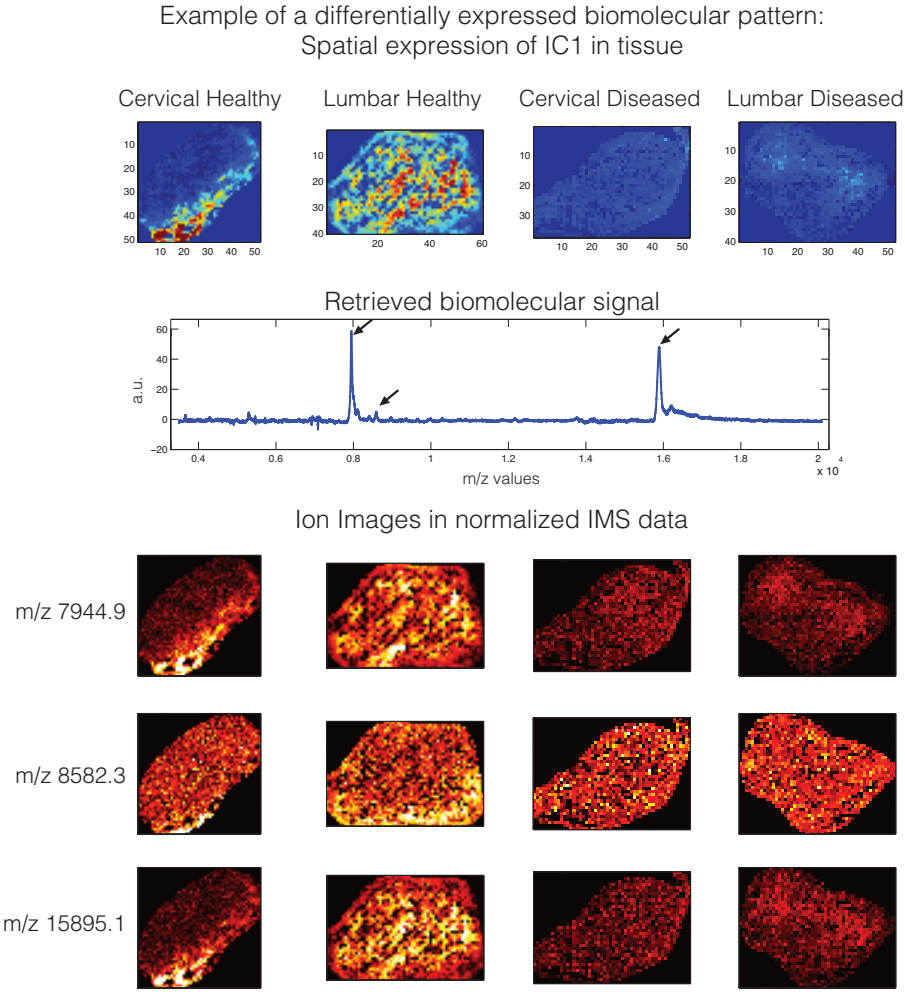

Figure 3.10: In-depth look at component IC1. The spatial expression of IC 1 in the mousespine tissue is shown, along with the accompanying biomolecular signal. Below this, the ion images for the 3 highest peaks in the retrieved signal are shown, which confirms that these are indeed almost exclusively expressed in the healthy mouse spine tissues.

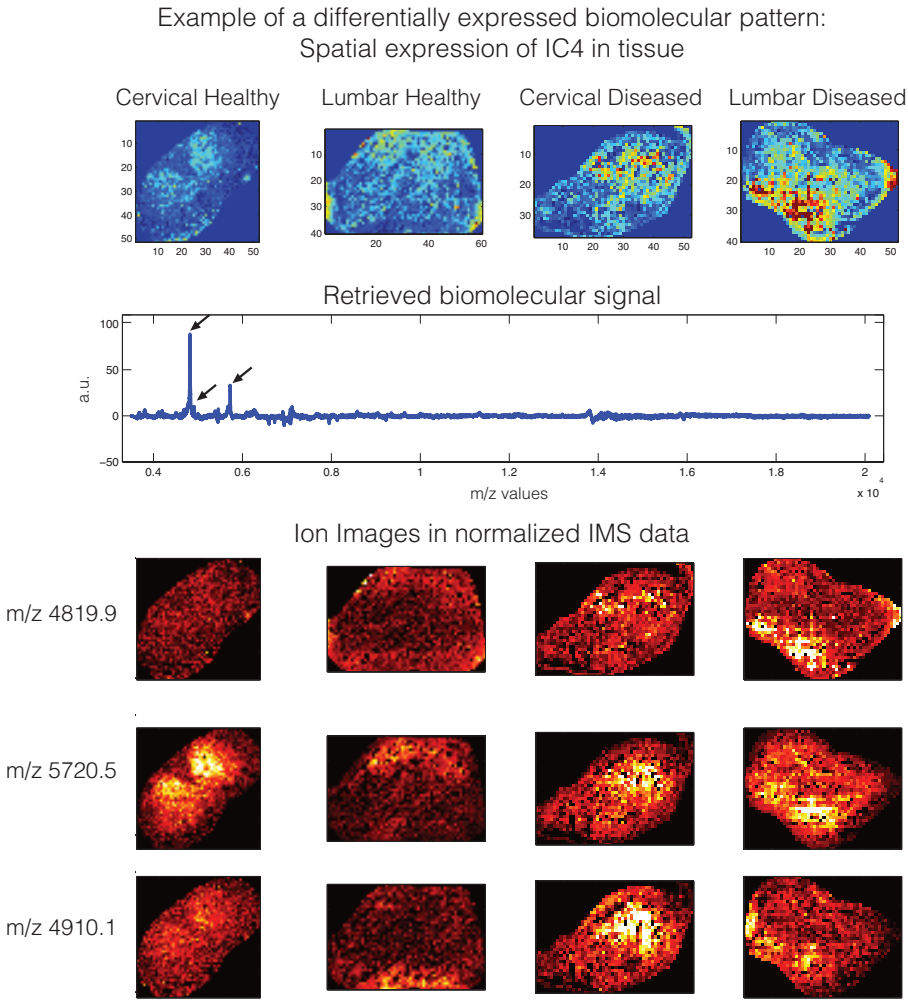

Figure 3.11: In-depth look at component IC4, which is expressed much more strongly in the diseased tissue than in the healthy tissue. Ion images of the 3 highest peaks are shown. These ions are expressed in the anterior horn, a region known to be affected in ALS.

method, on the other hand, gives more insight into the impact of leaving out an additional component, but is a lot more computation intensive, in the order of minutes to hours, depending on how many components one wants to use for reconstruction. An interesting observation is that Allen *et al.* [77] report an over-estimation of the number of components when using MDL for estimation of components in fMRI data. In fact Allen *et al.* [77] have added additional corrections to the MDL selection to correct for this overestimation. Judging by the data obtained from the data reconstruction method, we do not observe this overestimation, but rather observe an underestimation of the number of components. Nonetheless, MDL still gives a very good estimate of the number of components.

Comparison PCA and GICA

When performing regular PCA on the mousespine data, we observe the same mixing of components as in the artificial dataset, making it hard to find truly differentiating components. Figure 3.12 shows the results of the PCA analysis. PC 6 shows a differential component with a spatial expression similar to GICA component 1. The two largest peaks of GICA component 1 are present in PC 6, along with several additional peaks. A component similar to differential GICA component 4 could not be retrieved, however, and seems to be captured in PC 19 and PC3. Without prior knowledge these components would probably remain undetected as being differential. Due to the modest size of this dataset, (around 700MB), the analysis could be run on a standard desktop computer.

3.3.3 Results retina

In this section we present the results of the analysis of the mouse retina samples. These retina samples contain several very thin structures such as the retinal pigment epithelium (RPE) which is only a single cell layer thick [95]. It thus presents an ideal test case to see how the GICA method can retrieve such small structures.

Tissue-specific PCA & group-wide PCA

We apply component selection methods after performing group-wide PCA: 90% threshold (6), scree plot method (23), MDL (99), data reconstruction (139) and Laplace method (263). We observe the same trend as before, with the 90% threshold selecting a very small number of components and the Laplace method selecting a very large one. Given the poor performance of the scree

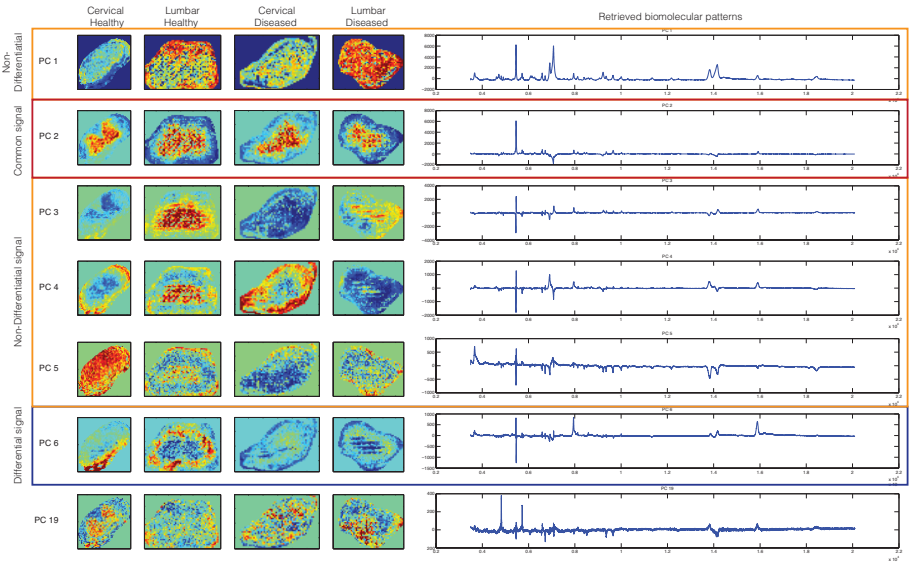

Figure 3.12: Results of the PCA analysis on the mouse spine data. PC 6 shows a differential component similar to GICA component 1. A component similar to GICA component 4 could not be retrieved, and seems to be captured in PC 19, however this component would probably not be detected as differential.

plot in the mouse spine dataset we did not calculate the GICA analysis for this number of components, nor for the 90% threshold which recommends an even lower number of components and which will almost certainly discard small components in the data. Figure 3.13 shows the data reconstruction summary for the retina data. Given the prior knowledge that these datasets contain very small structures, we aim to capture as much as possible of the structured variation and therefore select 139 components as the threshold, based on the 99 percentile change per variable plot in pane B. Using the large number of components recommended by the Laplace method once again led to failure for the ICA algorithm to converge. We perform the GICA analysis using both the 99 components selected by MDL and the 139 components selected through data reconstruction.

GICA results

The GICA analysis for the full retina dataset (12 GB) could be run on a regular desktop computer with 8GB of RAM, with a peak memory requirement of 5.5 GB, which was the memory necessary to calculate the PCA for an individual

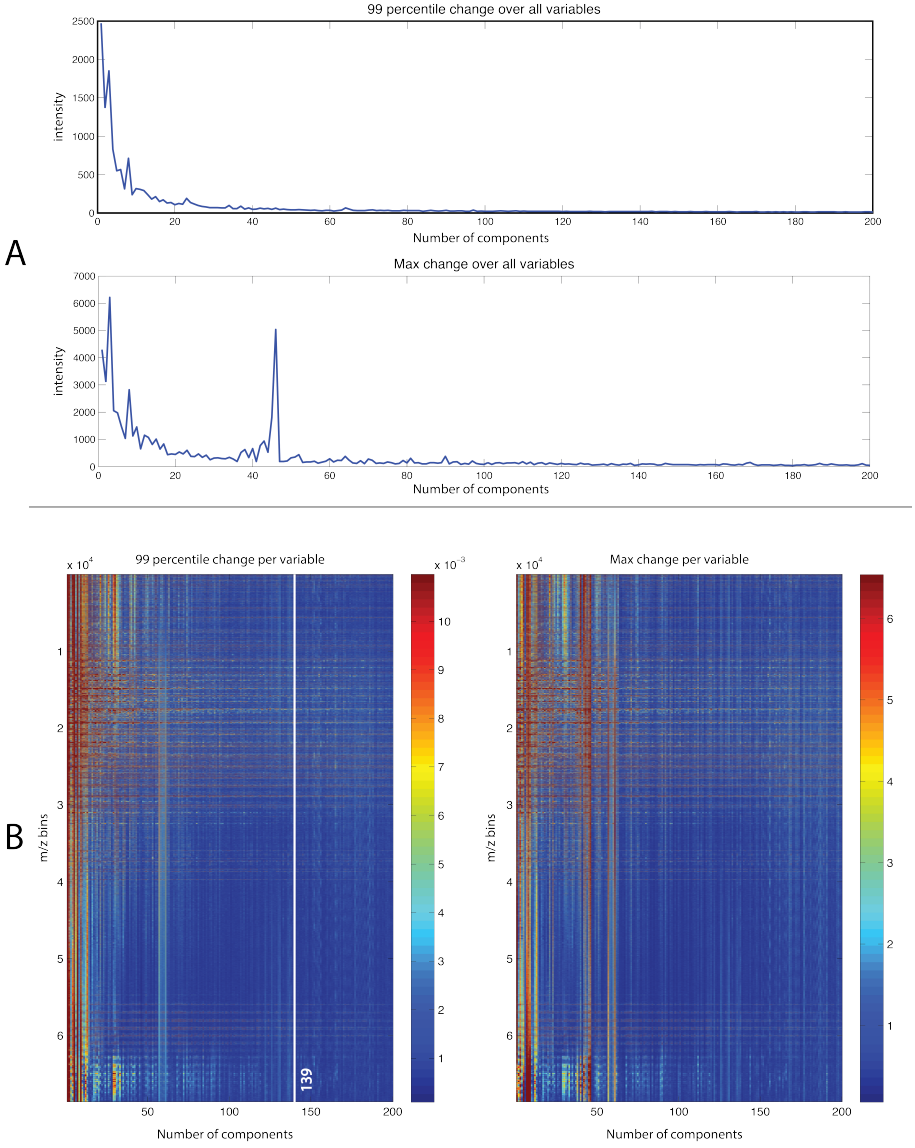

Figure 3.13: Data reconstruction for component selection in the mouse retina samples. Pane A shows the maximum and 99 percentile absolute change in ion counts per added component. Pane B shows this information split up per m/z bin. As we aim to capture as much of the structured variation as possible, we select 139 components as the threshold, based on the 99 percentile change per variable plot in pane B.

tissue. On average, the data size for a single tissue was around 2 GB. Only for a very large amount of samples would these memory requirements change, as the largest dimensionality reduction occurs at the single tissue level. Calculation of the full dataset on the desktop computer took approximately 3 hours, with PCA on the single tissues taking up most of the calculation time, followed by the ICA analysis which is repeated 10 times. As mentioned previously, however, the PCA analyses can be run in parallel if a fast analysis is necessary.

Figure 3.14 shows IC 34, which was retrieved through the GICA analysis, and its corresponding spatial expression. This component shows expression in the RPE, a structure only one cell layer thick, which illustrates the sensitivity of the algorithm, given that even these very small structures are retained. Below the retrieved biomolecular signal, we show the expression of this component in the IMS data. From both the expression of IC 34 and the ion images these ions seem to be absent in the Abca4 knockout, making the panel of biomolecular ions retrieved with GICA interesting candidates for further research. By looking at the expression of IC34, however, one can see that the first tissue of the Rpe65 knockout and the first tissue of the Abca4 knockout seem to miss several retinal layers at the top due to the difficulty of cutting these small samples. Additionally, as shown in Figure 3.15, IC9 shows a molecular profile that is expressed exclusively in a single tissue of the Rpe65 knockout and which could be detected due to the exploratory nature of GICA. Due to the extent to which this profile was present, this tissue could be classified as an outlier using GICA.

For these reasons, we repeated the GICA analysis with a subset of the retina samples, including an intact sample of each class. A summary of the most important components is shown in Figure 3.16. Again we find a number of differential and common components over the sample classes. One of the most notable differences with the mouse spine data is the large number of correlating peaks that are contained in the retrieved biomolecular patterns, demonstrating the richness of these lipid data. A first differential components, IC2 expresses a biomolecular pattern that is almost exclusively expressed in the Rpe65 knockout. An in-depth look at this pattern and the accompanying ion images can be found in Figure 3.17, confirming that this component is indeed expressed in the true IMS data. Looking at IC3 and IC9, on the other hand, we see that this biomolecular pattern is not expressed in the Abca4 knockout. Figure 3.18 zooms in on IC3 and shows ion images of the 4 largest peaks in the retrieved pattern. We again observe a biomolecular pattern that is differentially expressed in the RPE, which is known to be affected in retinal degeneration diseases.

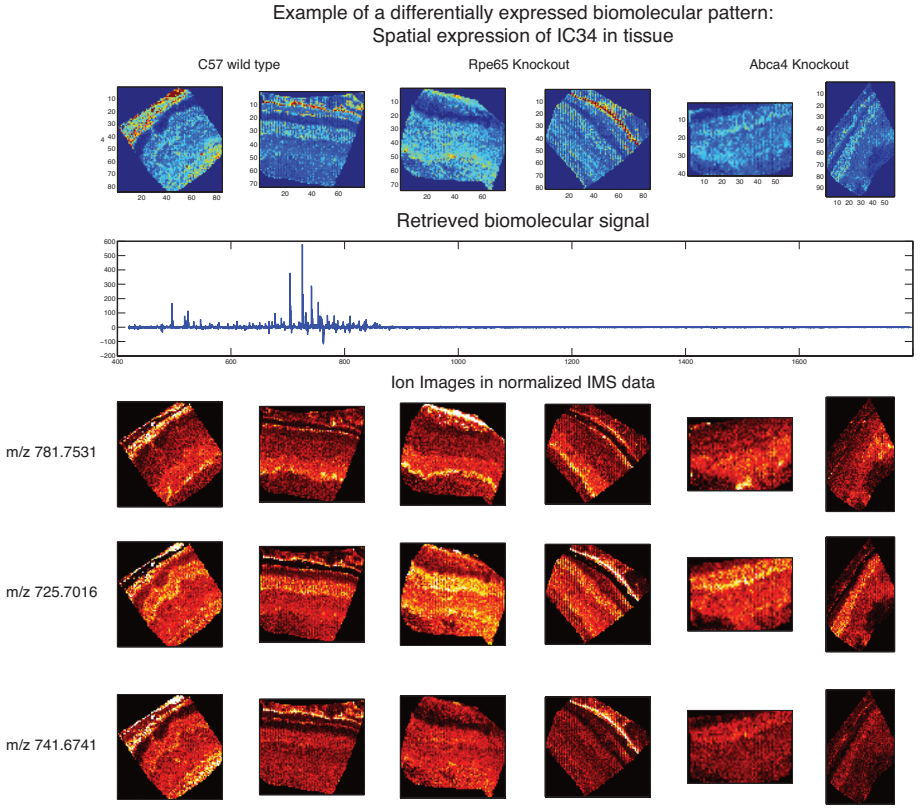

Figure 3.14: Spatial expression of component 34 in the tissue. These ions seem to be absent in the retinal pigment epithelium (RPE) of the Abca4 knockout. Despite the numerous dimensionality reduction steps, GICA is able to retrieve this small structure and extract its differential signature.

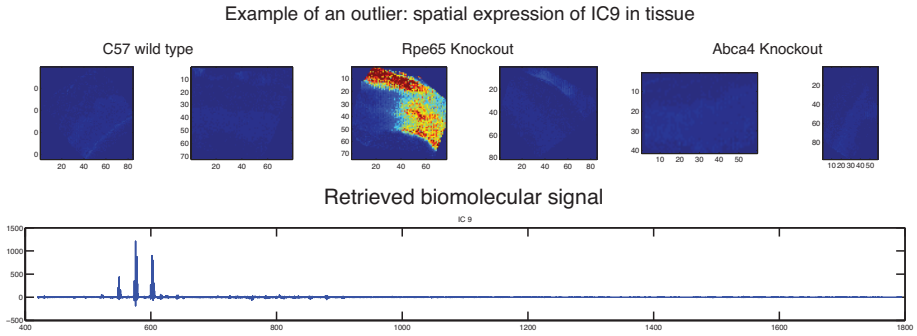

Figure 3.15: IC9 shows a molecular profile that is expressed exclusively in one of the Rpe65 knockout tissues. This component is expressed throughout the entire tissue and can therefore this tissue was considered an outlier. Using GICA we could easily detect this due to the explorative character of this analysis.

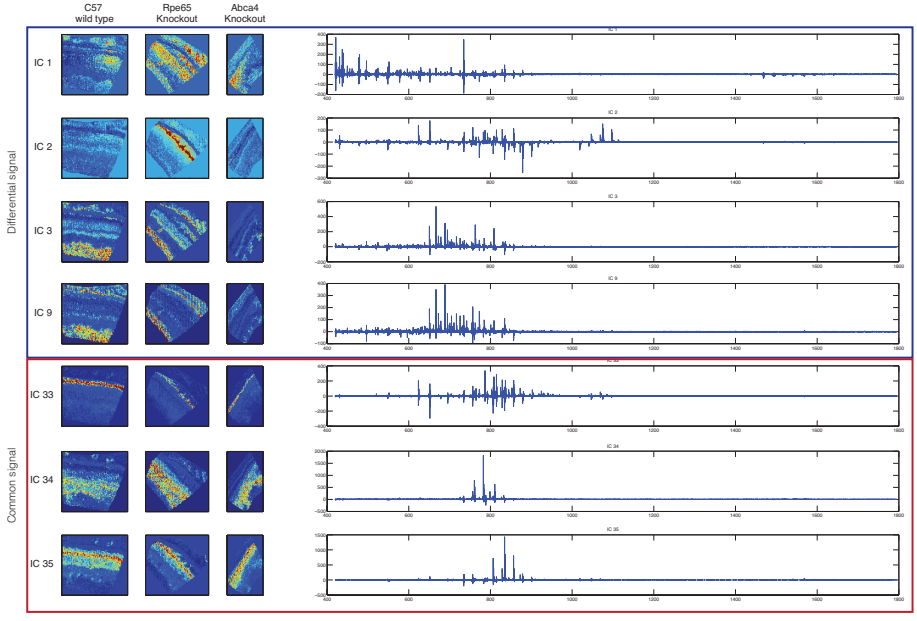

Figure 3.16: Overview of the results obtained from performing GICA on a subset of the retina samples. Results are sorted according to largest difference in spatial expression (left). The independent components are shown on the right.

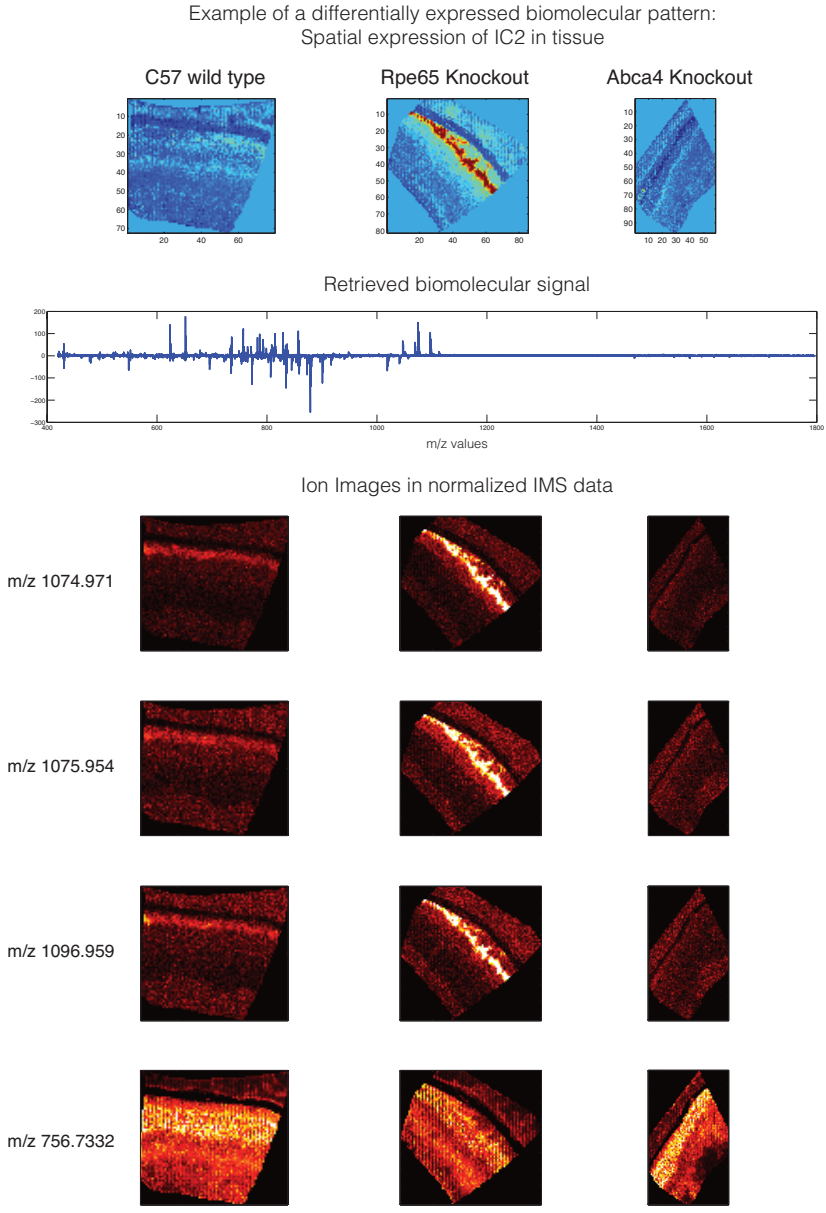

Figure 3.17: Zoom-in of the expression of component 2 retrieved from the retina samples (sub-set). This component is almost exclusively expressed in the Rpe65 knockout. We confirm the expression with ion images of the highest peaks in the retrieved biomolecular signature.

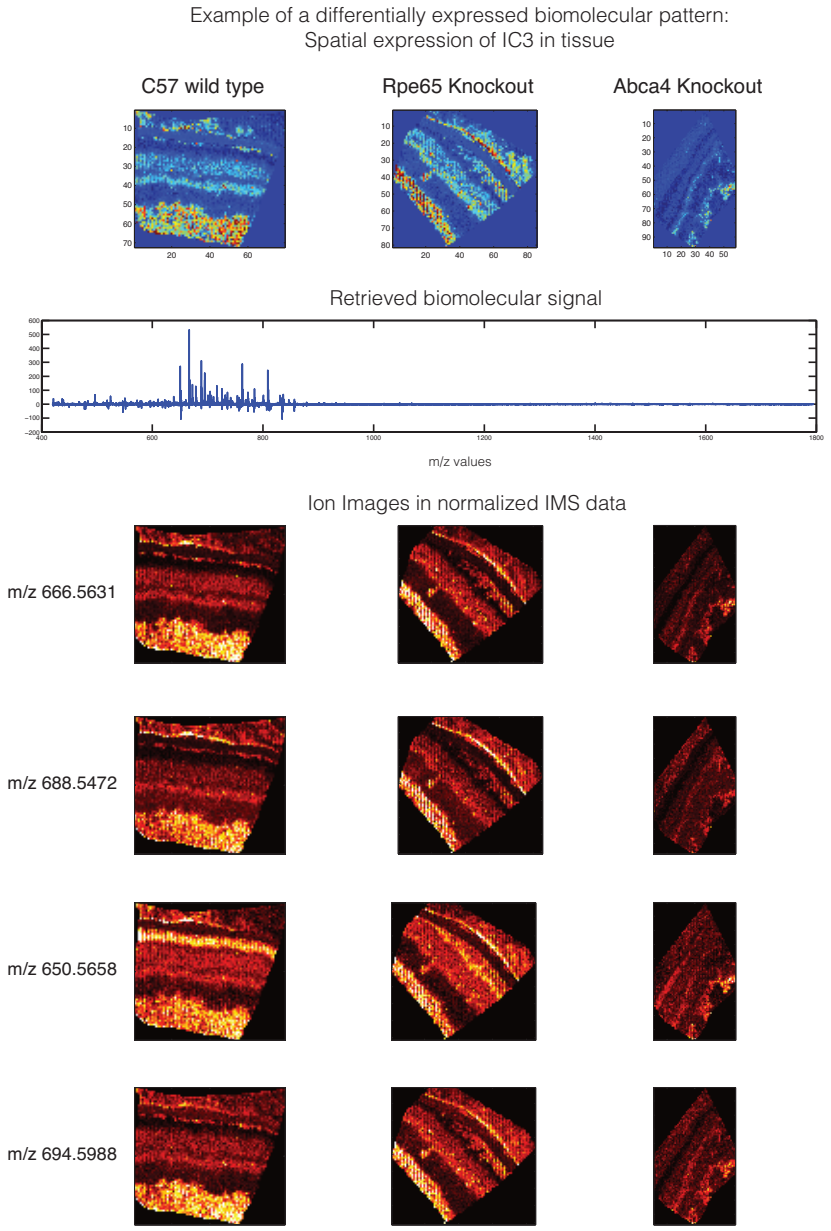

Figure 3.18: Zoom-in of the expression of IC3 retrieved from the retina samples (sub-set). This component is almost completely absent in the Abca4 knockout. We confirm the expression with ion images of the highest peaks in the retrieved biomolecular signature and clearly see the presence of this ion in the RPE of the C57 wild type tissue and the Rpe65 knockout, while being completely gone in the Abca4 knockout.

Comparison PCA and GICA

The PCA analysis of the full dataset could only be performed on a dedicated computational server and resulted in several out of memory errors in MATLAB, which occur when the data to be analyzed is too large to fit in computer memory. The analysis required over 35GB of RAM and several hours of computation time on this server. PCA results (not shown) followed the same trends as those observed previously, namely a strong mixing of components which made discovery of differential patterns between tissues difficult.

3.4 Conclusions & Future work

In this chapter we have introduced a new method for differential analysis of IMS datasets, an adaptation of the Group Independent Component Analysis (GICA) method. The method enables concurrent computational analysis of a large number of IMS-measured tissue samples and with reduced risk for information loss. By performing a pattern extraction step already at the level of a single tissue sample, the method promotes tissue-specific features which could be lost in an immediate simultaneous analysis over all tissues. We demonstrated correct operation of the algorithm on an artificial dataset, and showed near perfect retrieval of the ground truth differential components between the tissues. We applied the method to two genuine IMS case studies, one focused on mouse spine and another on retinal degeneration in mouse retina. These patterns suggested by the algorithm were verified by examining the corresponding ion images directly in the IMS data and were found to be in line with what a human would consider common and differential patterns. The true impact of the method lies in automating the discovery of such patterns, rather than relying on manual exploration of the often massive IMS datasets.

Towards the future, one area for expanding this work is in constructing a non-negative variant of the current GICA adaptation, such that the dimensionality reduction steps follow the inherent non-negative nature of ion counts even more closely. On the computational resources side, it is valuable to consider the advantage that the heaviest computational load of a GICA-based analysis lies in the sample-specific PCA analysis at the single tissue sample level. This enables our method to handle large groups of samples in a parallel way rather than having to perform an immediate massive cross-dataset analysis, which often lies beyond the limits of the computational resources available. Additionally, this method structure has also advantages for single tissue sample analysis. Since single IMS datasets are ever-increasing in data size (e.g. datasets over 100GB per tissue sample are becoming more common), single tissue datasets

are becoming so large that performing PCA on a single sample is sometimes no longer feasible. Our method, however, could be used to enable such an exploration by dividing the tissue area into separate areas and treating each area as a separate sample, essentially adding an additional hierarchical layer to the analysis.

Chapter 4

Automated Anatomical Interpretation of Imaging Mass Spectrometry Data using Curated Atlases

This chapter is based on the following publication:
N. Verbeeck, J. Yang, B. De Moor, R.M. Caprioli, E. Waelkens, R. Van de Plas, "Towards Automated Anatomical Interpretation of Ion Distributions in Tissue: Linking Imaging Mass Spectrometry to Curated Atlases", *Analytical Chemistry*, vol. 86, no. 18, pp. 8974-8982, 2014.

In this chapter we explore the potential of integrating IMS data with information from additional data sources and by doing so, creating richer datasets and enabling deeper insights into the IMS data. A first way of performing such integration is by combining IMS data with data from other imaging techniques or modalities, which is commonly referred to as multimodal imaging. Here, the same or a neighboring tissue slice is measured using IMS and multiple other imaging technique and the resulting data are analyzed together. By combining different imaging modalities, the strengths of one modality can help alleviate the weakness of the other, enabling even more information to be extracted. For example, the high spatial resolution but low biochemical information of microscopy can be combined with the, comparatively, low spatial resolution but high biochemical information of IMS. In this way, Van de Plas *et al.* [96] have used microscopy measurements to predict ion distributions at a spatial

resolution that exceeds that of measured ion images by ten times or more. Several other studies have demonstrated the power of multimodal imaging. Chughtai *et al.* have combined SIMS and MALDI IMS data in the imaging of musculoskeletal tissues [47]. In this way the high spatial resolution but limited mass range of SIMS could be complimented with the extended mass range of IMS enabling detection of higher mass phospho-lipids and proteins. Masyuko *et al.* [97] used Raman spectroscopy in combination with IMS. In contrast with IMS, Raman spectroscopy is both nondestructive and amenable to ambient conditions and can thus be used to perform live cell imaging. Andersson *et al.* [98] have studied the combination of magnetic resonance imaging (MRI) with IMS with a focus on 3D reconstruction of MSI data. Wehrl *et al.* [99] have studied choline metabolism in mouse glioma utilizing Magnetic Resonance Spectroscopy (MRS) and Positron Emission Tomography (PET) with SIMS. Here the limited chemical information of *in-vivo* MRS and PET is complemented with the much richer *ex-vivo* chemical information of SIMS.

Multimodal imaging with IMS is an exciting new evolution in IMS and offers many interesting new research perspectives, however, the work presented in this chapter focuses on a second way of integrating IMS data with additional data sources, namely by combining IMS with curated data sources (in contrast to the measured data discussed above). Although computational methods have made it practically feasible to search large IMS datasets for relevant findings, these methods lack access to an important source of information that many interpretations rely upon: anatomical insight. In the work presented below, we address this need by (1) integrating a curated anatomical data source with an empirically acquired IMS data source, establishing an algorithm-accessible link between them; and (2) demonstrating the potential of such an IMS-anatomical atlas link by applying it towards automated anatomical interpretation of ion distributions in tissue.

The concept is demonstrated with mouse brain tissue, using the Allen Mouse Brain Atlas as the curated anatomical data source that is linked to MALDI-based IMS experiments. We first develop a method to spatially map the anatomical atlas to the IMS datasets using non-rigid registration techniques. Once a mapping is established, a second computational method, called correlation-based querying, gives an elementary demonstration of the link by delivering basic insight into relationships between ion images and anatomical structures. Finally, a third algorithm moves further beyond both registration and correlation by providing automated anatomical interpretation of ion images. This task is approached as an optimization problem that deconstructs ion distributions as combinations of known anatomical structures. We demonstrate that establishing a link between an IMS experiment and an anatomical atlas enables automated

anatomical annotation, which can serve as an important accelerator both for human and machine-guided exploration of IMS experiments.

4.1 Introduction

Understanding the spatial context in which molecular interactions take place is becoming increasingly important in the study of biological and pathological processes in living organisms. The spatial distribution of biomolecules and the localization of biochemical interactions throughout tissue often hold crucial clues towards determining the biological functions of these biomolecules. Imaging mass spectrometry (IMS) [13] [14] is a molecular imaging technology that can deliver such spatial information with high chemical specificity for various classes of biomolecules, including metabolites, lipids, peptides, and proteins. IMS has gained considerable momentum in recent years, primarily in the field of tissue biomarkers [15, 16] and drug delivery [17, 18], and has been successfully applied to tissues of various origin, including insect [19], mammalian [20], and human tissue [21–23]. IMS makes it possible to monitor many hundreds of biomolecules simultaneously, making it a prime technology for exploratory studies. However, this exploratory advantage is hampered by the large amount of data that a single IMS experiment can deliver, making interpretation and analysis difficult. Previous work has employed both supervised and unsupervised computational methods [49] [100] [101] [50], such as hierarchical clustering, principal component analysis [65] [66] and probabilistic latent semantic analysis [82], to perform comprehensive analysis of these large datasets. While these methods aid human interpretation by reducing the data size and complexity, they often operate in a blind fashion in the sense that they lack access to an important source of information that many human interpretations rely upon: anatomical information on the tissue under study. This information is available in textbooks and through publically accessible anatomical atlases for various organ types and organisms. However, the use of such information in IMS studies remains largely restricted to manual comparison [102] [103] [71], which poses a practical challenge for larger multi-experiment studies and brings with it a risk of introducing human bias into the analysis. In order to fully utilize this body of anatomical insight for the interpretation of ion distributions in IMS data, a computer-traversable bridge between IMS and curated anatomical information is essential. Abdelmoula *et al.* [104] have taken a first step towards this integration by developing a workflow that performs automated spatial registration of IMS data to anatomical information through microscopy. Moreover, in recent work, Abdelmoula *et al.* [105] have demonstrated that it is possible to generically register IMS data to the microscopy data of a neighboring tissue slice in a fully automated way. This is done through use of nonlinear stochastic

embedding to visualize the high-dimensional IMS data in 2D, and then use this visualization to perform the registration with the microscopic image. We extend this line of research further by focusing on the applications that become possible once a registration is available. In order to do so, we first (1) establish and demonstrate our own algorithm-accessible link between curated anatomical data and empirically acquired IMS data, and then (2) move beyond registration by applying the established IMS-anatomical atlas link towards automated anatomical interpretation of the ion images obtained through IMS. Since a substantial amount of IMS research focuses on rodent brain [106], our case studies use MALDI-TOF IMS data from mouse brain tissue as the empirical data source and the Allen Mouse Brain Atlas [107] as the curated anatomical data source. Both data types have been used in studies of neurodegenerative diseases such as Alzheimer's, Parkinson's, and healthy mouse brain [108] [109]. However, the methods developed in this paper are not specific to these case studies, a particular species, disease model, or atlas. They can be readily applied to any IMS-atlas combination that makes sense within the context of a particular study. The Methods section introduces the two data sources and describes the three computational methods we developed: (i) registration, (ii) correlation-based querying, and (iii) automated anatomical interpretation. The Results & Discussion section applies the developed methods to both a protein-focused and a lipid-focused case study, with complementary details in the Supplementary Information.

4.2 Methods

The first objective, the integration of the two data sources, entails development of two computational methods. The first method spatially registers the IMS data to the anatomical atlas. Registration is a necessary step that precedes anatomical interpretation and makes direct mapping of findings across data sources possible by establishing a common coordinate system [110] [111] [112]. Since the methods that follow are independent of how this spatial mapping is established and are quite robust against registration errors, we provide a relatively basic registration implementation. A more advanced and automated registration procedure is available in Abdelmoula *et al.* [104]. Since the registration needs to account for cutting artifacts, tissue deformations due to extraction and freezing of the brain, and other spatial perturbations, non-rigid registration techniques [110] [111] play a central role here. The second method utilizes the established link to interrogate the combined data sources for correlations. Correlation-based queries deliver insight into the spatial correlations between ion images in IMS and anatomical areas in the atlas. In previous studies the authors, as well as several other groups, have successfully demonstrated the potential of correlation-based

approaches to guide the user towards relevant findings [104,113] [114] [115] [116]. When the integration objective is complete, anatomical regions are implicitly annotated with the biochemical information obtained with mass spectrometry, and bio-molecular distributions are inherently mapped to a set of anatomical definitions.

The second objective moves beyond registration and simple correlation and uses the established IMS-anatomical atlas link to develop an automated anatomical interpretation method for IMS data. We define the anatomical interpretation of an ion image as decomposing the ion distribution into a combination of known anatomical areas that are tentatively tied to that specific ion. Anatomical interpretation becomes possible once a registration of IMS data to an atlas is available. Since it is independent of the particular method that was utilized to attain such registration, and it is assumed that registration errors are always present to some extent, automated anatomical interpretation methods can be developed largely orthogonal to any registration advancements. The interpretation method we develop can therefore be used in any setting where IMS is coupled to an atlas. Although the correlation-based queries demonstrate the functioning of the IMS-to-atlas mapping and provide a first step towards exploration of the combined data, they are insufficient to power automated anatomical interpretation of ion images. The main reason is their inadequate handling of ions that are present in several anatomical areas simultaneously. An automated anatomical interpretation of ion images therefore needs to be able to handle membership of an ion to multiple anatomical structures, and preferably should include a measure of abundance. To this end, the third computational method of this work develops an algorithm that uses the IMS-atlas link to automatically interpret any ion image in the IMS dataset as a combination of atlas-provided anatomical structures, without the need for human intervention.

4.2.1 Anatomy Data

The publicly accessible Allen Mouse Brain Atlas (AMBA) is used as the anatomical data source. This atlas is based on the brain of a 56-day-old C57BL/6J mouse, and has a user base that exceeds 10000 users per month [108]. The brain is dissected into 528 coronal tissue sections at 25 μm separation, which are Nissl stained, registered to each other, and assembled into a reconstructed brain volume. A low resolution (25 μm voxel width) gray-scale version of this reconstructed brain volume is accessible through the application programming interface provided by the AMBA website [117], and is imported into MATLAB 2012b (The Mathworks Inc., Natick, MA) to establish a local copy for further computation (Figure 4.1). Of the 528 coronal tissue sections, 132 are hand annotated and combined to create a 3-D anatomical reference atlas (Figure

4.1). This atlas is registered to the brain volume constructed from the Nissl stains and contains over 800 anatomical structures, which are organized in a hierarchical fashion. It can be consulted using the Brain Explorer 2 software (AMBA website).

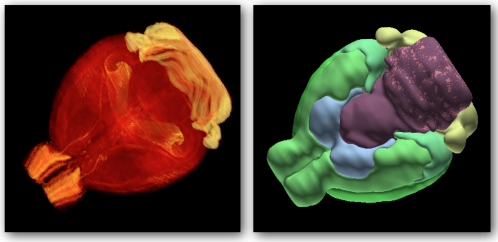

Figure 4.1: Registered microscopy and anatomy in the Allen Mouse Brain Atlas. (left) The microscopy volume contains 528 coronal Nissl stained tissue sections. These data are imported into and visualized using MATLAB, color-coding intensity from red to yellow. (right) The 3-D anatomical reference atlas contains over 800 anatomical structures. These data are visualized using the Brain Explorer 2 software.

4.2.2 IMS Data

Coronal tissue sections of 12 μm thickness were acquired from a healthy adult mouse brain that had been frozen in liquid nitrogen. Two neighboring sections from this brain were selected for IMS measurement, and mounted on ITO-coated glass slides. A third neighboring section was mounted on a glass slide and Nissl-stained for matching against the Nissl-stains of the AMBA. We give an overview of the staining and IMS measurements below and refer to Supplementary Information for full details. One IMS measurement focuses on protein imaging, acquiring ions between m/z 3000 and 22000. The other IMS measurement focuses on lipids with a m/z range from 400 to 1000. The tissue sections were sublimated with sinapinic acid (protein-oriented) and 1,5-diaminonaphthalene (lipid-oriented). The measurements were acquired on a Bruker Autoflex Speed MALDI-TOF mass spectrometer in positive linear mode with a laser spot size of 80 μm on target and at a pitch of 100 μm (protein-oriented) and in negative reflector mode with a laser spot size of 30 μm on target and at a pitch of 80 μm (lipid-oriented), using FlexControl 3.3 software. Approximately 100 shots/spot were acquired at a 1 kHz repetition rate using a Smartbeam II Nd:YAG laser. Image acquisition was carried out using FlexImaging 2.1 and further processing

took place in MATLAB. The spectra were normalized on the basis of their common ion current, disregarding differential peaks [55]. They were baseline-corrected using a spline approximation of the baseline at the 10%-quantile of ion intensities, and employing window sizes of 500 and 50 and step sizes of 250 and 25 for the protein and lipid-focused spectra respectively. The spectra were also optimally aligned along the m/z axis to reduce peak drift, allowing a maximum m/z shift of 12 and 0.5 for protein and lipid-focused spectra respectively. Both steps were performed using the Bioinformatics Toolbox of MATLAB (The Mathworks Inc., Natick, MA).

4.2.3 Registration of Data Sources

Coupling IMS data to the atlas requires the two data sources to be registered to each other, thus transforming them to a common coordinate system, in which their pixel locations describe the same space and can be directly compared [110, 111]. The IMS-atlas registration process requires multiple steps and uses both rigid and non-rigid registration to handle the complexities that are commonly encountered in tissue (e.g. deformation during extraction and freezing of the brain, differences between individual mouse brains, cutting artifacts, etc.). To deal with these complex deformation cases, non-rigid registration techniques [118, 119] are essential.

The IMS data are registered to the anatomical atlas via a modality common to both data sources: stained microscopy (Figure 4.2). The registration process entails: (1) rigid registration of IMS data to the experiment histology, (2) rigid registration of atlas data to the reference histology, and (3) non-rigid registration of the experiment histology to the reference histology. Although it is technically possible to register IMS data directly to the atlas, it is preferable to go through the histology data since the IMS data and the atlas data are of a substantially different nature with very differing spatial resolutions (resp. 80-100 μm and 0.3 μm). Using microscopy as an intermediate offers several advantages: (i) registration is more straightforward within a single modality, certainly given the complexity involved in non-rigid registration, (ii) the resolutions are high and comparable, and (iii) histological microscopy is readily available for most state-of-the-art IMS experiments. Details of the registration process can be found in Supplementary Information. To summarize, (1) the rigid registration between IMS data and experiment histology is performed in MATLAB through manual selection of fiducial markers, (2) the rigid registration from atlas to reference histology is provided by the AMBA, and (3) the non-rigid registration between experiment histology and reference histology is performed using the Medical Image Registration Toolbox (MIRT) by Myronenko [112], making use of a free form deformation (FFD) model.

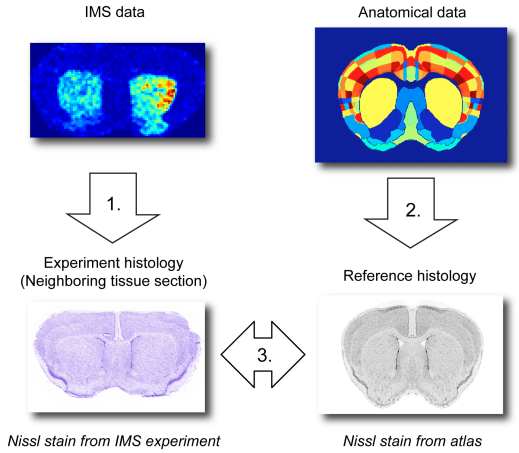

Figure 4.2: Workflow of the registration process. Spatially registering IMS data to anatomical data consists of 3 individual registration steps: (1) rigid registration of IMS data to experiment histology, (2) rigid registration of anatomical data to reference histology, and (3) non-rigid registration of experiment histology to reference histology.

4.2.4 Correlation-based Querying

Once registration is complete, it becomes possible to find out which anatomical structures of interest correlate with a measured ion distribution, or which ions of interest correlate with a certain anatomical zone of interest. To enable this simple form of correlation-based querying, we generate an anatomical structure image for each individual anatomical structure that is present in the tissue slice. Such an anatomical structure image (caudoputamen example at top in Figure 4.4A) represents the spatial location of a single anatomical annotation, and contains ones in locations where the anatomical structure is present, and zeros elsewhere. To soften the binary assignments somewhat at the edges of the structure, we apply a Gaussian filter, which eliminates crisp borders. Next, the spatial correlation between an ion image and an anatomical structure image is obtained by calculating the Pearson correlation coefficient between the intensities of both images over all IMS measurement locations. Only pixels for which both types of data are available are part of the correlation analysis to prevent bias. The analysis is performed by reshaping both 2D images to 1D vectors, removing any rows for which only one type of information is available, and then calculating the correlations between the resulting vectors. The correlation coefficient for each possible anatomical structure/ion image combination is calculated and collected into a correlation table for easy visualization and querying.

4.2.5 Automated Anatomical Interpretation

The goal of the anatomical interpretation method is to examine the pattern in an ion image, and, without human intervention, determine which anatomical structures are involved and what their ion intensity contribution is. In other words, once an ion image is mapped to the atlas (using the proposed registration pipeline or an automated variant thereof [104]), the interpretation method takes that ion distribution pattern as an input and then automatically decomposes it into a combination of atlas-provided anatomical structures.

At its core, anatomical interpretation is a problem of approximating the spatial pattern of an ion with a combination of patterns selected from a provided vocabulary of anatomical patterns. The model we employ in our algorithm considers an ion image to be a sum of products, each product multiplying a pattern from the finite set of anatomical patterns with its contribution coefficient. Since we know both the ion image and the anatomical patterns and their relationship is established by the model, the search for the optimal anatomical contribution coefficients (and thus the optimal anatomical interpretation) can be approached as a multivariate optimization problem.

Consider the set of anatomical structure images previously defined in the section on correlation-based querying, which contain high values in pixels that are a member of the structure and low values in pixels that are not. We represent each of those anatomical images ϕ^m as a vector of length $K \in \mathbb{N}_0$ where K denotes the total number of pixels in the image, or all of them combined as

$$\{\phi^m \in \mathbb{R}_+^K\}_{m=1}^M \tag{4.1}$$

where $M \in \mathbb{N}$ is the total number of anatomical images. Note that the anatomical images are positive by construction since they represent membership to an anatomical structure. For our examples this formulation can even be simplified to $\{\phi^m \in [0,1]^K\}_{m=1}^M$. Similarly, let the ion image we want to decompose be described by a positive vector $q = [(q_1, ..., q_K)]^T \in \mathbb{R}_+^K$ of length K. Due to the IMS-atlas link, the anatomical images ϕ^1 to ϕ^M contain the same number of pixels as the ion image q. The ion image q is also positive by construction since its pixels encode ion counts. The algorithm now seeks the optimal (and smallest) combination of anatomical images that, when multiplied by their contribution coefficients, add up to the target ion image. It is this optimal profile of intensity contribution coefficients p that answers our anatomical decomposition question. Anatomical images, and thus structures, corresponding to high absolute coefficients are important for approximating the ion distribution of interest, and are therefore considered part of the anatomical interpretation of that ion image. Anatomical structures with low absolute coefficients are

considered unrelated to the ion image of interest. The following linear model is adopted for any $k = 1, ..., K$

$$q_k = \sum_{m=1}^{M} \phi_k^m p_m + \epsilon_k \tag{4.2}$$

where the coefficients $p = (p_1, ..., p_M) \in \mathbb{R}^M$ encode the assumption that the ion image is essentially a weighted sum of the anatomical images/structures, up to the residual $\epsilon = (\epsilon_1, ..., \epsilon_K)^T \in \mathbb{R}^K$. Our implementation uses a classical approach to approximate the linear coefficients we are looking for, minimizing the squared norm of the residuals, or

$$p^* = \arg\min_p \sqrt{\sum_{k=1}^{K} \big(\sum_{m=1}^{M} \phi_k^m p_m - q_k \big)^2} \tag{4.3}$$

This optimization problem considers the optimal contribution coefficients to be the ones that minimize the difference between the measured ion image and its approximation using anatomical patterns. However, since the anatomical structures are stored in the Allen Mouse Brain Atlas in the form of a hierarchical tree, many anatomical patterns are very similar to each other, exhibiting large overlap and often differing by only a small sub-area. In the current formulation, this could cause multiple similar structures to be selected, dividing that area's contribution among them. This could result in unnecessarily complex anatomical interpretations consisting of a large numbers of structures. Instead, we want the algorithm to select only the best fitting anatomical structures for its interpretation, disregarding similar but less optimal ones, and keeping the total number of anatomical structures in the interpretation to a minimum. To accomplish this, we introduce an L1 regularization term for an appropriate choice of $\gamma \geq 0$, turning the anatomical interpretation into

$$p_\gamma^* = \arg\min_p \left[\sqrt{\sum_{k=1}^{K} \big(\sum_{m=1}^{M} \phi_k^m p_m - q_k \big)^2} + \gamma \sum_{m=1}^{M} |p_m| \right] \tag{4.4}$$

This formulation is based on the LASSO [120] algorithm. However, our approach differs from the formulation of the original LASSO algorithm in that it aims to minimize the Euclidean distance between the measured ion image and its approximation, whereas the original formulation minimizes the squared errors between the two (discarding the square root). Essential in this approach, however, is that the 1-norm makes the optimization problem pursue sparsity in the solution of coefficients. This can be interpreted as follows: if the optimization problem does not know which solution p^* to prefer up to a numerical quantity,

then choose the solution with the smallest 1-norm. Parameter γ regulates what is meant by such a numerical quantity: if γ is large, a sparse solution containing only a few significant anatomical contribution coefficients will be preferred over an exact least squares fit of the ion image, while the reverse holds true for when γ is small. In our case studies, we set $\gamma = 1$. Given the number of pixels versus the number of anatomical structures involved in these datasets, this setting ensures that approximation of the ion image is more important than sparsity in the anatomical interpretation, while the sparsity term does keep the anatomical explanation to a minimum if multiple valid interpretations exist.

Our implementation uses CVX, a package for specifying and solving convex programs [121, 122], to solve the optimization problem for each ion image we want to interprete. The anatomical patterns are used as building blocks to construct an approximation of each ion image, and the coefficients specify how each anatomical structure contributes to the overall approximation. A nice feature of the method is that an anatomical contribution coefficient tends to be proportional to the ion intensity in that anatomical structure, inherently assigning a notion of importance to each anatomical zone involved. Anatomical images, and thus structures, corresponding to high absolute coefficients are important for approximating the ion distribution of interest, and are therefore considered part of the anatomical interpretation of that ion image.

Note that the coefficients are not constrained to positive values in this formulation. This allows the anatomical interpretation to say things like "the ion seems to be present in zone A plus zone B minus zone C." Although it will usually result in a more intricate interpretation using lots of little structures, if desired, it is possible to constrain the algorithm to only consider additive interpretation by setting a positivity constraint on the coefficients in the optimization problem.

4.3 Results & Discussion

We demonstrate the developed methods both in a protein-focused and a lipid-focused case study of coronal mouse brain sections, illustrating the potential of incorporating anatomical information into IMS analysis.

4.3.1 Results on Registration of the Data Sources

A non-rigid registration algorithm is used to register the experiment histology (Fig. 4.3A.a) to the atlas reference histology (Fig. 4.3A.d). Although the experiment histology is strongly deformed compared to the reference

histology, the registration results (Fig. 4.3A.c) demonstrate that the non-rigid registration can deal well with these soft-tissue deformations. The FFD (Free Form Deformation) mesh of control points (Fig. 4.3A.b) that constitutes the transformation between the two images, shows how the experiment histology image is 'warped' to register to the atlas histology. By using a FFD transformation mesh we ensure that neighboring IMS pixels remain neighbors, and prevent excessive distortion of the original data. Figure 4.3B shows how the transformation established by the non-rigid registration process is used to transform and project data from one data source to the other and vice versa. The non-rigid transformation is reversible (up to a rounding error) and there is a one-to-one relationship between locations in "IMS space" and locations in "atlas space". This means that we are able to project IMS data onto the reference histology (Fig. 4.3B top right). Such a projection effectively brings empirical MS measurements acquired from imperfectly sectioned or deformed tissue into the reference shape of the mouse brain, and it directly links these observations to reference atlas annotations. Similarly, we can project the atlas annotations onto the experiment histology (Fig. 4.3B lower left). This direction of projection draws annotations that are typically defined in an ideally shaped version of the mouse brain into the practical tissue sample that was sectioned. Casting the anatomy information to the experiment histology also enables quick visual verification of the correctness of the projection since the neighborhood structure of the different subareas should be retained regardless of tissue deformation.

4.3.2 Results on Querying the Correlation between Anatomy and Ions

Once registration of the two data sources is complete, we construct a correlation table by calculating the spatial correlation between the anatomical structure images and the ion images of the peak picked IMS protein data. The correlation table (available in Supplementary Information) allows two types of queries: anatomical queries and ion queries.

Anatomical Query

The anatomical query provides an answer to the question "Which ions are specific to anatomical region X?". To demonstrate, we use the caudoputamen as an example anatomical structure. Figure 4.4A shows the input, the anatomical structure image of the caudoputamen, and displays the part of the correlation table that pertains to this structure, highlighting its spatial correlation to the various ion images. Ions m/z 12650 (ρ=0.77) and m/z 21832 (ρ=0.70) are two

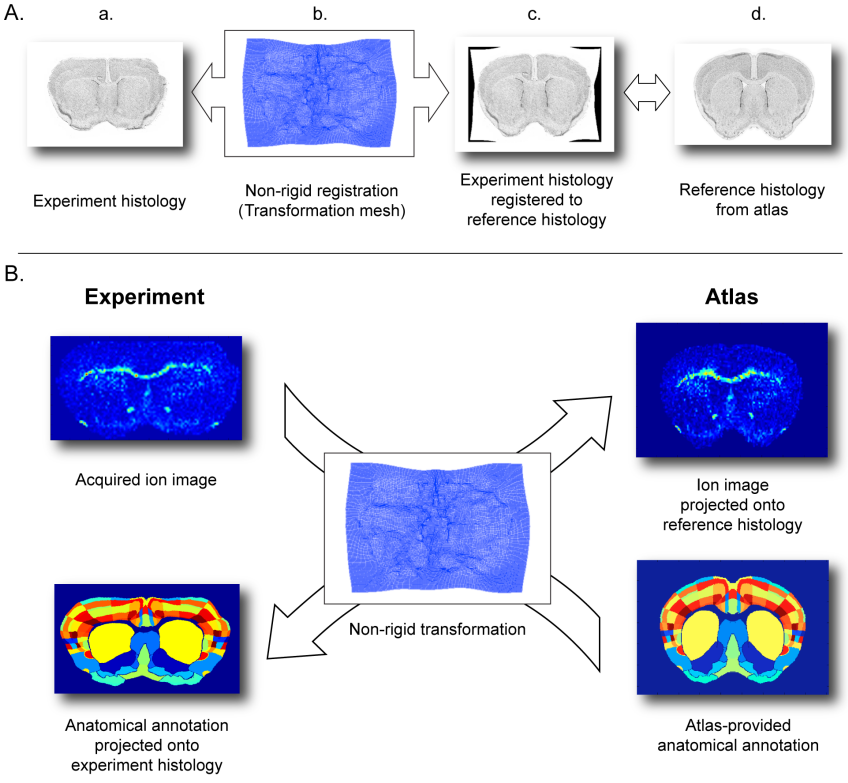

Figure 4.3: Overview of registration results. (Panel A) Non-rigid registration of the experiment histology onto the reference histology from the atlas. Despite the large initial deformation of the experiment histology (a), the non-rigid transformation (b) successfully compensates and registers (c) onto the reference histology (d). (Panel B) The non-rigid registration enables traversal from the coordinate system of one data source to the other: the acquired ion image can be projected onto the reference histology for direct comparison to the anatomical annotations. Inversely, the anatomical annotations, provided by the atlas, can be projected onto the experiment histology, to provide anatomical guidance within the sectioned tissue.

examples of ion images that exhibit high positive correlation in this case. These ions show a clear spatial overlap with the anatomical structure image of the caudoputamen, and are almost exclusively located therein.

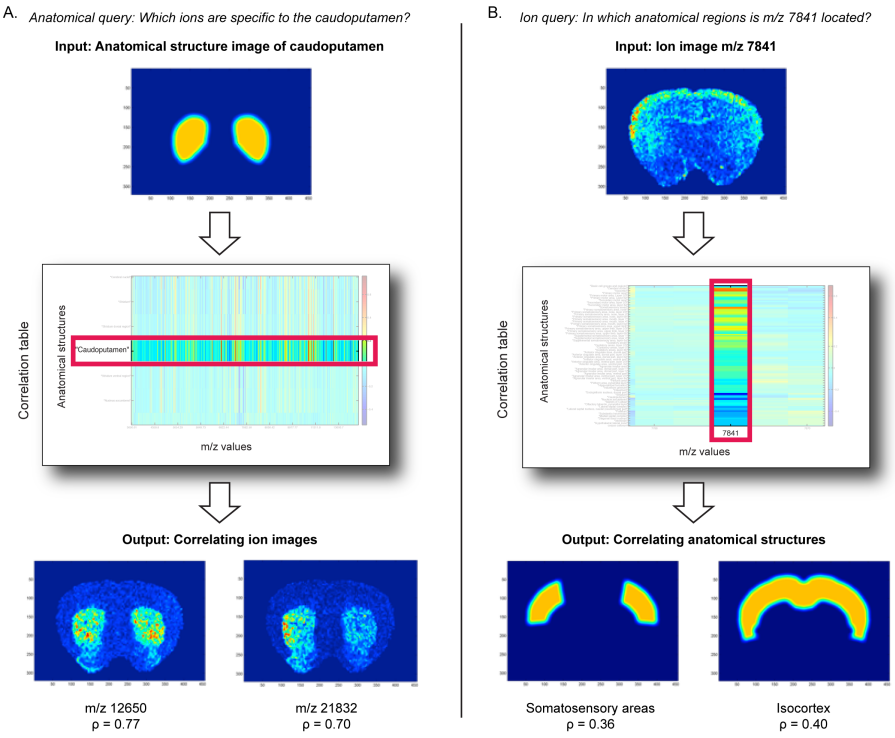

Figure 4.4: (Panel A) Example of an anatomical query, finding ions specific to the caudoputamen. The anatomical structure image of the caudoputamen is given as input and the correlation table returns the spatial correlation to this structure for each ion image. Two examples of ion images that positively correlate with the target anatomical structure are displayed. (Panel B) Example of an ion query, finding anatomical regions in which m/z 7841 is highly expressed. The ion image of m/z 7841 is given as an input and the correlation table returns the spatial correlation to this ion image for all the anatomical regions. Two examples of anatomical structure images that positively correlate with the target ion image are displayed.

Ion Query

The ion query provides an answer to the question "In which anatomical regions is ion Y located?". To demonstrate this query, we use the ion image for m/z 7841 as an example. Figure 4.4B shows the ion image as input and, analogous to the anatomical query, it displays the part of the correlation table that pertains to this ion. The isocortex (ρ=0.40) and the somatosensory areas (ρ=0.36), which are substructures of the isocortex, are two examples of anatomical structures that exhibit high positive correlation with m/z 7841. Additional examples are provided in Supplementary Information.

Correlation-based queries can deliver fast insight into relationships between ions and anatomical structures, but have several disadvantages. First, it is difficult to define a generic threshold to determine when these correlations become significant. Second, since several thousands of correlations are being calculated in parallel, the multiple testing problem needs to be considered when drawing conclusions from these results. However, the most important roadblock for using correlation towards automated anatomical interpretation is the concept of "multi-membership". Correlation considers only the relationship between a single anatomical structure and a single ion. An ion that appears in several anatomical structures simultaneously will exhibit a relatively low correlation to each of the individual anatomical structures that it is a member of. Such an ion will not give a strong signal in the correlation table, and could go undetected as a result. In these multi-membership situations, which are quite common in most biological tissue types, correlation-based querying falls short and is not capable of dealing with the complexities of the biology. In fact, in such complex cases, any univariate querying strategy will provide skewed results.

4.3.3 Results on Automated Anatomical Interpretation

Since the membership of an ion to multiple anatomical structures cannot be clearly captured by a univariate strategy, the use of correlation to drive automated anatomical interpretation of ion images is limited. Instead, a more advanced approach, using multivariate models to account for multi-membership, is necessary. The Methods section introduces a linear model, capable of capturing the multi-membership aspect effectively, to tie ion image patterns to anatomical structure patterns. By applying convex optimization to this model and the given ion and anatomy patterns, it is possible to obtain an optimal anatomical explanation for each ion image. Figure 4.5 shows the automated anatomical interpretation of several ion images from the lipid case study, using our method. Each ion image gets interpreted, automatically and without user-intervention, as an optimal combination of anatomical zones from the AMBA. The interpretation

is optimal in the sense that it selects the combination of atlas patterns that gives the closest approximation of the measured ion image. If there are multiple combinations that come equally close, the combination with the least amount of anatomical structures (and thus the simplest explanation) is selected. In short, the automated anatomical interpretation method provides (i) the closest approximation of the measured ion image using atlas structures (Fig. 4.5, right, middle), and (ii) an overview of the contributing anatomical structures, specifying name, reference location, and contributing intensity or weight in the interpretation (Fig. 4.5, right, middle). This visualization delivers quick insight into the major anatomical zones associated with an ion image, while also providing a notion of the relative contributions of each underlying structure involved. The example in panel A shows a clearly defined expression of the ion m/z 742.57 in the fiber tracts and ventricles (more specifically the corpus callosum, which is not annotated as a separate region in the AMBA), the pallidum, and the somatosensory areas. Using negative contribution coefficients, the algorithm also reports areas where the ion exhibits reduced presence, such as in the striatum. Panel B shows the ion m/z 723.53, which is specifically absent from the somatosensory areas. The strong congruence between empirically observed patterns in the ion distributions and anatomical structures extracted from the atlas, indicate good spatial mapping between these data sources, and strong biological signals in the IMS measurements. Additional examples of the in total 1405 ion images that we applied the automated anatomical interpretation methodology to are provided in Supplementary Information.

It should be noted that the heads-up-display type visualization of Figure 4.5 is meant for human consumption, and is but one of many possible representations of the interpretation results. In essence, the interpretation algorithm provides for each ion image a set of contribution coefficients, one for each structure in the atlas. Each contribution coefficient can be considered to report a degree of membership of an ion to a particular structure. Custom visualizations of these coefficients can be developed as demanded by the application. In case of a machine-based follow-up, no visualization will be necessary at all and the anatomical membership coefficients for each ion image in an IMS experiment can be passed on directly to the next computational step in the analysis. The automated breakdown of an ion image into contributing anatomical structures is a powerful tool, particularly for the histological non-expert. It supplies the researcher directly with the relationships between an ion and the anatomical structures in which it is expressed. The interpretation algorithm can substantially aid in unraveling the function of biomolecular ions. It can incorporate the body of pathological research that is currently publicly available into the analysis of an individual IMS experiment, and this without much additional effort. Essentially, the results of the analysis constitute a table of anatomical membership coefficients, linking each ion image to each anatomical structure. This table can be queried in the

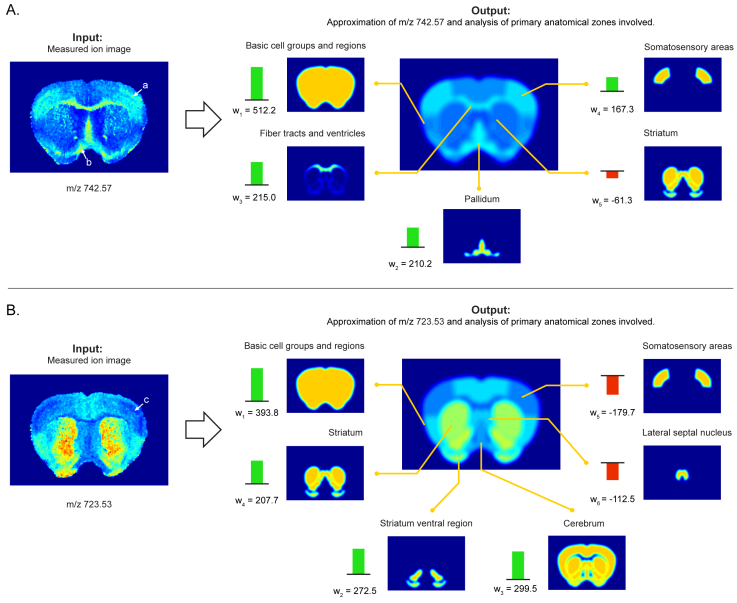

Figure 4.5: Examples of automated anatomical interpretation. When an ion image is given as input (left), the interpretation method provides an optimal anatomical explanation for the observed ion pattern (right), using the library of provided anatomical structures. Specifically, the ion intensity pattern is decomposed without user intervention into an optimal combination of contributing anatomical structures from the atlas. The interpretation method provides (i) the closest approximation of the measured ion image using atlas structures (right, middle), and (ii) an overview of the contributing anatomical structures, specifying name, reference location, and contributing intensity or weight in the interpretation (right, outer ring). This visualization delivers quick insight into the major anatomical zones associated with an ion image, while also providing a notion of the relative contributions of each underlying structure involved. Negative weights indicate a relative decrease of the ion in those areas. (panel A) Ion m/z 742.57 is highly expressed in the somatosensory areas, the fiber tracts, and the pallidum. (panel B) Ion m/z 723.53 shows a decrease specifically in the somatosensory areas, as indicated by the negative weight. The empirical ion distributions show good congruence with the boundaries of the anatomical structures defined in the atlas, indicating good spatial mapping between the data sources, and strong biological signals in the IMS measurements.

same way as the correlation table from the correlation-based queries, but does away with the disadvantages of a univariate approach. Most importantly, the multi-membership of ions to different anatomical zones is now taken into account, providing the user with a much more complete and reliable list of anatomical-structure-to-ion relationships. The greatest benefit of an automated anatomical interpretation method is the potential for parallelization. A computational interpretation method can deliver anatomical interpretations for every ion image in an experiment, even if the number of ion images runs into the hundreds or thousands. It can provide these interpretations concurrently, without human intervention, and in a single calculation. This avoids the labor and time-intensive step of having a histological expert manually interpret hundreds of ion images, and the potential risk for human bias that comes with it. It also enables the researcher to pursue broad exploratory measurements, after which the focus can be narrowed to only the most promising ions that have demonstrated a relationship to a particular anatomical structure of interest. A possible downside to using an anatomical atlas for the interpretation of IMS data is that zones that have not been previously discovered in pathological and anatomical research and thus are not present in the atlas, cannot be found in the ion images and cannot be part of their interpretation. However, examining those ions that cannot be adequately approximated with an atlas could provide a route to the automated discovery of new physiological or chemical sub-divisions within tissue areas that are considered to be homogeneous by anatomy. We elaborate on this in the Supplementary Information.

4.4 Conclusions

Establishing a link between an IMS experiment and an anatomical atlas can serve as an important accelerator both for human and machine-guided exploration of IMS experiments. As the number and complexity of IMS experiments increases, and studies aim to compare a growing number of IMS experiments to each other, the importance of automated approaches to filter the massive datasets for patterns of interest will become increasingly important. The automated anatomical interpretation of ion images can serve as a formidable catalyst for IMS analysis, due to its speed and ability to interpret thousands of ion images concurrently without human supervision. On a less application-specific scale, a mapping between curated anatomy and IMS data can make a body of anatomical research available to any IMS-related computational method, to integrate into its analysis and improve its performance. In the case of multiple IMS experiments and multiple mappings to the same atlas, one can compare the anatomical interpretations of multiple experiments to each other by using a common reference (atlas) space. This would allow, for example, the automated

detection of differences in activities of anatomical structures between healthy versus diseased tissues. Mapping multiple IMS datasets to a common reference coordinate system would also enable the creation of an IMS-based chemical reference atlas for lipidomics, proteomics, and metabolomics. In the specific case of the Allen Mouse Brain Atlas this offers several promising perspectives for the future, since IMS-based data could then even be combined with non-IMS data sources, such as gene expression and neuron connectivity measurements, which are currently being linked to the Allen Brain Atlas by other research initiatives. The combination of these data sources can serve as a valuable multi-modal resource for systems biology research, bringing together insights from many different technologies and fields.

4.5 Supplementary Information

4.5.1 IMS Data Acquisition and Microscopy Staining Details

Materials Ethanol, acetonitrile (ACN), and acetic acid were purchased from Fisher Scientific (Suwanee, GA), xylene was purchased from Acros (Morris Plains, NJ), chloroform, 1,5-diaminonaphthalene (DAN), and methanol were purchased from Sigma-Aldrich (Milwaukee, WI). Sinapinic acid (SA) was purchased from Oakwood Products, Inc (SC) and recrystallized twice with 70% ACN. Conductive indium tin oxide (ITO) coated microscope glass slides were purchased from Delta Technologies (Stillwater, MN). Cresyl Violet (for Nissl stain) was purchased from Electron Microscope Sciences (Hatfield, PA). Milli-Q water was from a Milli-Q Advantage A10 Ultrapure Water Purification System (Millipore, Billerica, MA). Carnoy's fluid was prepared from 60 mL of ethanol, 30 mL of chloroform, and 10 mL of acetic acid. Fresh frozen mouse brain was purchased from Pel-Freez Biologicals (Rogers, AZ), and was sectioned using a Leica CM3050 cryostat (Leica Microsystems GmbH, Wetzlar, Germany). Frozen tissue sections were thaw mounted on cold ITO coated microscope slides. Serial sections were prepared for MS imaging and optical imaging.

Sample preparation for imaging mass spectrometry of proteins Briefly, mouse brain sections of 12 μm thickness were rinsed with 70% ethanol 30 s, 100% ethanol 30 s, Carnoy's fluid 2 min, 100% ethanol 30 s, H2O 30 s, and 100% ethanol 30 s. The sections were then dried for 10 min under ambient conditions and sublimated with sinapinic acid to obtain a coating of 0.2 mg/cm^2. The SA coated slides were then treated with H2O: acetic acid (1 mL: 50 μL) vapor to recrystallize [25] the SA coating for 3.5 minutes under 85 degC.

Sample preparation for imaging mass spectrometry of lipids Tissue sections of 12 μm thickness were sublimated with DAN to obtain a matrix coating of 0.2 mg/cm^2 and were then treated with vapor of water:methanol:chloroform (40:30:10) at 85degC for 2 min. The procedure is similar as in the literature mentioned above, but with 2 min and a total volume of 30 μL of liquid.

Nissl staining of tissue sections for microscopy Literature reference: Histological and Histochemical Methods: Theory and Practice, 4th edition, edited by J.A. Kiernan, Publisher: Cold Spring Harbor Laboratory Press; 4th edition (March 1, 2008), ISBN-10: 1904842429

Procedure:

1.Dehydrate: 75% ethanol for 30 sec 95% ethanol for 30 sec 100% ethanol for 30 sec 100% ethanol for 30 sec Xylene for 30 sec Xylene for 30 sec Xylene for 30 sec

2.Rehydrate and stain: Xylene for 30 sec Xylene for 30 sec 100% ethanol for 30 sec 100% ethanol for 30 sec 95% ethanol for 30 sec 75% ethanol for 30 sec Milli-Q H2O for 30 sec Cresyl violet for 10 min

3.Destain: Milli-Q H2O 75% ethanol 30 sec 95% ethanol 30 sec 100% ethanol for 30 sec 100% ethanol for 30 sec Xylene for 30 sec Xylene for 30 sec Xylene for 45 min

4.Coverslip The stained sections were scanned using a Mirax slide scanner from Zeiss with 0.33 μm spatial resolution.

Mass spectrometry and data analysis MALDI MS analyses were performed on a Bruker Autoflex Speed MALDI TOF mass spectrometer in positive linear mode (proteins) and a Bruker Autoflex Speed TOF/TOF mass spectrometer in negative reflector mode (lipids) using FlexControl 3.3 software. Approximately 100 shots/spot were acquired with a 1 kHz repetition rate Smartbeam II Nd:YAG laser. Image acquisition was carried out using FlexImaging 2.1 and spectral analysis was performed with FlexAnalysis 3.3 and MATLAB.

4.5.2 Registration Process Details

This section gives a more detailed description of the registration process.

Registration of IMS data to experimental histology

The first registration step maps IMS locations (at 80 and 100 μm resolution for lipids and proteins respectively) to the experimentally acquired microscopy image of the neighboring Nissl-stained tissue section (0.3 μm res.), further referred to as the experiment histology.

While it is common practice in IMS experiments to overlay ion images on a neighboring tissue section for visualization purposes, extra care is taken since here this alignment will become part of a larger registration pipeline. The alignment is based on the use of an affine transformation matrix and involves several processing steps:

- Selecting an m/z that clearly co-localizes with a well defined anatomical structure; Here, ion m/z 18411 was selected, which clearly delineated the corpus callosum (Figure 4.6), a good reference structure for registration.
- Loading both the ion image (m/z 18411) and the experiment histology into MATLAB and selecting landmark points using the cpselect function.
- Using these landmarks to create the affine transformation matrix that determines the registration by employing the MATLAB function cp2tform.



Figure 4.6: Affine registration of m/z 18411 to the experimental histology.

Registration of atlas data to reference histology

The second step registers the anatomical atlas to the microscopy of the reference Nissl stain, which is included with the atlas and hereafter referred to as the reference histology. The registration from atlas to reference histology is taken care of by the Allen Mouse Brain Atlas, which provides these modalities already in a registered state, making step 2 for our purposes a straightforward identity transformation. Out of the reference stains provided with the atlas, the closest structural match to the experiment histology is manually selected, identical

to how a medical professional would select a reference tissue in the atlas for anatomical interpretation. Alternatively, the closest matching tissue can be selected automatically using methods such as those proposed by Abdelmoula *et al.* Automatic selection can be helpful in certain situation (e.g. when a large number of tissues needs to be registered at once), but is not always desirable (e.g. in the case of partially matching tissues due to a skewed cut, where a winner-takes-all mechanism may give incorrect results). Since the selection needs to happen only once for each IMS experiment, taking care of the spatial mapping for potentially thousands of ion images in a single action, the selection mechanism is typically not a bottleneck to the interpretation efforts and its implementation can be tailored to the particular study at hand. The reference stains are all in gray-scale and have been down-sampled to 25 μm resolution.

Registration of experimental histology to reference histology

The registration between the experiment histology and the reference histology is the most difficult step in the registration process, as it has to account for tissue deformation during extraction and freezing of the brain, and cutting artifacts. As can be seen in Figs. 4.7 and 4.9, the experiment histology has a relatively high deformation compared to the reference histology from the Allen Brain Atlas. While this may seem problematic for a correct registration of both images, this deformation can be accounted for by using non-rigid image registration techniques.

Preliminary processing

Difference in resolution As the reference histology is provided at 25 μm resolution, and the experiment histology has a resolution of 0.33 μm, an intermediate resolution of 5 μm is used for the registration. Thus, the experiment microscopy picture is down-sampled with a factor of 15 and the reference microscopy is up-sampled with a factor of 5, using the imresize function provided in MATLAB.

Difference in color scheme As the reference microscopy is provided in grayscale, the experiment microscopy is converted to grayscale as well.

Rigid registration The experiment histology is first centered and scaled to the reference histology using a rigid registration, in order to get a good starting point for the non-rigid registration. This is done in MATLAB by selecting

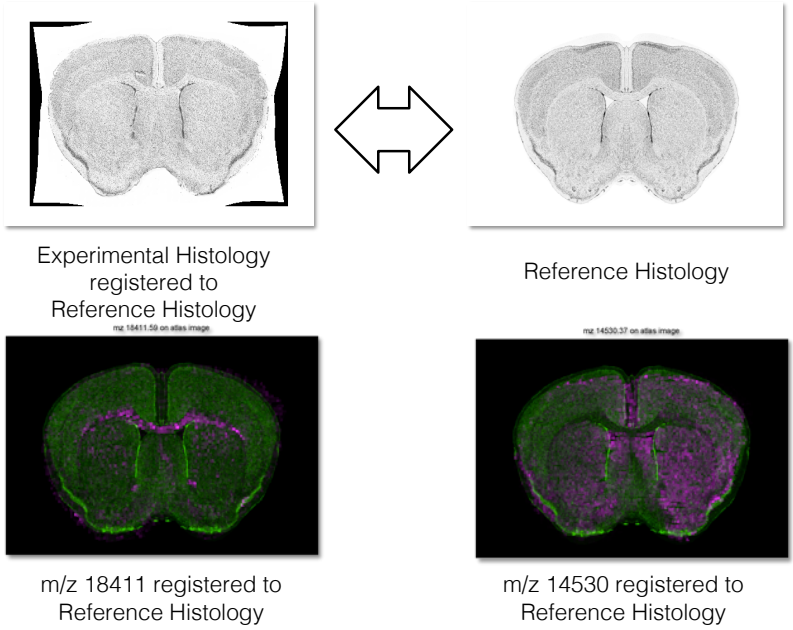

Figure 4.7: The experiment histology has a relatively high deformation compared to the reference histology in the Allen Brain Atlas. The need for non-rigid registration is apparent.

landmark points using the cpselect function. These landmarks are then used to create the affine transformation matrix that determines the registration by employing the MATLAB function cp2tform.

Non-rigid registration The non-rigid registration is performed using the algorithms from the Medical Image Registration Toolbox (MIRT) by Myronenko [112]. A non-rigid registration contains a similarity measure and a transformation model. The similarity measure defines how pixels are compared between images, while the transformation model determines the types of non-rigid transformations allowed.

Similarity measure The squared correlation coefficient is used as a similarity measure to compare the intensities of pixels across the images. The squared correlation coefficient is maximized when the images are linearly related, and is

defined as

$$E_{CC} = \frac{\left(\sum_{n=1}^{N}(I_N - \bar{I})(J_N - \bar{J^T})\right)^2}{\sum_{N=1}^{N}(I_N - \bar{I})^2 \sum_{N=1}^{N}(J_N - \bar{J^T})^2} \tag{4.5}$$

where $\bar{I}$ is the intensity mean of image I and $\bar{J}$ is the intensity mean of image J. The medical image processing literature shows that the best choice of similarity measure is dependent on the particularities of the data at hand. For these case studies, the squared correlation coefficient empirically showed sufficiently good results, so no further benchmarking of similarity measures needed to be pursued. Further improvements of the registration phase are welcome but not essential to the correct functioning of the anatomical interpretation phase, since its optimization formulation allows the interpretation output to degrade gracefully as the number of registration errors rises. The entire registration procedure was run on a desktop computer.

Transformation model As a transformation model we use the Free Form Deformation (FFD), which is a popular non-rigid registration model that uses a mesh of control points (example in Figure 4.8). In this mesh, each control point is connected to its neighbor, and each location in between these control points is interpolated via B-spline basis functions. The number of admissible transformations is limited by the spacing of the control points and by the smoothness of the B-spline basis functions. The degree of non-rigid deformation that can be modeled depends on the resolution of the mesh of control points, which in turn determines the computational complexity. In order to have an optimal trade off between computational complexity and admissible deformation, the MIRT uses a hierarchical multi-resolution approach, as suggested by Rueckert *et al.* [119], in which the resolution of the control mesh is increased, along with the image resolution, in a coarse to fine fashion. The initial mesh size we use consists of 8x8 control points, with 4 consecutive hierarchical levels. Using the results of the affine registration as a starting point, the experimental histology is registered to the reference histology using the above settings. The result of the registration can be seen in Figure 4.9, which shows very good matching despite the relatively large initial deformation.

4.5.3 Additional Correlation-based Querying Examples

Protein case study examples Figure 4.10 shows the correlation table that results from the spatial correlation analysis between the anatomical structure images and the ion images of the peak picked IMS protein data. The y-axis shows all the anatomical structures at this depth of the brain (119 in total). The x-axis shows the peak picked m/z values. Anatomical structure/ion abundance

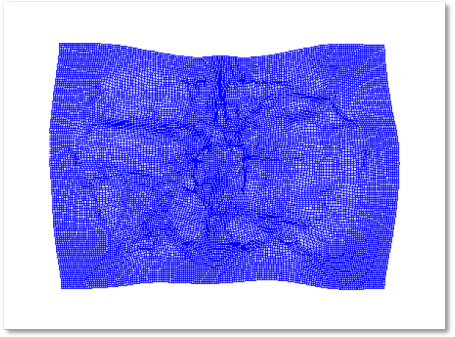

Figure 4.8: The Free Form Deformation registration method uses a mesh of control points to register the experimental histology to the reference histology. The number of admissible transformations is limited by the spacing of the control points and by the smoothness of the B-spline basis functions.

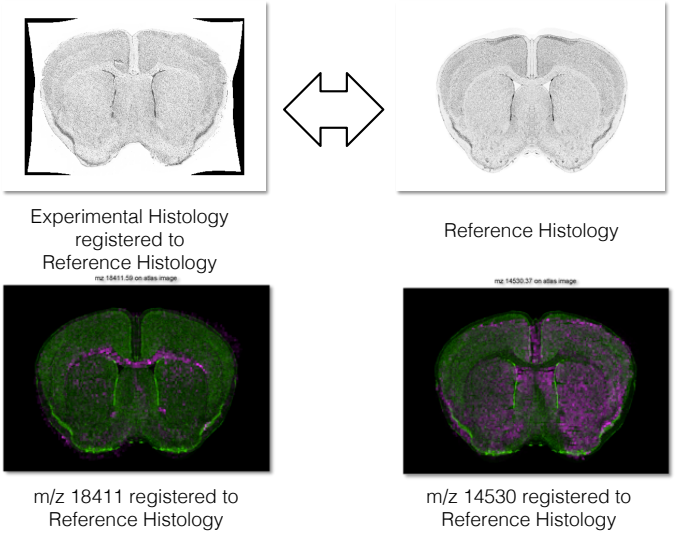

Figure 4.9: Results of the non-rigid registration process. Despite the high initial deformation of the experiment histology (see Fig. S-2), the registered version looks very similar to the reference histology. The lower images show examples of using this non-rigid transformation to register IMS data to the reference histology.

combinations that have a high positive correlation (range $\rho = [0.540.76]$) are shown in dark red. Combinations that have a high negative correlation are indicated in dark blue. The other combinations exhibit low correlation ($|\rho| < 0.15$).

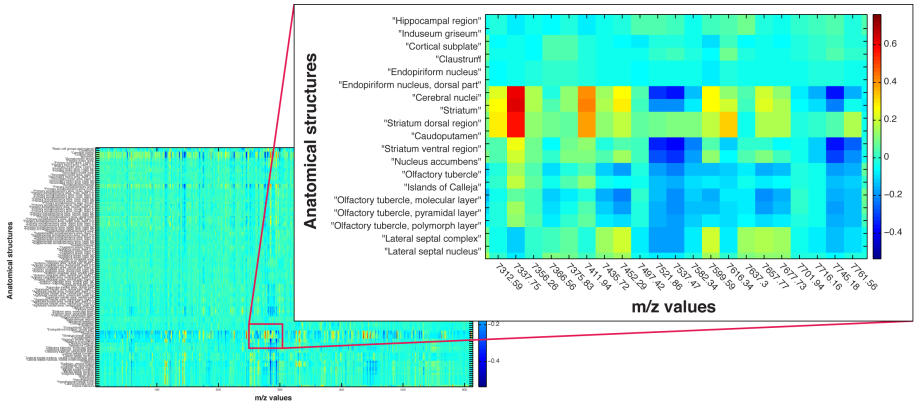

Figure 4.10: The correlation table displaying spatial correlation between the anatomical structure images and ion images of the peak picked IMS protein data. The y-axis shows all the anatomical structures that are present at the depth in the brain that this IMS experiment was acquired at. The x-axis shows the peak picked m/z values. Red = high positive correlation, blue = high negative correlation, green = low correlation.

Figure 4.11 shows an extended overview of the correlation-based queries presented in the main paper, including examples of negative and near-zero correlation. Ions m/z 6279 and m/z 11247 are two randomly selected ion image examples that exhibit low spatial correlation ($|\rho| < 0.15$) with the caudoputamen, while ions m/z 11842 and m/z 6961 describe ion images that have high negative correlation with the target structure, exhibiting a clear absence of these ions therein. The anti-correlating examples demonstrate an interesting feature of the algorithm, which is that the anatomical query can be used not only to find ions specifically present in the anatomical structure of interest, but that it can also be used to search for ions specifically absent from the region of interest.

Figure 4.12 shows several examples of positively correlated anatomical structure/ion image pairs in the protein dataset. Example 4.12A shows the correlation of the pallidum (ventral region) and m/z 4538, which is almost exclusively located in this area. The congruence of these relatively small areas demonstrates the accuracy of the registration. Figure 4.12C shows m/z 18483, which is well localized in the corpus callosum. Unfortunately, the corpus callosum is currently not available as a separate anatomical area in the Allen

Brain Atlas, and is a substructure of the fiber tracts and ventricles, which shows high correlation with this ion in the current tissue slice. Figure 4.12G shows the correlation between the lateral septal complex and m/z 9751. From this image we can see that m/z 9751 is not exclusively located in the septal complex, but rather in multiple anatomical structures. This example emphasizes the need for a multi-membership approach.

Lipid case study examples Figure 4.13 shows the correlation table for the lipid data (negative mode) with the atlas data. It shows areas of high positive, high negative, and low correlation. Figure 4.14 displays several examples of anatomical structure/ion abundance combinations with high positive correlation.

4.5.4 Additional Automated Anatomical Interpretation Examples

4.5.5 Sparsity in Automated Anatomical Interpretation

From the equation of the convex optimization formulation (eq.4.4) it is clear that γ is the only parameter in the anatomical interpretation model. Parameter γ controls how much emphasis is placed on the approximation of the measured ion pattern versus the simplicity (sparsity) of the anatomical explanation. This parameter functions essentially as a dial that allows the user to choose whether a broad overview anatomical interpretation is provided, capturing the major anatomical trends of the ion pattern, or whether a more detailed interpretation is returned, that breaks those larger anatomical zones into more precise sub-zones. The γ dial of the algorithm has a range from 0% sparsity (closest approximation, most detailed anatomical interpretation) to 100% sparsity (approximation of only the broader trends, less complex anatomical interpretation).

Which level of interpretation complexity is preferable is dependent on the particular purpose that the anatomical interpretation will serve. If it is meant for human consumption, it is often not preferable to have a very intricate interpretation using twenty or more anatomical labels. Instead, a medical professional typically wants a broad breakdown into the five or six major anatomical zones involved, unless they have a particular case-study reason to request otherwise. If the interpretation is to be used as annotations for a computational follow-up (e.g. classification or clustering that includes anatomical labels as extra information), there is little reason not to go for highly detailed interpretations since the computer will easily handle these. Regardless of the particular use case, the same model is capable of delivering these interpretations. Given these examples, it is clear that there is not one

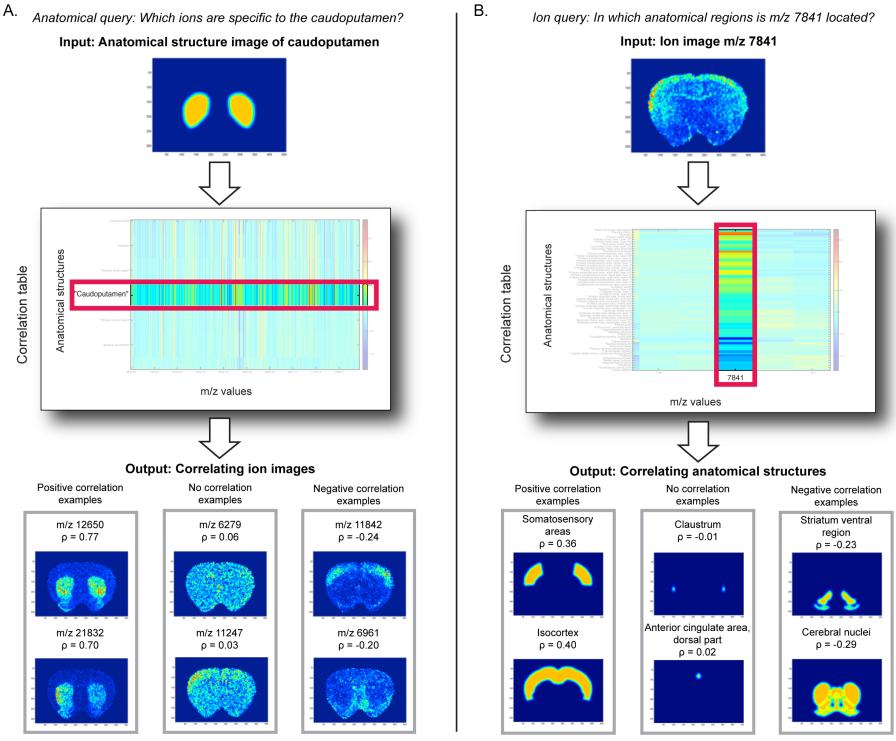

Figure 4.11: (Panel A) Example of an anatomical query, finding ions specific to the caudoputamen. The anatomical structure image of the caudoputamen is given as input and the correlation table returns the spatial correlation to this structure for each ion image. Two examples of ion images that positively correlate with the target anatomical structure are displayed on the left. Examples of no correlation are shown in the middle, with examples of negative correlation on the right. (Panel B) Example of an ion query, finding anatomical regions in which m/z 7841 is highly expressed. The ion image of m/z 7841 is given as an input and the correlation table returns the spatial correlation to this ion image for all the anatomical regions. Two examples of anatomical structure images that positively correlate with the target ion image are displayed on the left. Examples of no correlation are shown in the middle, with examples of negative correlation on the right.

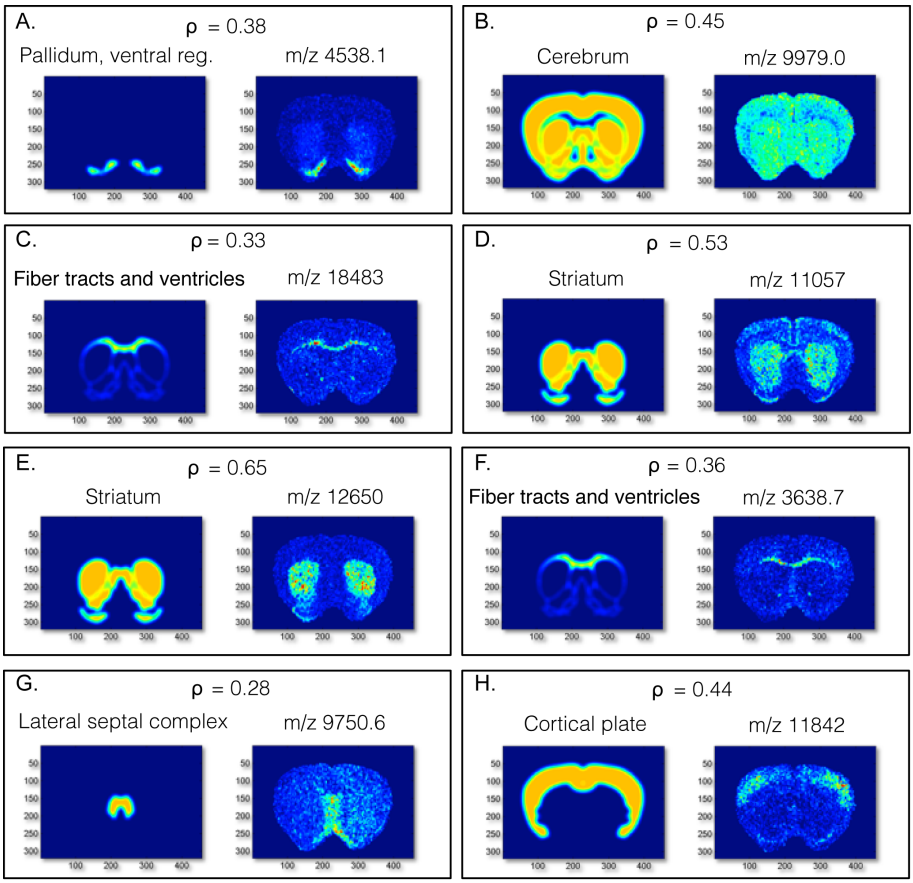

Figure 4.12: Examples of positively correlating anatomical structure/ion image combinations. In each image the anatomical structure image is shown on the left and the correlating ion image is shown on the right.

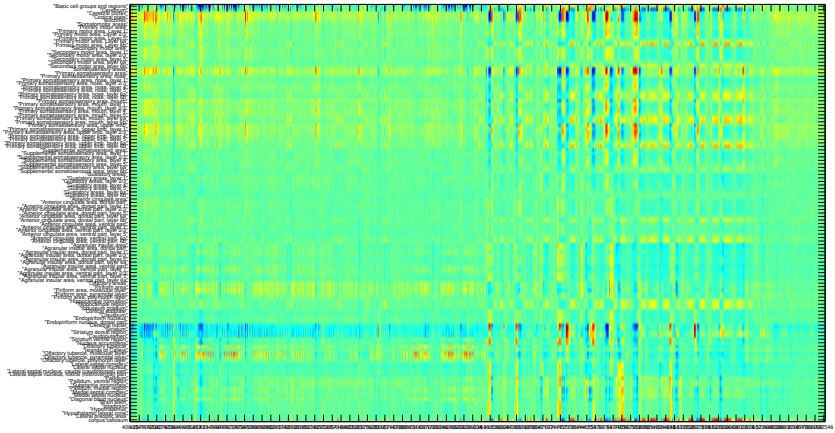

Figure 4.13: The correlation table resulting from the spatial correlation analysis between the anatomical structure images and the ion images of the peak-picked IMS lipid data. The y-axis shows all the anatomical structures that are present in the tissue slice. The x-axis shows the peak-picked m/z values. Red = high positive correlation, blue = high negative correlation, and green = low correlation.

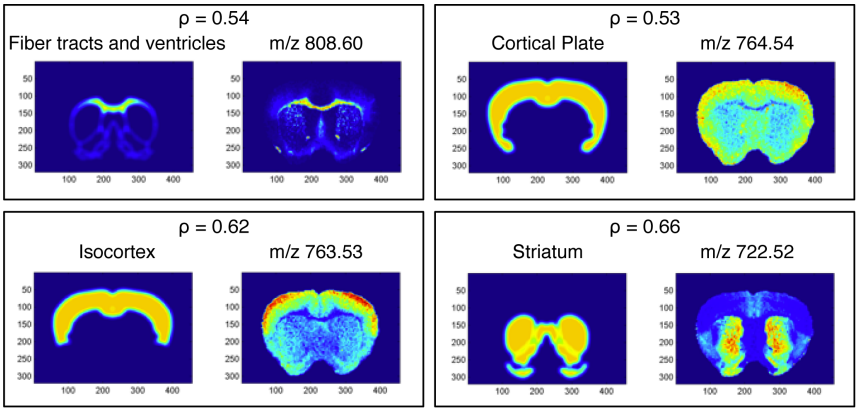

Figure 4.14: Examples of several positively correlating lipid ion images. In each image the anatomical structure image is shown on the left and the correlating ion image is shown on the right.

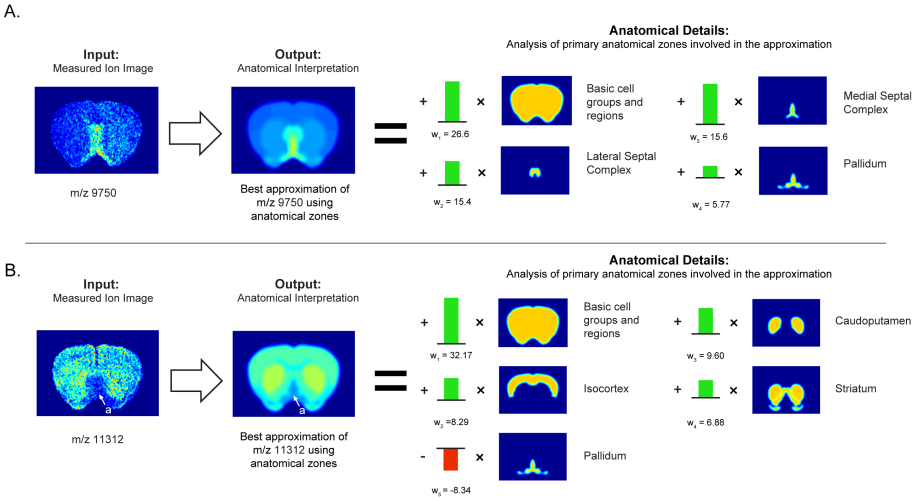

Figure 4.15: Examples of automated anatomical interpretation applied to the protein-focused IMS case study. When an ion image is given as input (left), the interpretation method provides an optimal anatomical explanation for the observed ion pattern (right), using the library of provided anatomical structures. Specifically, the ion intensity pattern is decomposed without user intervention into an optimal combination of contributing anatomical structures from the atlas. The interpretation method provides (i) the closest approximation of the measured ion image using atlas structures (middle), and (ii) a linear combination of the contributing anatomical structures, specifying name, reference location, and contributing intensity or weight in the interpretation (right). This visualization shows for each ion image the most contributing anatomical zones, preceded by their weight, which signifies their relative importance in the overall anatomical interpretation. Negative weights indicate a relative decrease of the ion in those areas. (panel A) Ion m/z 9750 is highly expressed in the medial septal complex and the lateral septal complex. (panel B) Ion m/z 11312 is specifically absent in the pallidum. This is indicated by its negative contribution coefficient or weight, and is also clearly visible in the ion image. The empirical ion distributions show good congruence with the boundaries of the anatomical structures defined in the atlas, indicating good spatial mapping between the data sources, and strong biological signals in IMS measurements also in the protein-focused case study.

particular 'best' value for γ, since the best value for one application is not necessarily the best for another. We therefore prefer to provide the user access to this dial, so that he/she may determine directly which sort of interpretation complexity is desired for a particular use case. However, note that the ability to request different levels of complexity in an anatomical interpretation does not jeopardize the analytical accuracy of the interpretation. This is because a less complex anatomical interpretation (higher γ) is directly related to a more complex interpretation (lower γ) in the sense that large anatomical zones are simply broken down into subdivisions that allow for a more granular description. As such, changing γ only changes the complexity of the list of anatomical labels involved, it does not change which anatomical space is tied to the ion pattern.

For the proof-of-concepts and introduction of the multivariate model and algorithm in this paper, we have set γ to one since it allows us to demonstrate the capabilities of the algorithm sufficiently to make our point. Due to the dimensions of our datasets (and in fact most MALDI IMS datasets) a γ of one ensures that the sparsity is secondary to the approximation. It basically ensures that only if two equally close anatomical interpretations are available, the sparser one wins.

Another note is that due to the speed and parallelism of the algorithm, it is perfectly feasible to explore multiple γs and determine the 'optimal' γ for a particular dataset and application. In fact, this is the common approach to setting such parameters in the optimization literature. An example of this is the use of 'optimal trade-off curves' in Boyd and Vandenberghe [123].

With regards to comparing anatomical interpretations between different ion images, it is also good to note that the γ parameter is equal across all the hundreds of ion images within an IMS experiment. This makes these interpretations directly comparable regardless of the particular value of γ, an advantage not available from human interpretation (where the level of interpretation complexity is a free-floating variable, subjectively controlled at best).

4.5.6 Automated Anatomical Interpretation and the Structure of Annotation Labels

With section 4.5.5 discussing the role of sparsity in the anatomical interpretation, it is valuable to also have a closer look at the structure of the AMBA and specifically how the structure of the annotations in the AMBA pertains to the interpretation. As mentioned in section 4.2.1, the AMBA is organized in a hierarchical fashion. This annotation structure sometimes does not accurately

reflect the region subdivisions encountered in nature, and since the anatomical interpretation method can only work with the anatomical annotations provided by the atlas, an optimal interpretation sometimes returns more anatomical zones than would be expected at first sight. Below, we illustrate this hierarchical subdivision of the atlas with several screenshots (Fig.4.16 to 4.18), taken from the interactive atlas viewer available on the AMBA website [117]. As an example, these figures show that the isocortex is hierarchically first sub-divided into "radial" sections, and only at a deeper level do these radial structures get split further into concentric sub-divisions. The hierarchical annotation structure prevents the AMBA from providing a single named annotation that represents all concentric regions of the isocortex as one, regardless of the radial locations involved. As such, the automated anatomical interpretation method does not have a concentric radial-independent sub-division pattern available in its vocabulary of anatomical structures, and can only represent such a structure as a combination of concentric sub-divisions of radial sub-divisions. Studies such as Ko *et al.* [124] on the other hand show that in biology a concentric radial-independent division exists in the mouse brain, and such a division can also be seen in ion image m/z 7841 of Figure 4.4B (the intensity of the ion is higher in the outer-most concentric layer of the brain).

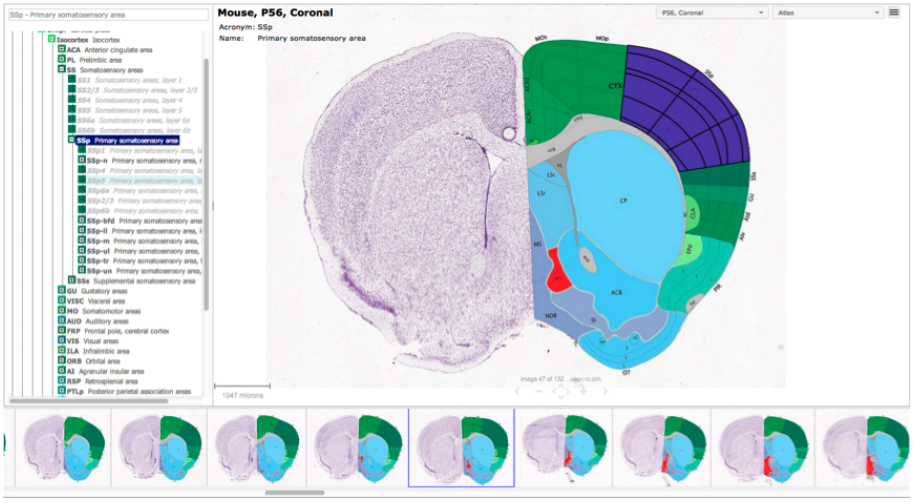

Figure 4.16: Primary somatosensory area

This example shows that ions specific to a concentric layer of the brain, but not specific to a radial sub-division of that brain, are difficult to approximate using the AMBA. The AMBA essentially forces a non-hierarchical anatomy structure into a hierarchical representation, and as a result loses some of the

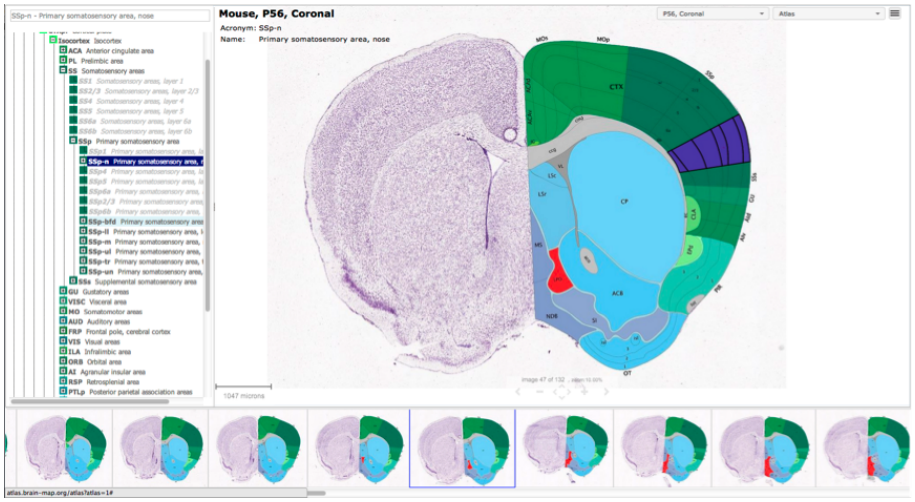

Figure 4.17: Primary somatosensory area, nose (a radial sub-area of the "Primary somatosensory area" parent shown in the previous figure)

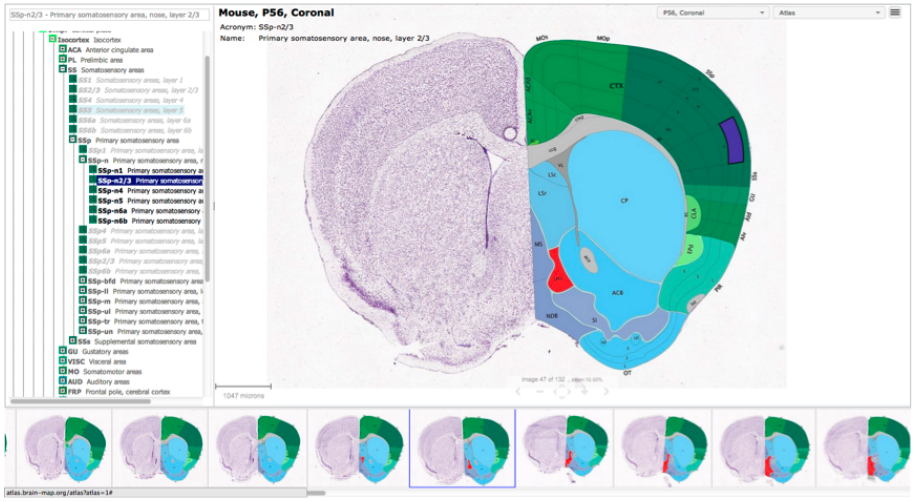

Figure 4.18: Primary somatosensory area, nose, layer 2/3 (a concentric sub-area within a radial sub-area of the "Primary somatosensory area")

parent-daughter region relationships. This is an example of a biological region that can be easily recognized, but is not explicitly represented in the atlas and therefor hard to label with a single name. This sort of biological region cannot be captured adequately using a univariate approach, and this example further illustrates the need for multivariate interpretation algorithms capable of handling multi-membership properly.

With a multi-membership aware interpretation method, such as the one developed in this paper, there are currently two ways of handling these types of regions:

- Complement the AMBA with additional patterns that group sub-patterns in different combinations (could be done exhaustively). For our proof-of-concept demonstrations we wanted to use the AMBA as-is, so this option is not explicitly discussed here.
- Use the regularization parameter γ of our method as a dial that can be set to "make the interpretation as specific as possible" (which would grab the cortical layers specifically) or "make the interpretation high-level" (which would grab the overall encompassing region), as discussed in section 4.5.5.
- It is valuable to note that due to the speed and parallelism of the interpretation method, it is perfectly feasible to collect interpretations using multiple regularization parameter values (from global all the way to very specific). It would just require storing more anatomical membership tables, but would be a means of 'auto-discovering' the sub-categorizations that have been lost due to the hierarchical structure of the AMBA.
- In short, although the hierarchical structure of the AMBA is a feature external to our method, our interpretation method provides the parameter to either ask for an interpretation of a particular kind (from global to more specific), or the speed and performance to interrogate the same IMS measurement from a bunch of different viewpoints and collect the data for assessment in a later phase.

4.5.7 Automated Anatomical Interpretation of Non-Conforming Ion Images

This section describes how the automated anatomical interpretation method handles ion images that do not conform to anatomy, or that do not exhibit uniform intensity distribution within known anatomical structures. The interpretation method we developed finds for each empirically acquired ion image in an IMS experiment, the best possible approximation using only the

vocabulary of reference anatomy patterns in the atlas. This means that for each ion image in the IMS experiment, there is a measure of 'distance' between the measured ion image and its approximation using anatomy from the atlas. This distance is expressed in terms of ion counts, and is implicitly provided by our algorithm for each individual ion species in the study. These distances can be directly compared across all ion images in an IMS experiment.

Two possible scenarios unfold:

- Small distance between ion image and anatomical approximation: The spatial distribution of this ion can be well approximated using the atlas ⇒ strong automated anatomical interpretation of this ion species is possible using the standard anatomy provided by the atlas.
- Large distance between ion image and anatomical approximation: The spatial distribution of this ion cannot be well approximated using the atlas ⇒ the anatomy atlas does not provide the tools to adequately describe the spatial distribution of this ion species.

In other words, the automated anatomical interpretation algorithm can be used as an automated filter that separates ion species that conform to known anatomy from those ion species that deviate from known anatomy.

This means that the algorithm we describe here does not only provide a means for automated anatomical interpretation of ion images, but implicitly also provides a means of discovering which ion species deviate substantially from known histology. As such, our method can be used to automatically detect which ion species are distributed into subdivisions, subareas, or gradients within anatomical regions that according to the atlas are supposed to be homogeneous. Essentially, our method gives an automated anatomical interpretation for those ions that can be explained using the atlas, and additionally highlights those ions in which the IMS measurements show patterns that cannot be explained by known anatomy (e.g. disease-specific chemical patterns, previously unknown chemical subdivisions within tissue areas, etc.).

Chapter 5

Conclusions & Future perspectives

5.1 Conclusions

Imaging mass spectrometry (IMS) is a powerful biomolecular imaging technology that can track hundreds to thousands of biomolecules simultaneously throughout a tissue slice in a single experiment. Due to ever-increasing technological advances, the data resulting from these experiments have become so large and complex that manual exploration has become infeasible, and datamining is becoming an increasingly vital component in extracting the rich information that is captured in these data, as evidenced by the steep increase in the number of groups focusing on the computational analysis of IMS data in recent years. In this thesis we have contributed to this community effort by focusing on three areas of IMS data analysis, namely data size and dimensionality reduction of IMS data, the comparison of IMS data originating from multiple tissue samples and the fusion of IMS data with additional data sources.

In our research on dimensionality reduction in Chapter 2 we investigated the use of the discrete wavelet transform DWT as an alternative to peak picking for dimensionality and size reduction of IMS data. Peak picking is a very harsh form of dimensionality reduction in that it discards a lot of information contained in the mass spectrum, such as peak shape. By using the DWT for dimensionality reduction we can keep all the useful information in the mass spectrum, while still achieving considerable reduction of the data size. We accomplished this by discarding only those variables that contain noise, and

which take up unnecessary storage space and hamper computational analysis. We differentiated between variables that contain noise and those that are the result of true biological signal by incorporating a key feature of IMS data, spatial information, into the selection procedure. We demonstrated two different ways of incorporating this spatial information, namely globally and locally, where both methods show improvement over traditional methods that do not incorporate this information, either through better retention of features that contain true signal or through better data compression by retaining only those features that contain true biological signal.

In Chapter 3 we introduced Group Independent Component Analysis as a new method for differential analysis of the heterogeneous IMS data originating from multiple tissue slices. This method allowed for an in-depth comparison of a large number of IMS datasets without the need for peak picking, extracting features from the full IMS data at the level of an individual tissue. By performing the pattern extraction at the level of a single tissue, GICA emphasizes small, tissue-specific features which could be lost in a simultaneous analysis over all tissues. Using an artificially generated dataset containing 15 artificial tissues divided over 3 different disease classes, we demonstrated that the algorithm allows for near-perfect retrieval of the differential components between tissues in the correct locations. We then continued by applying this technique to 2 real-life IMS datasets, a case study of ALS mouse spine and one of retinal degeneration disease in mouse retina. Application of the GICA algorithm to these datasets resulted in retrieval of multiple differential biomolecular patterns between the different tissue classes, which were expressed in biologically relevant areas for the diseases under study. These patterns could be retrieved in structures one cell-layer thick in the case of the retina samples. The retrieved biomolecular patterns were verfied by comparing the expression of the found peaks directly in the IMS data. Towards the future this technique could be used to extract features from larger datasets by dividing a single dataset up into various areas.

In our work on data fusion in Chapter 4 we explored the potential of integrating IMS data with information from additional data sources, thus enabling deeper insights into the IMS data. In this work we linked IMS data directly to an anatomical atlas, namely the Allen Mouse Brain Atlas (AMBA), which allowed for direct, computational access to an important resource that many human interpretations of IMS data will naturally rely upon, anatomical insights. We spatially mapped IMS data to the atlas using non-rigid image registration techniques and used the established link to investigate spatial correlations between ion images and the anatomical structures of the atlas, gaining basic insights into the relationships between the two. We finally moved beyond simple correlation and used the atlas information to provide automated anatomical interpretation of the ion images in the IMS data. This challenge was solved as a

convex optimization problem that deconstructs ion distributions as combinations of known anatomical structures. We demonstrated that establishing a link between an IMS experiment and an anatomical atlas can serve as an important accelerator both for human and machine-guided exploration of IMS experiments.

5.2 Future perspectives

The field of computational analysis for IMS has strongly grown over the last few years, and has evolved from the application of standard data analysis techniques to the development of datamining techniques specifically designed for analysis of IMS data. Nonetheless, there are still many exciting computational challenges to be addressed before IMS can achieve its full potential. Below we will focus on a few of these challenges which, when solved, would greatly benefit the field of IMS as a whole.

5.2.1 Data size and dimensionality of IMS data

A first challenge will be to handle the ever-increasing data sizes and dimensionality of IMS data. Datasets with data sizes in the terabyte range are now starting to appear, and with micron and sub-micron spatial resolutions around the corner and the growing usage of high-mass resolution mass spectrometers such as FT-ICR and Orbitrap for IMS experimentation, this trend of increasing data sizes is sure to continue for the years to come. A first way of dealing with these large datasets will be the further improvement of dimensionality reduction techniques such as peak picking and the more conservative DWT-based compression used in this thesis. Alternatively, fast and sensitive feature selection methods from other fields with high-dimensional data such as next-gen sequencing could be introduced. As we have shown, the incorporation of spatial information into these approaches can lead to an improvement in the number of relevant features retained. In this regard, spatial information in IMS data is still very much an under-utilized resource in the computational analysis of these data in general. Towards the future, it would therefore be interesting to see a deeper merging of already-established techniques from the computer vision field with techniques that can handle the high dimensionality of IMS data. This could for example allow for feats such as the automated extraction of anatomical structures based on morphology of the ion images alone.

A second way of handling these will be through the use of large scale and distributed computing, a field that is rapidly advancing with the coming of

'big data' over the last few years. Several groups are already introducing these techniques for use in IMS with the establishment of cloud-based data analysis services for IMS data such as openMSI [125] and the recent demonstration of distributed computing strategies for FT-ICR IMS data by Smith et al. [126]

5.2.2 Data integration

In our chapter on differential analsysis we saw how pattern extraction techniques can be used to extract differential biomolecular patterns between multiple tissues. The retrieved biomolecular patterns, while already presenting a focused view on the correlating ions in an area of interest, still contain tens to hundreds of biomolecular species. Making sense of these data is far from trivial and it is often unclear which peaks are the best targets for further investigation. In order to gain insight into these data, it would be of tremendous value to integrate these data with the vast body of already existing knowledge, available through literature and online data sources such as KEGG, UniProt, SwissProt, etc. Furthermore, a new IMS experiment is currently analyzed largely independent of previous IMS experiments, meaning that often only a fraction of the previously acquired data and knowledge is transferred to this new experiment. This not only happens within a single lab, but throughout the entire IMS field.

In an ideal world we would therefore be able to fully integrate information gained from other IMS experiments, together with knowledge gathered from external datasources directly into the analysis a new IMS experiment, allowing for automated annotation of previously identified biomolecules and taking into account information such as tissue type, clinical information, matrix used, etc. Enabling such integration will require an interplay of experimental and computational advances, and an overall community effort.

With respect to experimental advances one of the most important elements necessary for linking IMS data to external datasources is the identification of the retrieved biomolecular ions. Currently only a limited number of the observed ions can be easily identified and annotated, however high-mass resolution mass spectrometers are currently allowing for very accurate reporting of molecular masses for small molecules, and towards the future, approaches such as on-tissue trypsin digest may enable the high-throughput identification of proteins in IMS. Such on-tissue digestion of the proteins, which fragments the proteins into peptides, will however add a whole new layer of complexity to the computational analysis of these IMS data where the challenge will then be to match peptides to their original proteins.

Concerning the integration of data sources useful for such annotation, Van de Plas et al. [55] have shown the potential of combining IMS data of diabetes

models with hormone processing pathways extracted from the UniProt database. Here ions were mapped to these pathways based and thus without prior identification of the ions, using only the masses of the ions involved. By comparing the mappings of the retrieved a wild type mouse to different diabetes models, it could be determined where in the pathway there were irregularities.

Finally, there is a strong growth of the IMS community aimed at pushing this technology forward, with the European COST action encouraging many new collaborations throughout Europe. Given this strong community, there is hope for an open sharing of knowledge on IMS (and potentially data), which would greatly advance the field as a whole. Watrous et al. [127] have recently called for a community effort towards annotating IMS data and initiatives such as the MSiMass List database [128] allows for public annotations, have opened the door for such a community effort to be put in practice.

In our work on the Allen Mouse Brain Atlas we offered another way of integrating IMS data with existing resources. Towards the future, this work could be extended towards IMS imaging of a full mouse brain, which would allow for the construction of a full IMS-based protein and lipid atlas, which is presently non-existent. These data could then in turn be combined with the non-IMS data sources already linked to the AMBA, such as gene expression and neuron connectivity measurements. In the specific case of the Allen Mouse Brain Atlas this offers several promising perspectives for the future, since IMS-based data could then even be combined with non-IMS data sources. The combination of these data sources could serve as an immensely valuable multi-modal resource for systems biology research, bringing together insights from many different technologies and fields. In order to make full use of such an atlas, it would be necessary to annotate the biomolecules found in the IMS data, which ties with the challenges we described above.

A final observation is that tissue-focused IMS, which has long been confined to a research setting, is now making its way towards clinical application in the operating room [129] and towards clinical pathology [75] [130] [131]. Here it can make a direct impact on individual lives by providing clinicians with rich and objective chemical informal on a diseased tissue, which can improve disease diagnosis and consequently lead to a better patient treatment. With such promising new applications around the corner, and the ever-advancing IMS technology increasing the richness and complexity of the resulting data, there is a bright future ahead for IMS and its supporting datamining efforts.

Bibliography

[1] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Journal of Theoretical Biology*, vol. 22, pp. 437–467, Mar. 1969.

[2] I. Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, pp. 931–945, Oct. 2004.

[3] I. Ezkurdia, D. Juan, J. M. Rodriguez, A. Frankish, M. Diekhans, J. Harrow, J. Vazquez, A. Valencia, and M. L. Tress, "Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes.," *Human molecular genetics*, vol. 23, pp. 5866–5878, Nov. 2014.

[4] M. Wilhelm, J. Schlegl, H. Hahne, A. M. Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J.-H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, and B. Kuster, "Mass-spectrometry-based draft of the human proteome," *Nature*, vol. 509, pp. 582–587, May 2014.

[5] C. Walsh, *Posttranslational Modification of Proteins*. Expanding Nature's Inventory, Roberts Publishers, 2006.

[6] L. Martin, X. Latypova, and F. Terro, "Post-translational modifications of tau protein: implications for Alzheimer's disease.," *Neurochemistry international*, vol. 58, pp. 458–471, Mar. 2011.

[7] M. Anbalagan, B. Huderson, L. Murphy, and B. G. Rowan, "Post-translational modifications of nuclear receptors and human disease.," *Nuclear Receptor Signaling*, vol. 10, pp. 1–13, 2012.

[8] M. Karas and F. Hillenkamp, "Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons.," *Analytical Chemistry*, vol. 60, pp. 2299–2301, Oct. 1988.

[9] K. Tanaka, H. Waki, Y. Ido, and S. Akita, "Protein and polymer analyses up to m/z 100 000 by laser ionization time of flight mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 2, no. 8, pp. 151–153, 1988.

[10] J. Fenn, M. Mann, C. Meng, S. Wong, and C. Whitehouse, "Electrospray ionization for mass spectrometry of large biomolecules," *Science (New York, N.Y.)*, vol. 246, pp. 64–71, Oct. 1989.

[11] D.-E. van der Merwe, K. Oikonomopoulou, J. Marshall, and E. P. Diamandis, "Mass spectrometry: uncovering the cancer proteome for diagnostics.," *Advances in cancer research*, vol. 96, pp. 23–50, 2007.

[12] T. K. Sinha, S. Khatib-Shahidi, T. E. Yankeelov, K. Mapara, M. Ehtesham, D. S. Cornett, B. M. Dawant, R. M. Caprioli, and J. C. Gore, "Integrating spatially resolved three-dimensional MALDI IMS with in vivo magnetic resonance imaging.," *Nature methods*, vol. 5, pp. 57–59, Jan. 2008.

[13] R. M. Caprioli, T. B. Farmer, and J. Gile, "Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS," *Analytical Chemistry*, vol. 69, pp. 4751–4760, Dec. 1997.

[14] L. A. Mcdonnell and R. M. A. Heeren, "Imaging mass spectrometry.," *Mass spectrometry reviews*, vol. 26, pp. 606–643, July 2007.

[15] K. Chughtai and R. M. A. Heeren, "Mass Spectrometric Imaging for Biomedical Tissue Analysis," *Chemical Reviews*, vol. 110, pp. 3237–3277, May 2010.

[16] A. Walch, S. Rauser, S.-O. Deininger, and H. Höfler, "MALDI imaging mass spectrometry for direct tissue analysis: a new frontier for molecular histology.," *Histochemistry and cell biology*, vol. 130, pp. 421–434, Sept. 2008.

[17] N. Takai, Y. Tanaka, K. Inazawa, and H. Saji, "Quantitative analysis of pharmaceutical drug distribution in multiple organs by imaging mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 26, pp. 1549–1556, May 2012.

[18] P. J. Trim, C. M. Henson, J. L. Avery, A. McEwen, M. F. Snel, E. Claude, P. S. Marshall, A. West, A. P. Princivalle, and M. R. Clench, "Matrix-Assisted Laser Desorption/Ionization-Ion Mobility Separation-Mass Spectrometry Imaging of Vinblastine in Whole Body Tissue Sections," *Analytical Chemistry*, vol. 80, pp. 8628–8634, Nov. 2008.

[19] Y. Sugiura, Y. Konishi, N. Zaima, S. Kajihara, H. Nakanishi, R. Taguchi, and M. Setou, "Visualization of the cell-selective distribution of PUFA-containing phosphatidylcholines in mouse brain by imaging mass spectrometry," *The Journal of Lipid Research*, vol. 50, pp. 1776–1788, Aug. 2009.

[20] M. Stoeckli, P. Chaurand, D. E. Hallahan, and R. M. Caprioli, "Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues.," *Nature medicine*, vol. 7, pp. 493–496, Apr. 2001.

[21] S. Khatib-Shahidi, M. Andersson, J. L. Herman, T. A. Gillespie, and R. M. Caprioli, "Direct Molecular Analysis of Whole-Body Animal Tissue Sections by Imaging MALDI Mass Spectrometry," *Analytical Chemistry*, vol. 78, pp. 6448–6456, Sept. 2006.

[22] L. S. Eberlin, I. Norton, A. L. Dill, A. J. Golby, K. L. Ligon, S. Santagata, R. G. Cooks, and N. Y. R. Agar, "Classifying Human Brain Tumors by Lipid Imaging with Mass Spectrometry," *Cancer research*, vol. 72, pp. 645–654, Jan. 2012.

[23] L. H. Cazares, D. Troyer, S. Mendrinos, R. A. Lance, J. O. Nyalwidhe, H. A. Beydoun, M. A. Clements, R. R. Drake, and O. J. Semmes, "Imaging Mass Spectrometry of a Specific Fragment of Mitogen-Activated Protein Kinase/Extracellular Signal-Regulated Kinase Kinase Kinase 2 Discriminates Cancer from Uninvolved Prostate Tissue," *Clinical Cancer Research*, vol. 15, pp. 5541–5551, Aug. 2009.

[24] P. Chaurand, S. A. Schwartz, M. L. Reyzer, and R. M. Caprioli, "Imaging mass spectrometry: principles and potentials.," *Toxicologic pathology*, vol. 33, no. 1, pp. 92–101, 2005.

[25] Y. juunhai and R. M. Caprioli, "Matrix Sublimation/Recrystallization for Imaging Proteins by Mass Spectrometry at High Spatial Resolution," *Analytical Chemistry*, vol. 83, pp. 5728–5734, July 2011.

[26] M. Koestler, D. Kirsch, A. Hester, A. Leisner, S. Guenther, and B. Spengler, "A high-resolution scanning microprobe matrix-assisted laser desorption/ionization ion source for imaging analysis on an ion trap/Fourier transform ion cyclotron resonance mass spectrometer," *Rapid Communications in Mass Spectrometry*, vol. 22, pp. 3275–3285, Oct. 2008.

[27] M. L. Pacholski and N. Winograd, "Imaging with Mass Spectrometry," *Chemical Reviews*, vol. 99, pp. 2977–3006, Oct. 1999.

[28] R. M. A. Heeren, D. F. Smith, J. Stauber, B. Kükrer-Kaletas, and L. Macaleese, "Imaging Mass Spectrometry: Hype or Hope?," *JAM*, vol. 20, pp. 1006–1014, Nov. 2009.

[29] E. J. Lanni, S. J. B. Dunham, P. Nemes, S. S. Rubakhin, and J. V. Sweedler, "Biomolecular Imaging with a C60-SIMS/MALDI Dual Ion Source Hybrid Mass Spectrometer: Instrumentation, Matrix Enhancement, and Single Cell Analysis," *Journal of the American Society for Mass Spectrometry*, vol. 25, pp. 1897–1907, Sept. 2014.

[30] A. Carado, M. K. Passarelli, J. Kozole, J. E. Wingate, N. Winograd, and A. V. Loboda, "C 60Secondary Ion Mass Spectrometry with a Hybrid-Quadrupole Orthogonal Time-of-Flight Mass Spectrometer," *Analytical Chemistry*, vol. 80, pp. 7921–7929, Nov. 2008.

[31] Z. Takats, J. M. Wiseman, B. Gologan, and R. G. Cooks, "Mass spectrometry sampling under ambient conditions with desorption electrospray ionization.," *Science (New York, N.Y.)*, vol. 306, pp. 471–473, Oct. 2004.

[32] J. M. Wiseman, D. R. Ifa, A. Venter, and R. G. Cooks, "Ambient molecular imaging by desorption electrospray ionization mass spectrometry.," *Nature protocols*, vol. 3, no. 3, pp. 517–524, 2008.

[33] J. M. Wiseman, D. R. Ifa, Y. Zhu, C. B. Kissinger, N. E. Manicke, P. T. Kissinger, and R. G. Cooks, "Desorption electrospray ionization mass spectrometry: Imaging drugs and metabolites in tissues.," *Proceedings of the National Academy of Sciences*, vol. 105, pp. 18120–18125, Nov. 2008.

[34] W. Rao, A. D. Celiz, D. J. Scurr, M. R. Alexander, and D. A. Barrett, "Ambient DESI and LESA-MS Analysis of Proteins Adsorbed to a Biomaterial Surface Using In-Situ Surface Tryptic Digestion," *Journal of the American Society for Mass Spectrometry*, vol. 24, pp. 1927–1936, Sept. 2013.

[35] J. Laskin, B. S. Heath, P. J. Roach, L. Cazares, and O. J. Semmes, "Tissue Imaging Using Nanospray Desorption Electrospray Ionization Mass Spectrometry," *Analytical Chemistry*, vol. 84, pp. 141–148, Jan. 2012.

[36] J. S. Becker, M. V. Zoriy, C. Pickhardt, N. Palomero-Gallagher, and K. Zilles, "Imaging of copper, zinc, and other elements in thin section of human brain samples (hippocampus) by laser ablation inductively coupled plasma mass spectrometry," *Analytical Chemistry*, vol. 77, pp. 3208–3216, May 2005.

[37] P. Nemes and A. Vertes, "Atmospheric-pressure Molecular Imaging of Biological Tissues and Biofilms by LAESI Mass Spectrometry," *Journal of Visualized Experiments*, no. 43, 2010.

[38] P. Chaurand, "Imaging mass spectrometry of thin tissue sections: A decade of collective efforts," *Journal of proteomics*, vol. 75, pp. 4883–4892, Aug. 2012.

[39] D. M. Drexler, T. J. Garrett, J. L. Cantone, R. W. Diters, J. G. Mitroka, M. C. Prieto Conaway, S. P. Adams, R. A. Yost, and M. Sanders, "Utility of imaging mass spectrometry (IMS) by matrix-assisted laser desorption ionization (MALDI) on an ion trap mass spectrometer in the analysis of drugs and metabolites in biological tissues.," *Journal of Pharmacological and Toxicological Methods*, vol. 55, pp. 279–288, May 2007.

[40] P. D. Verhaert, M. C. P. Conaway, T. M. Pekar, and K. Miller, "Neuropeptide imaging on an LTQ with vMALDI source: The complete 'all-in-one'peptidome analysis," *International Journal of Mass Spectrometry*, vol. 260, no. 2, pp. 177–184, 2007.

[41] S. Shimma, M. Furuta, K. Ichimura, Y. Yoshida, and M. Setou, "Direct MS/MS analysis in mammalian tissue sections using MALDI-QIT-TOFMS and chemical inkjet technology," *Surface and Interface Analysis*, vol. 38, no. 12-13, pp. 1712–1714, 2006.

[42] B. Prideaux, V. Dartois, D. Staab, D. M. Weiner, A. Goh, L. E. Via, C. E. Barry, III, and M. Stoeckli, "High-Sensitivity MALDI-MRM-MS Imaging of Moxifloxacin Distribution in Tuberculosis-Infected Rabbit Lungs and Granulomatous Lesions," *Analytical Chemistry*, vol. 83, pp. 2112–2118, Mar. 2011.

[43] I. M. Taban, A. F. M. Altelaar, Y. E. M. van der Burgt, L. A. Mcdonnell, R. M. A. Heeren, J. Fuchser, and G. Baykut, "Imaging of peptides in the rat brain using MALDI-FTICR mass spectrometry.," *JAM*, vol. 18, pp. 145–151, Jan. 2007.

[44] S. K. Shahidi-Latham, S. M. Dutta, M. C. Prieto Conaway, and P. J. Rudewicz, "Evaluation of an Accurate Mass Approach for the Simultaneous Detection of Drug and Metabolite Distributions via Whole-Body Mass Spectrometric Imaging," *Analytical Chemistry*, vol. 84, pp. 7158–7165, Aug. 2012.

[45] S. R. Ellis, J. H. Jungmann, D. F. Smith, J. Soltwisch, and R. M. A. Heeren, "Enhanced Detection of High-Mass Proteins by Using an Active Pixel Detector," *Angewandte Chemie International Edition*, vol. 52, pp. 11261–11264, Sept. 2013.

[46] J. H. Jungmann, D. F. Smith, L. MacAleese, I. Klinkert, J. Visser, and R. M. A. Heeren, "Biological Tissue Imaging with a Position and Time

Sensitive Pixelated Detector," *Journal of the American Society for Mass Spectrometry*, vol. 23, pp. 1679–1688, July 2012.

[47] S. Chughtai, K. Chughtai, B. Cillero-Pastor, A. Kiss, P. Agrawal, L. MacAleese, and R. M. A. Heeren, "A multimodal mass spectrometry imaging approach for the study of musculoskeletal tissues," *International Journal of Mass Spectrometry*, vol. 325-327, pp. 150–160, July 2012.

[48] M. K. Passarelli and N. Winograd, "Characterizing in situ Glycerophospholipids with SIMS and MALDI Methodologies.," *Surface and interface analysis : SIA*, vol. 43, pp. 269–271, Jan. 2011.

[49] E. A. Jones, S.-O. Deininger, P. C. W. Hogendoorn, A. M. Deelder, and L. A. Mcdonnell, "Imaging mass spectrometry statistical analysis.," *Journal of proteomics*, vol. 75, pp. 4962–4989, Aug. 2012.

[50] T. Alexandrov, "MALDI imaging mass spectrometry: statistical data analysis and current computational challenges.," *BMC Bioinformatics*, vol. 13 Suppl 16, p. S11, 2012.

[51] J. L. Jianqing Fan, "A Selective Overview of Variable Selection in High Dimensional Feature Space," *Statistica Sinica*, vol. 20, p. 101, Jan. 2010.

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

[53] L. A. McDonnell, A. van Remoortere, N. de Velde, R. J. M. van Zeijl, and A. M. Deelder, "Imaging Mass Spectrometry Data Reduction: Automated Feature Identification and Extraction," *JAM*, vol. 21, pp. 1969–1978, Dec. 2010.

[54] R. Van de Plas, B. De Moor, and E. Waelkens, "Discrete wavelet transform-based multivariate exploration of tissue via imaging mass spectrometry," in *the 2008 ACM symposium*, (New York, New York, USA), p. 1307, ACM Press, 2008.

[55] R. Van de Plas, *Tissue Based Proteomics and Biomarker Discovery - Multivariate Data Mining Strategies for Mass Spectral Imaging, PhD. Dissertation.* PhD thesis, KU Leuven (Leuven, Belgium), Leuven, Jan. 2010.

[56] C. Bauer, R. Cramer, and J. Schuchhardt, "Evaluation of peak-picking algorithms for protein mass spectrometry.," *Methods in molecular biology (Clifton, N.J.)*, vol. 696, pp. 341–352, 2011.

[57] C. Yang, Z. He, and W. Yu, "Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis," *BMC Bioinformatics*, vol. 10, no. 1, p. 4, 2009.

[58] I. Daubechies, *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992.

[59] S. Mallat, *A wavelet tour of signal processing*. Academic press, 1999.

[60] R. Murenzi, *Wavelets*. Berlin: Springer-Verlag, 1988.

[61] S. Sanei and J. A. Chambers, *EEG signal processing*. Wiley, 2007.

[62] P. Du, W. A. Kibbe, and S. M. Lin, "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching.," *Bioinformatics*, vol. 22, pp. 2059–2065, Sept. 2006.

[63] G. Strang and T. Nguyen, *Wavelets and filter banks*. Wellesley-Cambridge Press, 1996.

[64] A. P. Bradley, "Shift-invariance in the discrete wavelet transform," in *Proceedings of VIIth Digital Image Computing: ...*, (Sydney), 2003.

[65] R. Van de Plas, F. Ojeda, M. Dewil, L. Van Den Bosch, B. De Moor, and E. Waelkens, "Prospective Exploration of Biochemical Tissue Composition via Imaging Mass Spectrometry Guided by Principal Component Analysis," in *Proceedings of the Pacific Symposium on Biocomputing (PSB)*, (Maui, Hawaii), pp. 458–469, Jan. 2007.

[66] G. McCombie, D. Staab, M. Stoeckli, and R. Knochenmuss, "Spatial and Spectral Correlations in MALDI Mass Spectrometry Images by Clustering and Multivariate Analysis," *Analytical Chemistry*, vol. 77, pp. 6118–6124, Oct. 2005.

[67] I. T. Jolliffe, *Principal Component Analysis*. Springer, Oct. 2002.

[68] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, "A Real-Time Algorithm for Signal Analysis with the Help of the Wavelet Transform," in *Wavelets*, pp. 286–297, Berlin, Heidelberg: Springer Berlin Heidelberg, 1990.

[69] D. L. DONOHO and J. M. JOHNSTONE, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[70] D. L. Donoho and I. M. Johnstone, "Minimax estimation via wavelet shrinkage," *The Annals of Statistics*, vol. 26, no. 3, pp. 879–921, 1998.

[71] T. Alexandrov and J. H. Kobarg, "Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering.," *Bioinformatics*, vol. 27, pp. i230–8, July 2011.

[72] M. Hanselmann, J. Röder, U. Köthe, B. Y. Renard, R. M. A. Heeren, and F. A. Hamprecht, "Active learning for convenient annotation and classification of secondary ion mass spectrometry images.," *Analytical Chemistry*, vol. 85, pp. 147–155, Jan. 2013.

[73] F. Vogt, S. Banerji, and K. Booksh, "Utilizing three-dimensional wavelet transforms for accelerated evaluation of hyperspectral image cubes," *Journal of Chemometrics*, vol. 18, no. 7-8, pp. 350–362, 2005.

[74] M.-C. Djidja, E. Claude, M. F. Snel, S. Francese, P. Scriven, V. Carolan, and M. R. Clench, "Novel molecular tumour classification using MALDI–mass spectrometry imaging of tissue micro-array," *Analytical and Bioanalytical Chemistry*, vol. 397, pp. 587–601, Mar. 2010.

[75] K. A. Veselkov, R. Mirnezami, N. Strittmatter, R. D. Goldin, J. Kinross, A. V. M. Speller, T. Abramov, E. A. Jones, A. Darzi, E. Holmes, J. K. Nicholson, and Z. Takats, "Chemo-informatic strategy for imaging mass spectrometry-based hyperspectral profiling of lipid signatures in colorectal cancer.," *Proceedings of the National Academy of Sciences*, vol. 111, pp. 1216–1221, Jan. 2014.

[76] E. A. Jones, A. van Remoortere, R. J. M. van Zeijl, P. C. W. Hogendoorn, J. V. M. G. Bovée, A. M. Deelder, and L. A. Mcdonnell, "Multiple statistical analysis techniques corroborate intratumor heterogeneity in imaging mass spectrometry datasets of myxofibrosarcoma.," *PloS one*, vol. 6, no. 9, p. e24913, 2011.

[77] E. A. Allen, E. B. Erhardt, Y. Wei, T. Eichele, and V. D. Calhoun, "Capturing inter-subject variability with group independent component analysis of fMRI data: A simulation study," *NeuroImage*, vol. 59, pp. 4141–4159, Feb. 2012.

[78] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications.," *Neural networks : the official journal of the International Neural Network Society*, vol. 13, pp. 411–430, May 2000.

[79] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular value decomposition and principal component analysis," in *A practical approach to microarray data analysis*, pp. 1–18, Springer, 2003.

[80] L. A. Klerk, P. Y. W. Dankers, E. R. Popa, A. W. Bosman, M. E. Sanders, K. A. Reedquist, and R. M. A. Heeren, "TOF-Secondary Ion Mass

Spectrometry Imaging of Polymeric Scaffolds with Surrounding Tissue after in Vivo Implantation," *Analytical Chemistry*, vol. 82, pp. 4337–4343, June 2010.

[81] P. R. Peres-Neto, D. A. Jackson, and K. M. Somers, "How many principal components? stopping rules for determining the number of non-trivial axes revisited," *Computational Statistics & Data Analysis*, vol. 49, pp. 974–997, June 2005.

[82] M. Hanselmann, M. Kirchner, B. Y. Renard, E. R. Amstalden, K. Glunde, R. M. A. Heeren, and F. A. Hamprecht, "Concise Representation of Mass Spectrometry Images by Probabilistic Latent Semantic Analysis," *Analytical Chemistry*, vol. 80, pp. 9649–9658, Dec. 2008.

[83] A. M. Race, R. T. Steven, A. D. Palmer, I. B. Styles, and J. Bunch, "Memory efficient principal component analysis for the dimensionality reduction of large mass spectrometry imaging data sets.," *Analytical Chemistry*, vol. 85, pp. 3071–3078, Mar. 2013.

[84] J. Ojanen, T. Miettinen, J. Heikkonen, and J. Rissanen, "Robust denoising of electrophoresis and mass spectrometry signals with minimum description length principle," *FEBS Letters*, vol. 570, pp. 107–113, July 2004.

[85] V. D. Calhoun, J. Liu, and T. Adali, "A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data," *NeuroImage*, vol. 45, pp. S163–S172, Mar. 2009.

[86] A. A. Ramos, "The minimum description length principle and model selection in spectropolarimetry," *The Astrophysical Journal*, vol. 646, no. 2, p. 1445, 2006.

[87] Y.-O. Li, T. Adali, and V. D. Calhoun, "Estimating the number of independent components for functional magnetic resonance imaging data.," *Human Brain Mapping*, vol. 28, pp. 1251–1266, Nov. 2007.

[88] T. P. Minka, "Automatic choice of dimensionality for PCA," *NIPS*, vol. 13, pp. 598–604, 2000.

[89] M. S. Lewicki, "A review of methods for spike sorting: the detection and classification of neural action potentials.," *Network (Bristol, England)*, vol. 9, pp. R53–78, Nov. 1998.

[90] M. S. Bartlett, *Face Image Analysis by Unsupervised Learning*. Springer Science & Business Media, 2001.

[91] A. J. Bell and T. J. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, Nov. 1995.

[92] J. Himberg and A. Hyvarinen, "Icasso: software for investigating the reliability of ICA estimates by clustering and visualization," in *IEEE XIII Workshop on Neural Networks for Signal Processing*, pp. 259–268, IEEE, 2003.

[93] E. B. Erhardt, S. Rachakonda, E. J. Bedrick, E. A. Allen, T. Adali, and V. D. Calhoun, "Comparison of multi-subject ICA methods for analysis of fMRI data," *Human Brain Mapping*, vol. 32, pp. 2075–2095, Dec. 2010.

[94] R. A. Fisher, "Statistical Methods for Research Workers. 5th ed.," *Oliver and Boyd*, 1934.

[95] B. Cassin and S. Solomon, *Dictionary of eye terminology*. Triad Publishing Company, 1997.

[96] R. Van de Plas, J. Yang, J. Spraggins, and R. M. Caprioli, "Fusion of mass spectrometry and microscopy: a multi-modality paradigm for molecular tissue mapping," *Nature methods*.

[97] R. Masyuko, E. J. Lanni, J. V. Sweedler, and P. W. Bohn, "Correlated imaging–a grand challenge in chemical analysis.," *The Analyst*, vol. 138, pp. 1924–1939, Apr. 2013.

[98] M. Andersson, M. R. Groseclose, A. Y. Deutch, and R. M. Caprioli, "Imaging mass spectrometry of proteins and peptides: 3D volume reconstruction," *Nature methods*, vol. 5, pp. 101–108, Jan. 2008.

[99] H. F. Wehrl, J. Schwab, K. Hasenbach, G. Reischl, G. Tabatabai, L. Quintanilla-Martinez, F. Jiru, K. Chughtai, A. Kiss, F. Cay, D. Bukala, R. M. A. Heeren, B. J. Pichler, and A. W. Sauter, "Multimodal Elucidation of Choline Metabolism in a Murine Glioma Model Using Magnetic Resonance Spectroscopy and 11C-Choline Positron Emission Tomography.," *Cancer research*, vol. 73, pp. 1470–1480, Mar. 2013.

[100] J. Bruand, T. Alexandrov, S. Sistla, M. Wisztorski, C. Meriaux, M. Becker, M. Salzet, I. Fournier, E. Macagno, and V. Bafna, "AMASS: Algorithm for MSI Analysis by Semi-supervised Segmentation," *Journal of Proteome Research*, vol. 10, pp. 4734–4743, Oct. 2011.

[101] J. M. Fonville, C. L. Carter, L. Pizarro, R. T. Steven, A. D. Palmer, R. L. Griffiths, P. F. Lalor, J. C. Lindon, J. K. Nicholson, E. Holmes, and J. Bunch, "Hyperspectral Visualization of Mass Spectrometry Imaging Data," *Analytical Chemistry*, vol. 85, pp. 1415–1423, Feb. 2013.

[102] K. J. Boggio, E. Obasuyi, K. Sugino, S. B. Nelson, N. Y. Agar, and J. N. Agar, "Recent advances in single-cell MALDI mass spectrometry imaging and potential clinical impact," *Expert Review of Proteomics*, vol. 8, pp. 591–604, Oct. 2011.

[103] S. N. Whitehead, K. H. N. Chan, S. Gangaraju, J. Slinn, J. Li, and S. T. Hou, "Imaging Mass Spectrometry Detection of Gangliosides Species in the Mouse Brain following Transient Focal Cerebral Ischemia and Long-Term Recovery," *PloS one*, vol. 6, p. e20808, June 2011.

[104] W. M. Abdelmoula, R. J. Carreira, R. Shyti, B. Balluff, R. J. M. van Zeijl, E. A. Tolner, B. F. P. Lelieveldt, A. M. J. M. van den Maagdenberg, L. A. Mcdonnell, and J. Dijkstra, "Automatic registration of mass spectrometry imaging data sets to the Allen brain atlas.," *Analytical Chemistry*, vol. 86, pp. 3947–3954, Apr. 2014.

[105] W. M. Abdelmoula, K. Škrášková, B. Balluff, R. J. Carreira, E. A. Tolner, B. P. F. Lelieveldt, L. van der Maaten, H. Morreau, A. M. J. M. van den Maagdenberg, R. M. A. Heeren, L. A. Mcdonnell, and J. Dijkstra, "Automatic Generic Registration of Mass Spectrometry Imaging Data to Histology Using Nonlinear Stochastic Embedding," *Analytical Chemistry*, vol. 86, pp. 9204–9211, Sept. 2014.

[106] K. Sköld, M. Svensson, A. Nilsson, X. Zhang, K. Nydahl, R. M. Caprioli, P. Svenningsson, and P. E. Andrén, "Decreased Striatal Levels of PEP-19 Following MPTP Lesion in the Mouse," *Journal of Proteome Research*, vol. 5, pp. 262–269, Feb. 2006.

[107] E. S. Lein, M. J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A. F. Boe, M. S. Boguski, K. S. Brockway, E. J. Byrnes, L. Chen, L. Chen, T.-M. Chen, M. C. Chin, J. Chong, B. E. Crook, A. Czaplinska, C. N. Dang, S. Datta, N. R. Dee, A. L. Desaki, T. Desta, E. Diep, T. A. Dolbeare, M. J. Donelan, H.-W. Dong, J. G. Dougherty, B. J. Duncan, A. J. Ebbert, G. Eichele, L. K. Estin, C. Faber, B. A. Facer, R. Fields, S. R. Fischer, T. P. Fliss, C. Frensley, S. N. Gates, K. J. Glattfelder, K. R. Halverson, M. R. Hart, J. G. Hohmann, M. P. Howell, D. P. Jeung, R. A. Johnson, P. T. Karr, R. Kawal, J. M. Kidney, R. H. Knapik, C. L. Kuan, J. H. Lake, A. R. Laramee, K. D. Larsen, C. Lau, T. A. Lemon, A. J. Liang, Y. Liu, L. T. Luong, J. Michaels, J. J. Morgan, R. J. Morgan, M. T. Mortrud, N. F. Mosqueda, L. L. Ng, R. Ng, G. J. Orta, C. C. Overly, T. H. Pak, S. E. Parry, S. D. Pathak, O. C. Pearson, R. B. Puchalski, Z. L. Riley, H. R. Rockett, S. A. Rowland, J. J. Royall, M. J. Ruiz, N. R. Sarno, K. Schaffnit, N. V. Shapovalova, T. Sivisay, C. R. Slaughterbeck, S. C. Smith, K. A. Smith, B. I. Smith, A. J. Sodt, N. N. Stewart, K.-R.

Stumpf, S. M. Sunkin, M. Sutram, A. Tam, C. D. Teemer, C. Thaller, C. L. Thompson, L. R. Varnam, A. Visel, R. M. Whitlock, P. E. Wohnoutka, C. K. Wolkey, V. Y. Wong, M. Wood, M. B. Yaylaoglu, R. C. Young, B. L. Youngstrom, X. F. Yuan, B. Zhang, T. A. Zwingman, and A. R. Jones, "Genome-wide atlas of gene expression in the adult mouse brain.," *Nature*, vol. 445, pp. 168–176, Jan. 2007.

[108] A. R. Jones, C. C. Overly, and S. M. Sunkin, "The Allen Brain Atlas: 5 years and beyond," *Nature Reviews Neuroscience*, vol. 10, pp. 821–828, Oct. 2009.

[109] J. Hanrieder, N. T. N. Phan, M. E. Kurczy, and A. G. Ewing, "Imaging mass spectrometry in neuroscience.," *ACS chemical neuroscience*, vol. 4, pp. 666–679, May 2013.

[110] L. G. Brown, "A survey of image registration techniques," *ACM Computing Surveys*, vol. 24, pp. 325–376, Dec. 1992.

[111] W. R. Crum, "Non-rigid image registration: theory and practice," *British Journal of Radiology*, vol. 77, pp. S140–S153, Dec. 2004.

[112] A. Myronenko, *Non-rigid Image Registration, PhD. Dissertation.* PhD thesis, Oregon Health & Science University, Department of Science & Engineering School of Medicine, 2010.

[113] J. Bruand, S. Sistla, C. Meriaux, P. C. Dorrestein, T. Gaasterland, M. Ghassemian, M. Wisztorski, I. Fournier, M. Salzet, E. Macagno, and V. Bafna, "Automated Querying and Identification of Novel Peptides using MALDI Mass Spectrometric Imaging," *Journal of Proteome Research*, vol. 10, pp. 1915–1928, Apr. 2011.

[114] F. Suits, T. E. Fehniger, Á. Végvári, G. Marko-Varga, and P. Horvatovich, "Correlation Queries for Mass Spectrometry Imaging," *Analytical Chemistry*, vol. 85, pp. 4398–4404, May 2013.

[115] R. Van de Plas, K. Pelckmans, B. De Moor, and E. Waelkens, "Spatial Querying of Imaging Mass Spectrometry Data: A Nonnegative Least Squares Approach ," in *The NIPS workshop on Machine Learning in Computational Biology (NIPS MLCB)*, pp. 1–4, Aug. 2007.

[116] L. A. Mcdonnell, A. van Remoortere, R. J. M. van Zeijl, and A. M. Deelder, "Mass Spectrometry Image Correlation: Quantifying Colocalization," *Journal of Proteome Research*, vol. 7, pp. 3619–3627, Aug. 2008.

[117] Allen Institute for Brain Science, "Allen Brain Atlas Website," tech. rep., 2013.

[118] R. D. Eastman, J. Le Moigne, and N. S. Netanyahu, "Research issues in image registration for remote sensing," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, May 2007.

[119] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: application to breast MR images," *IEEE Transactions on Medical Imaging*, vol. 18, no. 8, pp. 712–721, 1999.

[120] R. Tibshirani, "JSTOR: All Volumes and Issues - Browse - Journal of the Royal Statistical Society. Series B (Methodological)," *Journal of the Royal Statistical Society Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[121] M. Grant and S. Boyd, "CVX: Matlab Software for Disciplined Convex Programming, version 2.0 beta," tech. rep., 2012.

[122] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control* (V. Blondel, S. Boyd, and H. Kimura, eds.), pp. 95–110, Springer-Verlag Limited, 2008.

[123] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, UK: Cambridge University Press, Aug. 2004.

[124] Y. Ko, S. A. Ament, J. A. Eddy, J. Caballero, J. C. Earls, L. Hood, and N. D. Price, "Cell type-specific genes show striking and distinct patterns of spatial expression in the mouse brain.," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 3095–3100, Feb. 2013.

[125] O. Rübel, A. Greiner, S. Cholia, K. Louie, E. W. Bethel, T. R. Northen, and B. P. Bowen, "OpenMSI: A High-Performance Web-Based Platform for Mass Spectrometry Imaging," *Analytical Chemistry*, vol. 85, pp. 10354–10361, Nov. 2013.

[126] D. F. Smith, C. Schulz, M. Konijnenburg, M. Kilic, and R. M. A. Heeren, "Distributed computing strategies for processing of FT-ICR MS imaging datasets for continuous mode data visualization," *Analytical and Bioanalytical Chemistry*, pp. 1–7, Oct. 2014.

[127] J. D. Watrous, T. Alexandrov, and P. C. Dorrestein, "The evolving field of imaging mass spectrometry and its impact on future biological research," *Journal of Mass Spectrometry*, vol. 46, pp. 209–222, Jan. 2011.

[128] L. A. Mcdonnell, A. Walch, M. Stoeckli, and G. L. Corthals, "MSiMass list: a public database of identifications for protein MALDI MS imaging.," *Journal of Proteome Research*, vol. 13, pp. 1138–1142, Feb. 2014.

[129] J. Balog, L. Sasi-Szabo, J. Kinross, M. R. Lewis, L. J. Muirhead, K. Veselkov, R. Mirnezami, B. Dezso, L. Damjanovich, A. Darzi, J. K. Nicholson, and Z. Takats, "Intraoperative Tissue Identification Using Rapid Evaporative Ionization Mass Spectrometry," *Science Translational Medicine*, vol. 5, pp. 194ra93–194ra93, July 2013.

[130] K. Schwamborn and R. M. Caprioli, "Molecular imaging by mass spectrometry — looking beyond classical histology," *Nature Reviews Cancer*, vol. 10, pp. 639–646, Aug. 2010.

[131] J. L. Norris and R. M. Caprioli, "Imaging mass spectrometry: a new tool for pathology in a molecular age.," *Proteomics. Clinical applications*, vol. 7, pp. 733–738, Dec. 2013.

List of publications

5.3 International Journal Publications

N. Verbeeck, J. Yang, B. De Moor, R.M. Caprioli, E. Waelkens, R. Van de Plas, "Towards Automated Anatomical Interpretation of Ion Distributions in Tissue: Linking Imaging Mass Spectrometry to Curated Atlases", *Analytical Chemistry*, vol. 86, no. 18, pp. 8974-8982, 2014.

A. Fassbender, N. Verbeeck, D. Boernigen, C.M. Kyama, A. Bokor, A. Vodolazkaia, K. Peeraer, C. Tomassetti, C. Meuleman, O. Gevaert, R. Van de Plas, B. De Moor, Y. Moreau, E. Waelkens, T. D'Hooghe, "Combined Microarray and Proteomic analysis of eutopic endometrium of women with and without endometriosis.", *Human Reproduction*, vol. 27, no. 7, pp. 2020-2029, 2011.

A. Fassbender, E. Waelkens, N. Verbeeck, C.M. Kyama, A. Bokor, A. Vodolazkaia, R. Van de Plas, C. Meuleman, K. Peeraer, C. Tomassetti, O. Gevaert, F. Ojeda, B. De Moor, T. D'Hooghe, "Proteomics Analysis of Plasma for Early Diagnosis of Endometriosis.", *Obstetrics & Gynecology*, vol. 119, no. 2, pp. 276–285, Feb. 2012.

5.4 Oral Talks at National and International Conferences or Workshops

16 Apr 2010 – 16 Apr 2010, *Belgian Society for Mass Spectrometry, Annual Meeting 2010*, Belgium, Brussels, BSMS. N. Verbeeck, R. Van de Plas, B. De Moor, and E. Waelkens, "Exploration of Tissue: Enabling Multivariate Analysis On Large Mass Spectral Imaging Experiments Through Improved Wavelet Compression"

29 March 2012 – 30 March 2012, *NVMS/BSMS Rolduc Conference*, Netherlands, Kerkrade, NVMS. Flash presentation. N. Verbeeck, R. Van de Plas, Y. El Aalamat, B. De Moor, and E. Waelkens, "Inter-tissue Comparison of Mass Spectral Imaging Data Using Group Independent Component Analysis (GICA)"

3 September 2012 – 5 September 2012, *Ourense Conference on Imaging Mass Spectrometry*, Spain, Ourense, Ourcon. N. Verbeeck, R. Van de Plas, Y. El Aalamat, B. De Moor, E. Waelkens "Group Independent Component Analysis (GICA) as a Tool for Discovering Dissimilarities between Tissues in Mass Spectral Imaging"

18 February 2013, *SyBioMa Seminar*, Leuven, Belgium. N. Verbeeck, R. Van de Plas, Y. El Aalamat, B. De Moor, and E. Waelkens. "Imaging Mass Spectrometry: advanced data processing of large and complex data sets"

13 December 2013, *IAP Phase-VII Program Network 13: Integration of cellular signaling pathways in health and disease (ICePath) Annual Meeting*, Belgium, Brussels, Universit Catholique de Louvain. N. Verbeeck, R. Van de Plas, Y. El Aalamat, B. De Moor, and E. Waelkens, "Recent Advances in Mass Spectrometry Imaging"

8 October 2014 - 10 October 2014, *Bioimage Informatics Conference*, Belgium, Leuven, VIB. N. Verbeeck, R. Van de Plas, J. Yang, R.M. Caprioli, B. De Moor, and E. Waelkens, "Guided exploration of imaging mass spectrometry data using the Allen Mouse Brain Atlas"

5.5 Posters at National and International Conferences or Workshops

6 Apr 2010 - 10 Apr 2010, *Interdisciplinary Workshop Sparsity and Modern Mathematical Methods for High Dimensional Data*, Belgium, Brussels, VUB. N. Verbeeck, R. Van de Plas, B. De Moor, and E. Waelkens "Discrete Wavelet Transform Analysis of MSI Data for Feature Selection and Data Compression"

30 Apr 2010 - 30 Apr 2010, *IUAP on Signal integration mechanisms in health and disease*, Belgium, Gembloux, Belgian Science Policy. N. Verbeeck, R. Van de Plas, Janssens, V., Louis, J.V., B. De Moor, and E. Waelkens, "Improved Wavelet Analysis of Mass Spectral Imaging Data for Feature Selection and Data Compression through Incorporation of Spatial Information"

23 May 2010 - 27 May 2010, *58th ASMS Conference on Mass Spectrometry and Allied Topics*, USA, Utah, Salt Lake City, ASMS. N. Verbeeck, R. Van de Plas, B. De Moor, and E. Waelkens "Improved Wavelet Analysis of MSI Data

for Feature Selection and Data Compression through Incorporation of Spatial Information"

26 September 2010 – 29 September 2010, *European Conference on Computational Biology 2010*, Belgium, Gent, KU Leuven. N. Verbeeck, R. Van de Plas, B. De Moor, and E. Waelkens "Compression of mass spectral imaging data using discrete wavelet transform guided by spatial information"

6 Apr 2010 - 10 Apr 2010, *Interdisciplinary Workshop Sparsity and Modern Mathematical Methods for High Dimensional Data*, Belgium, Brussels, VUB, N. Verbeeck, R. Van de Plas, B. De Moor, and E. Waelkens, "Discrete Wavelet Transform Analysis of MSI Data for Feature Selection and Data Compression"

30 Apr 2010 - 30 Apr 2010, *IUAP on Signal integration mechanisms in health and disease*, Belgium, Gembloux, Belgian Science Policy. N. Verbeeck, R. Van de Plas, V. Janssens, J.V. Louis, B. De Moor, and E. Waelkens, "Improved Wavelet Analysis of Mass Spectral Imaging Data for Feature Selection and Data Compression through Incorporation of Spatial Information"

23 May 2010 - 27 May 2010, *58th ASMS Conference on Mass Spectrometry and Allied Topics*, USA, Utah, Salt Lake City, ASMS. N. Verbeeck, R. Van de Plas, B. De Moor, and E. Waelkens, "Improved Wavelet Analysis of MSI Data for Feature Selection and Data Compression through Incorporation of Spatial Information"

26 September 2010 – 29 September 2010, *European Conference on Computational Biology 2010*, Belgium, Gent, KU Leuven. N. Verbeeck, R. Van de Plas, B. De Moor, and E. Waelkens "Compression of mass spectral imaging data using discrete wavelet transform guided by spatial information"

16 December 2010 – 16 December 2010, Flanders Proteomics KVCV-P, Belgium, Antwerp, Flanders Proteomics. N. Verbeeck, R. Van de Plas, B. De Moor, and E. Waelkens "Compression of Mass Spectral Imaging Data Using Discrete Wavelet Analysis with Incorporated Spatial Information." **Best Poster Award**

21 March 2011 – 21 March 2011, *BioMAGNet Annual Meeting*, Belgium, Brussels, ULB. N. Verbeeck, R. Van de Plas, B. De Moor, and E. Waelkens, "Compression of Mass Spectral Imaging Data Using Discrete Wavelet Analysis with Incorporated Spatial Information."

5 June 2011 - 9 June 2011, *59th ASMS Conference on Mass Spectrometry and Allied Topics*, USA, Colorado, Denver, ASMS. N. Verbeeck, B. De Moor, E. Waelkens, and R. Van de Plas "Improved Peak Picking for Mass Spectral Imaging by Integrating Localized Wavelet Information"

12 December 2011 – 13 December 2011, *Benelux Bioinformatics Conference 2011*, Luxembourg, Luxembourg, BBC. N. Verbeeck, B. De Moor, E. Waelkens, and R. Van de Plas "A Custom-Designed Peak Picking Algorithm for Mass Spectral Imaging Data"

29 March 2012 – 30 March 2012, NVMS/BSMS Rolduc Conference, Netherlands, Kerkrade, NVMS. N. Verbeeck, R. Van de Plas, Y. El Aalamat, B. De Moor, and E. Waelkens, "Inter-tissue Comparison of Mass Spectral Imaging Data Using Group Independent Component Analysis (GICA)" **Best Poster Award**

29 March 2012 – 30 March 2012, *NVMS/BSMS Rolduc Conference*, Netherlands, Kerkrade, NVMS. Y. El Aalamat, R. Van de Plas, N. Verbeeck, E. Lerut, J. Berkers, B. De Moor, and E. Waelkens, "Hierarchical Application of Principal Component Analysis in Mass Spectral Imaging Analysis"

20 May 2012 - 24 May 2012, 60th ASMS Conference on Mass Spectrometry and Allied Topics, Canada, BC, Vancouver, ASMS. N. Verbeeck, Y. El Aalamat, B. De Moor, E. Waelkens and R. Van de Plas, "Automated Discovery of Inter-tissue Differential Patterns in Mass Spectral Imaging Data Using Group Independent Component Analysis (GICA)"

3 May 2013, *BSMS (Belgian Society for MS) 2013*, Antwerpen, Belgium. N. Verbeeck, R. Van de Plas, J. Yang, B. De Moor , R. M. Caprioli, E. Waelkens. "Linking Mass Spectrometry Imaging and Mouse Brain Anatomy using the Allen Brain Atlas"

June 9 - 13, 2013, *61st ASMS Conference on Mass Spectrometry and Allied Topics*, Minneapolis, Minnesota, ASMS. N. Verbeeck, Junhai Yang, Bart De Moor, Richard Caprioli, Etienne Waelkens, Raf Van de Plas. "Deeper Insights Through Integration: Linking Imaging Mass Spectrometry Data with Anatomical Data"

9-10 December 2013, *Benelux Bioinformatics Conference 2013*, Brussels, Belgium. Y. El Aalamat, R. Van de Plas, N. Verbeeck, E. Waelkens, B. De Moor "Tensor Decomposition for Data Reduction In Mass Spectrometry Imaging".

9-10 December 2013, *Benelux Bioinformatics Conference 2013*, Brussels, Belgium. N. Verbeeck, R. Van de Plas, Yang J., Caprioli R.M., E. Waelkens, B. De Moor, "Guided Exploration of Mass Spectrometry Imaging Data through Integration with Anatomical Information" **Best Poster Award**



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 Heverlee
nico.verbeeck@esat.kuleuven.be
http://www.esat.kuleuven.be/stadius/